

Universidade Federal do Piauí
Centro de Ciências da Natureza
Programa de Pós-Graduação em Ciência da Computação

Apparent Attention in Classrooms: Probing Vision Models with Synthetic Data

Flávio José Schaefer Ferlin

Flávio José Schaefer Ferlin

Apparent Attention in Classrooms: Probing Vision Models with Synthetic Data

Master's dissertation submitted to the Graduate Program in Computer Science at the Federal University of Piauí (area of concentration: Computer Vision), in partial fulfillment of the requirements for the degree of Master of Science in Computer Science.

Universidade Federal do Piauí – UFPI

Centro de Ciências da Natureza

Programa de Pós-Graduação em Ciência da Computação

Supervisor: Prof. Dr. Ivan Saraiva Silva

Co-supervisor: Prof. Dr. Kelson Rômulo Teixeira Aires

Teresina-PI

Fevereiro de 2026

FICHA CATALOGRÁFICA
Universidade Federal do Piauí
Biblioteca Comunitária Jornalista Carlos Castello Branco
Divisão de Representação da Informação

F357a Ferlin, Flávio José Schaefer.
Apparent Attention in Classrooms : Probing Vision Models with Synthetic Data / Flávio José Schaefer Ferlin. -- 2026.
153 f.

Dissertação (Mestrado) – Universidade Federal do Piauí, Programa de Pós-Graduação em Ciência da Computação, Teresina, 2026.
"Orientador: Prof. Dr. Ivan Saraiva Silva".

1. Atenção visualmente aparente. 2. Metodologia de geração de base sintética. 3. Geometria da Informação. 4. *Representation Learning*. 5. *Vision Foundation Models*. I. Silva, Ivan Saraiva. II. Título.

CDD 006.37

Bibliotecária: Francisca das Chagas Dias Leite – CRB3/1004

Flávio José Schaefer Ferlin

Apparent Attention in Classrooms: Probing Vision Models with Synthetic Data

Master's dissertation submitted to the Graduate Program in Computer Science at the Federal University of Piauí (area of concentration: Computer Vision), in partial fulfillment of the requirements for the degree of Master of Science in Computer Science.

Trabalho aprovado. Teresina-PI, 20 de fevereiro de 2026:

Documento assinado digitalmente
gov.br
IVAN SARAIVA SILVA
Data: 19/03/2026 14:33:01-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Ivan Saraiva Silva
Orientador

Documento assinado digitalmente
gov.br
KELSON ROMULO TEIXEIRA AIRES
Data: 20/03/2026 09:32:38-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Kelson Rômulo Teixeira Aires
Coorientador

Documento assinado digitalmente
gov.br
BRUNO MOTTA DE CARVALHO
Data: 25/03/2026 08:58:37-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Bruno Motta de Carvalho (UFRN)
Examinador Externo

Documento assinado digitalmente
gov.br
VINICIUS PONTE MACHADO
Data: 23/03/2026 11:37:28-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Vinícius Ponte Machado (UFPI)
Examinador Interno

Teresina-PI
Fevereiro de 2026

À minha amada esposa, Carmem Luiza e Silva Nascimento.
Aos meus amados pais, Romualdo Benigno Ferlin e
Zulmira Margarida Maria Schaefer Ferlin, *in memoriam*.

Agradecimentos

À minha esposa, Carmem Luiza e Silva Nascimento — minha parceira e melhor amiga — pelo companheirismo, pelo incentivo constante e pelo apoio firme ao longo desta jornada. Seu sorriso lindo me enche de felicidade e ilumina tudo o que faço, e a generosidade do seu imenso e bondoso coração faz cada esforço valer a pena. Agradeço pela compreensão em todos os momentos, por acreditar em meu potencial mais do que qualquer pessoa neste Universo, pelo amor e pela dedicação que nos unem diariamente, na busca de evoluir em nossa condição humana por meio do conhecimento.

Aos meus pais, Romualdo Benigno Ferlin e Zulmira Margarida Maria Schaefer Ferlin, *in memoriam*, pela dedicação, pelo amor, pelos sacrifícios e pelos valores e exemplos que permanecem como guia em minha vida. Minha mãe, exemplo de ternura e afeto, dona de uma voz tão bela quanto todo o seu ser, que abdicou de tudo para dedicar-se aos filhos. Meu pai, homem sem vaidades, de cultura ímpar, cuja dedicação me inspira a tentar ser, ainda que jamais alcance, um pouco do que ele foi. Amor e gratidão a eles, que repousam junto a Deus, nosso Criador e Senhor.

Aos meus orientadores, Professor Dr. Ivan Saraiva Silva e Professor Dr. Kelson Rômulo Teixeira Aires, pela orientação, pela confiança e pelas discussões que contribuíram decisivamente para o desenvolvimento deste trabalho, registro minha profunda gratidão e admiração pela dedicação às causas do ensino e da pesquisa. Destaco, em especial, a generosidade, a paciência e a lealdade ao longo desta jornada, que me deram condições de seguir adiante e aprofundar minha formação: considero uma honra e um privilégio ímpares ter contado com sua orientação e saber.

Aos membros da banca examinadora, Professor Dr. Bruno Motta de Carvalho e Professor Dr. Vinícius Ponte Machado, pela disponibilidade em avaliar esta dissertação e pelas valiosas contribuições apresentadas.

À *alma mater* Universidade Federal do Piauí, pelo ambiente acadêmico que possibilitou a realização desta pesquisa.

Ao legislador Piauiense, por insculpir no ordenamento da nossa terra querida a previsão de concessão de horário especial para o servidor estudante, e de licença para capacitação de noventa dias por quinquênio de efetivo exercício.

Ao Ministério Público do Estado do Piauí, pela concessão de horário especial para estudo e de licença para capacitação de cento e cinquenta dias, que tornaram possível a dedicação necessária à realização deste trabalho.

Você sabe de onde eu venho?
Venho do morro, do Engenho,
Das selvas, dos cafezais,
Da boa terra do coco,
Da choupana onde um é pouco,
Dois é bom, três é demais,
Venho das praias sedosas,
Das montanhas alterosas,
Dos pampas, do seringal,
Das margens crespas dos rios,
Dos verdes mares bravios
Da minha terra natal.
Por mais terras que eu percorra,
Não permita Deus que eu morra
Sem que volte para lá;
Sem que leve por divisa
Esse "V" que simboliza
A vitória que virá:
Nossa vitória final,
Que é a mira do meu fuzil,
A ração do meu bornal,
A água do meu cantil,
As asas do meu ideal,
A glória do meu Brasil.
(Trechos da Canção do Expedicionário - Guilherme de Almeida e Spartaco Rossi, 1944)

Dai-me, Senhor meu Deus, o que Vos resta;
Aquilo que ninguém Vos pede.
Não Vos peço o repouso nem a tranquilidade,
Nem da alma nem do corpo.
Não Vos peço a riqueza nem o êxito nem a saúde;
Tantos Vos pedem isso, meu Deus,
Que já não Vos deve sobrar para dar.
Dai-me, Senhor, o que Vos resta,
Dai-me aquilo que todos recusam.
Quero a insegurança e a inquietação,
Quero a luta e a tormenta.
Dai-me isso, meu Deus, definitivamente;
Dai-me a certeza de que essa será a minha parte para sempre,
Porque nem sempre terei a coragem de Vo-la pedir.
Dai-me, Senhor, o que Vos resta,
Dai-me aquilo que os outros não querem;
Mas dai-me, também, a coragem,
a força e a fé.
(Oração do Paraquedista - Aspirante a Oficial Paraquedista André Zirnheld, 1942)

Resumo

Ao modelar a atenção visualmente aparente em cenários de sala de aula, um conjunto sintético de dados, rotulado por meio de modelos visão–linguagem e construído sob restrições computacionais, fornece uma base empírica sólida para investigar, em condições controladas, o comportamento de representações visuais contemporâneas pré-treinadas. Os resultados reforçam uma interpretação operacional da utilidade das representações, segundo a qual estas são consideradas eficazes na medida em que permitem a extração eficiente de informações relevantes para a tarefa por meio de classificadores simples. Nesse contexto, análises sensíveis à variância e induzidas pela tarefa da geometria dos *embeddings* indicam que, sob os mecanismos de sondagem e a função de perda adotados, a discriminação supervisionada entre estados atento e desatento induz uma estrutura de sensibilidade efetivamente unidimensional em espaços de *embeddings* de alta dimensão. Essa estrutura fortemente anisotrópica de sensibilidade da perda já está presente ao operar sobre representações congeladas e pode ser exposta de forma confiável com uma quantidade muito reduzida de dados, como evidenciado pela rápida saturação das curvas de aprendizado e pela limitada sensibilidade ao aumento da capacidade do *readout*, indicando que o sinal supervisionado necessário para essa discriminação binária é acessível sem adaptação específica do *backbone* à tarefa.

Motivado pela hipótese de que *Vision Foundation Models* organizam variação relevante para tarefas associadas ao comportamento atencional visualmente aparente quando combinados com *readouts* leves, este trabalho adota uma perspectiva centrada na representação, que prioriza a análise geométrica em detrimento da otimização ponta a ponta. O reconhecimento da atenção visualmente aparente é, assim, formulado como uma investigação das propriedades geométricas induzidas por objetivos supervisionados em espaços de *embeddings* pré-treinados, e não apenas como um problema de decodificação.

Para testar essa hipótese em condições controladas, é construído um conjunto de dados sintético de sala de aula — disponibilizado publicamente — como instrumento experimental para a perturbação seletiva de indícios atencionais visualmente aparentes em uma variedade de atividades em sala de aula e configurações comportamentais, mantendo-se um rótulo binário de atento versus desatento e sem tratar dados sintéticos como substitutos de distribuições do mundo real. As anotações comportamentais para os alvos binários são fornecidas por modelos multimodais visão–linguagem. *Embeddings* provenientes de *Vision Foundation Models* e de um *baseline* convolucional congelados são então avaliados como objetos geométricos sob regimes de *readout* com capacidade variável. O desempenho é avaliado por meio de métricas padrão de classificação (Macro-F1 e ROC–AUC), enquanto as propriedades geométricas induzidas pela tarefa são examinadas por meio de

projeções sensíveis à covariância (Análise Discriminante Linear) e de espectros empíricos da Informação de Fisher.

Os resultados mostram que a separabilidade sob o objetivo binário atento versus desatento emerge com poucas amostras e satura rapidamente, enquanto o aumento da capacidade do *readout* principalmente intensifica a sensibilidade ao longo da mesma direção discriminativa dominante, em vez de induzir direções adicionais estatisticamente independentes induzidas pela tarefa. O *baseline* convolucional apresenta um perfil geométrico induzido pela tarefa qualitativamente semelhante, embora com menor eficiência amostral e desempenho inferior nas etapas subsequentes. Em conjunto, esses resultados fornecem evidência quantitativa e não anedótica de que, para a distinção comportamental estudada e dentro das condições explicitamente definidas do presente projeto experimental — incluindo o protocolo de supervisão, o regime de tamanho amostral explorado e a família de mecanismos de *readout* avaliados —, representações visuais pré-treinadas e mantidas congeladas já tornam acessível a maior parte da variação relevante para a tarefa por meio de adaptadores leves, com benefício marginal limitado decorrente do aumento da complexidade do *readout*. De forma mais ampla, esses achados motivam o uso de diagnósticos geométricos e baseados na teoria da informação como ferramentas para caracterizar como objetivos supervisionados interagem com representações pré-treinadas, esclarecendo tanto seus regimes efetivos de operação quanto suas limitações.

Palavras-chaves: atenção visualmente aparente; base de dados sintética; metodologia de geração de base sintética; Informação de Fisher; Geometria da Informação; Análise Discriminante Linear; *representation learning*; *Vision Foundation Models*.

Abstract

When modeling visually apparent attention in classroom scenes, a synthetically generated dataset labeled via vision–language models—constructed under constrained computational resources—provides a strong empirical basis for probing the behavior of contemporary pretrained visual representations under controlled conditions. The results reinforce an operational interpretation of representation usefulness, in which representations are considered effective insofar as they enable efficient extraction of task-relevant information by simple classifiers. In this context, variance-aware, task-induced analyses of embedding geometry indicate that, under the chosen probe and loss, the supervised attentive versus inattentive discrimination induces an effectively one-dimensional sensitivity structure in high-dimensional embedding spaces. This strongly anisotropic loss sensitivity structure is already present when operating on frozen representations and can be reliably exposed with very little data, as reflected in the rapid saturation of learning curves and the limited sensitivity to increased readout capacity, indicating that the supervised signal required for this binary discrimination is accessible without task-specific adaptation of the backbone.

Motivated by the hypothesis that Vision Foundation Models can support task-relevant variation for visually apparent attentive behavior when paired with lightweight readouts, this work adopts a representation-centric perspective that prioritizes geometric analysis over end-to-end optimization. Visually apparent attention recognition is thus framed as an investigation into the geometric properties induced by supervised objectives in pretrained embedding spaces, not solely as a decoding task.

To test this hypothesis under controlled conditions, a synthetic classroom dataset—made publicly available—is constructed as an experimental instrument for selectively perturbing visually apparent attentional cues across diverse classroom activities and behavioral configurations, while retaining a binary attentive versus inattentive target label and without treating synthetic data as a surrogate for real-world distributions. Behavioral annotations for the binary targets are provided by multimodal vision–language models. Embeddings from frozen Vision Foundation Models and a convolutional baseline are then evaluated as geometric objects under fixed readout regimes with varying capacity. Performance is assessed using standard classification metrics (Macro-F1 and ROC–AUC), while task-induced geometric properties are examined through covariance-aware projections (Linear Discriminant Analysis) and empirical Fisher Information spectra.

Results show that separability under the binary attentive versus inattentive objective emerges with few samples and saturates rapidly, while increasing readout capacity primarily sharpens sensitivity along the same dominant discriminative direction instead of inducing

additional statistically independent task-induced directions. The convolutional baseline exhibits a qualitatively similar task-induced geometric profile, albeit with reduced sample efficiency and lower downstream performance. Taken together, these results provide quantitative, non-anecdotal evidence that, for the behavioral distinction studied and within the explicitly defined conditions of the present experimental design — including the supervision protocol, the explored sample-size regime, and the family of readout mechanisms evaluated — frozen pretrained visual representations already make most of the task-relevant variation accessible to lightweight downstream adapters, with limited marginal benefit from increasing readout complexity. More broadly, these findings motivate the use of geometric and information-theoretic diagnostics as principled tools for characterizing how supervised objectives interact with pretrained representations, clarifying both their effective operating regimes and their limitations.

Keywords: visually apparent attention; synthetic dataset; synthetic dataset generation methodology; Fisher Information; Information Geometry; Linear Discriminant Analysis; representation learning; Vision Foundation Models.

List of Figures

Figure 1 – Dataset Generation Pipeline 31
Figure 2 – Image for the behavior ”absorbed”, labeled as inconclusive by VLMs. . 34
Figure 3 – Image for the behavior ”absorbed”, labeled as attentive by VLMs. . . . 34
Figure 4 – Example of a VLM-evaluation record used to derive the label for Figure 2. 42
Figure 5 – Example of a VLM-evaluation record used to derive the label for Figure 3. 43
Figure 6 – "attentive", seed 8 50
Figure 7 – "curious", seed 86 50
Figure 8 – "engaged", seed 23 50
Figure 9 – "focused on the teacher with concentrated face", seed 29 50
Figure 10 – "listening attentively seated at desk", seed 2 50
Figure 11 – "looking at the teacher with focused expression", seed 20 50
Figure 12 – Synthetic image created by ChatGPT, as CCTV-like image of college student seated, looking towards the back 51
Figure 13 – Synthetic image created by ChatGPT, as student seated in the classroom, looking towards the back 51
Figure 14 – "leaning forward with eyes half closed appearing tired", seed 3 52
Figure 15 – "rubbing temples with closed eyes at desk", seed 1 52
Figure 16 – "supporting head with both hands looking bored", seed 39 52
Figure 17 – "blowing a bubble with gum at the desk", seed 3 52
Figure 18 – "checking a phone", seed 55 52
Figure 19 – "having a side conversation", seed 41 52
Figure 20 – behavior ”focused” 53
Figure 21 – behavior ”intellectually engaged” 53
Figure 22 – Image generated with behavior "ready to learn" and seed 75 53
Figure 23 – Learning curve of Macro-F1 for a linear classification head, as a function of training size, comparing different frozen visual backbones. 57
Figure 24 – Learning curve of Macro-F1 for a shallow MLP classification head, as a function of training size, comparing different frozen visual backbones. . 58
Figure 25 – Learning curve of ROC-AUC for a linear classification head, as a function of training size, comparing different frozen visual backbones. 59
Figure 26 – Learning curve of ROC-AUC for a shallow MLP classification head, as a function of training size, comparing different frozen visual backbones. 59
Figure 27 – Mean paired difference in Macro-F1 between an MLP classification head (512–2) and a linear probe (Δ = MLP − Linear), as a function of training size. 60

Figure 28 – Mean paired difference in ROC-AUC between an MLP classification head (512–2) and a linear probe (Δ = MLP − Linear) as a function of training size. 61

Figure 29 – Mean paired difference in test loss between an MLP classification head (512–2) and a linear probe (Δ = MLP − Linear) as a function of training size. 61

Figure 30 – Mean paired difference in Macro-F1 between an MLP classification head (512–2) with no dropout and the baseline MLP (Δ = No-Dropout − Baseline), as a function of training size. 62

Figure 31 – Mean paired difference in ROC-AUC between an MLP classification head (512–2) with no dropout and the baseline MLP (Δ = No-Dropout − Baseline), as a function of training size. 63

Figure 32 – Mean paired difference in test loss between an MLP classification head (512–2) with no dropout and the baseline MLP (Δ = No-Dropout − Baseline), as a function of training size. 63

Figure 33 – Probability that an MLP classification head achieves higher Macro-F1 than a linear probe (Δ = MLP - Linear) as a function of training size, computed over repeated splits for each frozen backbone. 65

Figure 34 – Probability that an MLP classification head (512–2) with no dropout achieves higher Macro-F1 than the baseline MLP (Δ = No-Dropout - Baseline) as a function of training size, computed over repeated splits for each frozen backbone. 65

Figure 35 – Projection of test-set embeddings onto the one-dimensional Fisher discriminant direction. Despite irregular learning curves, the projection exhibits clear separation between attentive and inattentive samples, with a linear decision boundary lying in a region of low density. 67

Figure 36 – Learning curve of Macro-F1 for Linear Discriminant Analysis, as a function of sample size, comparing different frozen visual backbones. . 67

Figure 37 – Learning curve of ROC-AUC for Linear discriminant analysis, as a function of sample size, comparing different frozen visual backbones. . 68

Figure 38 – Top-1 spectral energy (E1) of the representation-space Fisher Information Matrix on the test split, as a function of training set size, for different frozen visual backbones. Higher values indicate stronger concentration of Fisher information along a single dominant direction, suggesting increased anisotropy in the embedding geometry. 71

Figure 39 – Shannon effective rank of the representation-space Fisher Information spectrum on the test split, as a function of training set size, for different frozen visual backbones. The effective rank summarizes how Fisher information is distributed across spectral components, with lower values indicating greater spectral concentration. 72
Figure 40 – Trace of the Fisher Information Matrix computed in representation space on the test split, as a function of training set size, for different frozen visual backbones. The trace measures the total Fisher information accumulated across all representation dimensions, providing a global summary of the sensitivity of the classifier to perturbations in the embedding space. 74
Figure 41 – labeled as attentive, level 5 78
Figure 42 – labeled as inattentive, level 2 78
Figure 43 – labeled as attentive, level 5 78
Figure 44 – label from json file 79
Figure 45 – label from json file 79
Figure 46 – label from json file 79
Figure 47 – Confusion Matrix (Counts) 80
Figure 48 – absorbed 141
Figure 49 – actively listening 142
Figure 50 – attentive 143
Figure 51 – concentrated 144
Figure 52 – curious 145
Figure 53 – checking a phone 146
Figure 54 – dozing 147
Figure 55 – hunched over 149
Figure 56 – with silent attention 150
Figure 57 – immersed in the task 151

List of Tables

Table 1 – Dataset composition before and after deduplication and similarity-based filtering. 54
Table 2 – Synthetic image groups by prompt family. 95
Table 3 – N50 – Train→Test activity overlap, Repetition 1. *Train*: number of training examples; *Test*: occurrences in test set. A dash (–) indicates the activity is absent from that split. 99
Table 4 – N200 – Train→Test activity overlap, Repetition 1. *Train*: number of training examples; *Test*: occurrences in test set. A dash (–) indicates the activity is absent from that split. 104
Table 5 – N400 – Train→Test activity overlap, Repetition 1. *Train*: number of training examples; *Test*: occurrences in test set. A dash (–) indicates the activity is absent from that split. 109
Table 6 – N1600 – Train→Test activity overlap, Repetition 1. *Train*: number of training examples; *Test*: occurrences in test set. A dash (–) indicates the activity is absent from that split. 115
Table 7 – Activities as distributed among classes by Vision-Language Models . . . 121
Table 8 – Fisher metrics for **CNN** representations on the **Linear Probe** adapter architecture 134
Table 9 – Fisher metrics for **CNN** representations on the **MLP** adapter architecture 135
Table 10 – Fisher metrics for **CLIP** representations on the **Linear Probe** adapter architecture 136
Table 11 – Fisher metrics for **CLIP** representations on the **MLP** adapter architecture 137
Table 12 – Fisher metrics for **DINO** representations on the **Linear Probe** adapter architecture 138
Table 13 – Fisher metrics for **DINO** representations on the **MLP** adapter architecture 139

Contents

1 INTRODUCTION . . . 1
1.1 Problem Definition and Motivation . . . 1
1.2 *Attention* as a Semantic Hypothesis . . . 3
1.3 Scope and Deliberate Constraints . . . 3
1.4 Vision Foundation Models and Representation Power . . . 4
1.5 Synthetic Data as an Experimental Instrument . . . 5
1.6 Objectives . . . 6
1.7 Research Hypotheses . . . 6
1.8 Contributions . . . 6
1.9 Structure of the Dissertation . . . 7

I RESEARCH PREPARATION 9

2 THEORETICAL FRAMEWORK . . . 11
2.1 Visual Foundation Models and Embedding Spaces . . . 11
2.2 Global and Local Representations in Vision Transformers . . . 12
2.3 Latent Diffusion Models as Experimental Instruments . . . 12
2.4 Vision–Language Models . . . 13
2.5 Linear Discriminant Analysis and Fisher's Criterion . . . 13
2.5.1 Linear Discriminant Analysis versus Logistic Regression . . . 14
2.6 Fisher-Based Analysis of Representation Geometry . . . 15
2.6.1 Empirical Fisher in Representation Space . . . 15
2.6.2 Spectral Geometry of Representation Sensitivity . . . 15
2.6.3 Spectral Summaries . . . 16
2.6.4 Practical Computation . . . 16
2.7 Synthesis . . . 17

3 RELATED WORK . . . 19
3.1 Attention and Engagement Analysis in Educational Contexts . . . 19
3.2 Visual Proxies for Attention and Behavioral Cues . . . 20
3.3 Representation Learning and Probing of Pretrained Visual Models . . . 21
3.4 Vision–Language Models in Affective and Expression Recognition . . . 22
3.5 Synthetic Data and Generative Models as Experimental Instruments 23
3.5.1 Definition and Conventional Use of Synthetic Data . . . 23
3.5.2 Synthetic Data as an Enabling Resource . . . 23

3.5.3 Legal and Regulatory Constraints . . . 24
3.5.4 Data Scarcity and Rare Phenomena . . . 24
3.5.5 Biases and Dataset Composition . . . 25
3.5.6 Advances in Generative Models for Image Synthesis . . . 25
3.5.7 Epistemological Position: Synthetic Data as Instrument . . . 25
3.6 Vision–Language Models as Evaluators . . . 26
3.7 Summary of Research Trajectories . . . 26

II PROPOSED APPROACH 27

4 METHODOLOGY . . . 29
4.1 Proposed Dataset . . . 29
4.1.1 *Synthetic Data* as a Lens into Representation Spaces . . . 29
4.1.2 *Scope Delimitation*: Biases and Demographic Factors . . . 30
4.1.3 Dataset Generation Pipeline Overview . . . 30
4.1.4 Prompt Taxonomy and Behavioral Proxies . . . 31
4.1.5 Image Generation . . . 32
4.1.6 Automated Labeling with Vision–Language Models . . . 33
4.1.7 Construction of Train, Validation, and Test Splits . . . 35
4.1.7.1 Fixed test split . . . 35
4.1.7.2 Nested training sets . . . 35
4.1.7.3 Validation split construction . . . 36
4.1.7.4 Repetitions and seed control . . . 36
4.1.7.5 Class balance and stratification . . . 36
4.1.7.6 Leakage prevention and auditing . . . 36
4.1.7.7 Stability and overlap diagnostics . . . 36
4.1.7.8 Rationale . . . 37
4.1.7.9 Representative Samples for the Training Splits . . . 37
4.2 Representation-First Experimental Methodology . . . 37
4.2.1 Frozen Backbones and Lightweight Adapters . . . 38
4.2.2 Learning Curves and Sample Efficiency . . . 38
4.2.3 Linear Discriminant Analysis as a Geometric Probe . . . 39
4.2.4 Empirical Fisher Information and Task-Induced Geometry . . . 39
4.2.5 Why Empirical Fisher Instead of the Hessian . . . 39
4.2.6 Unifying View: Fisher Criterion, LDA, and Information Geometry . . . 40
4.2.7 Cross-Backbone Comparison . . . 41
4.2.8 Methodological Scope . . . 41

III EXPERIMENTAL EVALUATION AND CONCLUSIONS 45

5 RESULTS . . . 47
5.1 Dataset Results . . . 47
5.1.1 Vision Foundation Models, the Probing Paradox, and Instrumental Data for Identifiable Representation Analysis . . . 47
5.1.2 Dataset Composition . . . 49
5.1.3 Deduplication and Similarity-Based Filtering . . . 52
5.1.4 Assessments by Vision–Language Models . . . 54
5.1.5 Summary and Dataset Limitations . . . 54
5.2 Experimental Probing Results . . . 55
5.2.1 Evaluation Protocol . . . 55
5.2.2 Learning Curves as a Function of Data and Probe Capacity . . . 56
5.2.2.1 Macro-F1 learning curves with linear probes . . . 56
5.2.2.2 Effect of shallow non-linearity on Macro-F1 . . . 56
5.2.2.3 ROC-AUC learning curves and metric consistency . . . 57
5.2.2.4 Paired Analysis of Probe Architectures . . . 58
5.2.2.4.1 MLP versus linear probe . . . 58
5.2.2.4.2 Effect of dropout regularization . . . 60
5.2.2.4.3 Win-rate analysis across repetitions . . . 62
5.2.3 Closed-form Linear Probes: Fisher Linear Discriminant Analysis . . . 64
5.2.3.1 LDA learning curves under frozen visual representations . . . 64
5.2.4 Probing Representation Geometry with Fisher Information . . . 66
5.2.4.1 Metrics and reading guide . . . 68
5.2.4.2 Motivation for MLP Adapter Analysis . . . 69
5.2.4.3 Top-1 spectral energy (E1): Is the task effectively one-dimensional in representation space? . . . 70
5.2.4.4 Shannon effective rank: How many representation dimensions meaningfully contribute to the decision? . . . 70
5.2.4.5 A dominant task direction emerges quickly and persists across backbones. . . . 71
5.2.4.6 Fisher trace: How sensitive is the classifier to perturbations in representation space? 72
5.2.4.7 Interpreting the convergence of effective rank as a function of sample size in the MLP regime . . . 73
5.2.4.8 Operational use: Fisher as a representation audit . . . 75
5.2.4.8.1 Audit Question 1: Is the task already geometrically aligned with the representation? . . . 75
5.2.4.8.2 Audit Question 2: Does additional capacity reveal structure or merely amplify it? . . . 75
5.2.4.8.3 Audit Question 3: How energetically is the task encoded? . . . 76
5.2.4.8.4 Minimal protocol . . . 76
5.2.5 Summary of Findings . . . 76
5.3 Results with an Out-of-Domain Dataset . . . 77

6 CONCLUSIONS 83

BIBLIOGRAPHY 87

ANNEX 93

ANNEX A – CONTROLLED VOCABULARIES USED FOR SYNTHETIC IMAGE GENERATION 95
A.1 Attentive synonyms (4,300 images) 95
A.2 Attentive behavioral proxies (2,700 images) 96
A.3 Inattentive activities (6,900 images) 96
A.4 Inattentive behavioral proxies (1,000 images) 98

ANNEX B – SAMPLE TRAIN, TEST SPLITS COMPOSITION . . 99

ANNEX C – ACTIVITIES AS DISTRIBUTED AMONG CLASSES BY VISION-LANGUAGE MODELS 121

ANNEX D – FISHER METRICS 133

ANNEX E – SAMPLE IMAGES, AS LABELED BY VISUAL LANGUAGE MODELS 141

1 Introduction

1.1 Problem Definition and Motivation

Attention is a fundamental yet elusive concept. In *The Principles of Psychology*, William James provides a definition that remains influential more than a century later (JAMES, 2018):

> Every one knows what attention is. It is the taking possession by the mind, in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought. Focalization, concentration, of consciousness are of its essence. It implies withdrawal from some things in order to deal effectively with others, and is a condition which has a real opposite in the confused, dazed, scatter-brained state which in French is called *distraction*, and *Zerstreutheit* in German.

James further emphasizes the profound consequences of attention for both individual and collective life:

> Its remote effects are too incalculable to be recorded. The practical and theoretical life of whole species, as well as of individual beings, results from the selection which the habitual direction of their attention involves. [...] Suffice it meanwhile that each of us literally **chooses**, by his ways of attending to things, what sort of a universe he shall appear to himself to inhabit. The immediate effects of attention are to make us: a) perceive; b) conceive; c) distinguish; d) remember better than otherwise we could–both more successive things and each thing more clearly. It also e) shortens 'reaction-time.'

These classical formulations situate attention as a core organizing function of mental life, but they also underscore its intrinsically subjective character. Classical psychological accounts have long emphasized that, as a phenomenon of consciousness, attention is accessible only through introspection and report, and does not admit direct external observation. It resists precise operational definition, despite being intuitively understood at a phenomenological level. Empirical study therefore proceeds indirectly, by examining the behavioral and perceptual consequences through which attentional orientation becomes externally manifest.

From a broader interdisciplinary perspective, Mancas et al. (MANCAS et al., 2016) emphasize the ubiquity of attention, describing it as the first step of perception in living

beings, active from fetal development through wakefulness, sleep, and dreaming. The authors argue that attentional mechanisms emerge in parallel with sensory systems and are deeply intertwined with perception, cognition, and behavior. They further note that attentional processes are significantly affected in various cognitive disorders, such as autism and schizophrenia, reinforcing the centrality of attention across biological and psychological domains.

Mancas et al. characterize attention as a foundational component of perception, operating continuously across the lifespan and across states of consciousness. They further note that, despite its centrality, attention only became a systematic object of scientific investigation relatively recently. Finally, they emphasize the expansion of attention research from Philosophy and Psychology into Neuroscience and Computer Science.

Within the domain of *computational attention*[1], the authors suggest that attentional mechanisms may play a key role in the development of more autonomous and adaptive artificial systems. Unlike classical computers that process all inputs uniformly, an attentive system selectively focuses on information deemed relevant, enabling more efficient prediction, pattern discovery, and adaptive behavior. In this view, computational attention is not merely a technical optimization, but a foundational mechanism that may support curiosity-driven exploration and, ultimately, creativity.

This progression—from philosophical treatments of attention, through cognitive and neuroscientific perspectives, to computational formulations—underscores the conceptual context for the present analysis. These observations situate attention as both a fundamental and a methodologically challenging construct, providing the conceptual backdrop against which the present analysis of visual proxies is developed. Because the underlying attentional processes are not directly accessible to observation, empirical studies must rely on externally observable behavioral traces as correlates.

In educational settings, the literature describes attention-related behaviors such as body posture, gaze orientation, and interaction with learning materials, as contextual indicators of attentional orientation. Because contemporary visual encoders are trained on large-scale image corpora to capture general-purpose visual structure, their embedding spaces may implicitly encode patterns related to such observable behaviors.

This domain is therefore a natural testbed for probing how, and to what extent, pretrained visual embeddings encode separable structure associated with such proxies. At the same time, the collection and sharing of real classroom imagery involving minors is heavily constrained, which motivates the use of synthetic classroom-like scenes as an

[1] Mancas et al. discuss attention primarily in the context of saliency and perceptual selection models, grounded in cognitive and biological theories of attention. This usage is conceptually distinct from the dot-product attention mechanism introduced in Transformer architectures (Vaswani et al., 2017), which is a specific computational operation for weighting token interactions, not a model of perceptual or cognitive attention.

experimental instrument which should not be construed as a substitute for real-world observation.

1.2 *Attention* as a Semantic Hypothesis

This work adopts a deliberately conservative stance: it does not intend to claim to detect attention as a cognitive state, but focuses on *apparent attention*, understood as a set of visually observable configurations that are commonly interpreted as indicative of attentiveness or distraction. This distinction is essential, as visually attentive behavior does not guarantee cognitive engagement, and conversely, behaviors that appear inattentive may coexist with deep mental focus.

Also, attention is not treated as a target variable to be learned, but as a semantic hypothesis about structure within pretrained visual representations. Attention is not defined as a ground-truth label whose prediction accuracy is to be optimized: the study asks whether visually apparent attentional configurations—operationalized through observable proxies such as posture, gaze direction, and object interaction—correspond to coherent organization in embedding space.

Under this framing, attention functions as a conceptual reference point that motivates the analysis, without imposing a fixed operational definition or a single evaluation criterion. Model-based judgments are used instrumentally to structure and inspect variation, but should not be interpreted as defining attention as a supervised learning target. Large language models and latent diffusion models are thus employed to instantiate a hypothesis space of what visually apparent attention may look like, which is rendered into images and subsequently interrogated through the geometry of pretrained visual embeddings.

The central question is therefore not whether any single model can correctly identify attention, but whether hypothesized visual distinctions associated with attentional states induce consistent and interpretable structure within representation space.

1.3 Scope and Deliberate Constraints

Numerous modalities have been explored in the literature for attention assessment, including electroencephalography, electrocardiography, eye tracking, and functional neuroimaging. While such approaches may provide richer signals, they also introduce practical, ethical, and economic constraints that limit their applicability in real educational environments.

This dissertation deliberately restricts its scope to **static image-based visual representations**, without relying on temporal dynamics, wearable sensors, or physiological

measurements. This choice is motivated in part by the nature of the dataset constructed here, which is composed of static images by design, and by the reliance on CLS tokens as the primary representational unit — whose capacity to compress holistic scene information into a single vector has been well-documented in the vision transformer literature. While this entails a loss of temporal and physiological richness, it also isolates the representational signal available from visual cues alone, clarifying what can be inferred from a single frame before layering in more complex modalities.

1.4 Vision Foundation Models and Representation Power

Over the past decade, Computer Vision has undergone a substantial transformation with the emergence of Transformer-based architectures and large-scale pretraining. Vision Foundation Models, such as Vision Transformers and multimodal models trained on image–text pairs, have demonstrated representational capacities that extend beyond the object-centric classification paradigm exemplified by the ImageNet Large Scale Visual Recognition Challenge (RUSSAKOVSKY et al., 2015).

Notably, recent multimodal systems can generate detailed textual descriptions of complex scenes and synthesize realistic images from textual prompts, suggesting that their internal representations encode high-level semantic and relational information. These developments raise a natural question: to what extent do such pretrained visual representations already contain signal relevant to abstract behavioral concepts, such as apparent attention, without task-specific fine-tuning?

We do not assert that these models "understand" attention, for this work treats their embeddings as *objects of empirical investigation*. The central interest lies in assessing whether these representations exhibit separability, consistency, and structure when confronted with carefully constructed visual stimuli associated with different apparent attention states.

From recognition to richer descriptions. A useful historical perspective is that, even at the peak of CNN-based recognition systems, influential voices in the field emphasized the gap between label prediction and human-level scene understanding. For instance, in the context of Computer Vision education and research circa 2019, it was still common to illustrate how an image may be trivially labeled (e.g., "person", "scale", "locker room") while still requiring rich commonsense knowledge to explain what is happening and why it is socially meaningful [2] [3] . This contrast highlights a key limitation of purely discriminative pipelines: strong performance on classification benchmarks does not necessarily imply the ability to represent interactions, intentions, or context.

[2] <https://karpathy.github.io/2012/10/22/state-of-computer-vision/>
[3] <https://youtu.be/dJYGatp4SvA?si=US_vlUdA-xj8W9w4&t=2619>

In the following years, large-scale pretraining and the rise of Transformer-based Vision Foundation Models substantially expanded what can be extracted from visual representations, particularly when combined with language supervision (DOSOVITSKIY et al., 2020; LIU et al., 2023). Although such systems should not be assumed to comprehend scenes in the human sense, their embeddings encode semantic regularities that extend beyond object categories. In Vision Transformers, local image patches (e.g., 14×14 pixels in the input space) are mapped to continuous token embeddings, which are then globally integrated through attention-based interactions. This process enables the emergence of structured, relational representations that support complex downstream inferences, despite being grounded entirely in statistical learning, not symbolic reasoning.

This observation motivates the methodological use of synthetic data generation (ROMBACH et al., 2022) as an experimental instrument to examine pretrained visual representations, in their relationships to visually apparent attention in classroom-like settings, as opposed to using the dataset for fine-tuning.

1.5 Synthetic Data as an Experimental Instrument

A central methodological challenge in studying attention is the lack of reliable, unbiased, and ethically deployable datasets, particularly when minors are involved. Real-world datasets are often limited in size, affected by demographic biases, or constrained by privacy considerations. As explicitly acknowledged in the classroom behavior analysis literature, many of the datasets underpinning recent advances in this area remain unpublished and inaccessible, as they are derived from real classroom monitoring data and cannot be publicly released due to ethical and legal constraints. As a result, a substantial portion of prior work relies on proprietary or non-disclosable datasets, limiting reproducibility, cross-study comparability, and independent verification of reported findings (e.g., Student Classroom Behavior Detection based on YOLOv7-BRA and Multi-Model Fusion (YANG et al., 2024)).

To address these issues, this dissertation employs **synthetic image generation** using established latent diffusion models, selected for their stability and feasibility under constrained computational resources. Synthetic data is not introduced as a substitute for real-world observations, nor as ground truth, but as an *instrumental probe*. By inducing, but not strictly controlling, visual factors such as posture, gaze direction, object interaction, and classroom context through prompt-based generation, synthetic images enable a systematic and reproducible exploration of how Vision Foundation Model embeddings respond to variations commonly associated with apparent attention and distraction—an analysis that would be infeasible under the access and disclosure constraints governing real classroom data.

1.6 Objectives

The primary objective of this dissertation is to jointly assess the contribution of controlled dataset design and pretrained visual representations to the task of discriminating visually apparent attentive and inattentive states in classroom-like scenarios, without fine-tuning the underlying models. In particular, the study examines whether supervision provided by contemporary Vision–Language Models—used as high-cost semantic annotators—can be distilled into lightweight linear probes operating on frozen visual embeddings, *yielding comparable discriminative behavior at a fraction of the computational cost.*

The work defines a conservative operational notion of apparent attention grounded in observable visual attributes and introduces a synthetic image dataset constructed as an experimental instrument for systematically probing these attributes under controlled variations.

Using this dataset, the study then takes the assessments produced by the VLMs and evaluates embeddings extracted from frozen Vision Foundation Models and a convolutional baseline when paired with lightweight downstream classifiers, examining both predictive performance and task-induced geometric structure. By treating the dataset and the representations as co-equal components of the experimental framework, the analysis characterizes the extent to which apparent attention can be supported by pretrained visual representations, once VLM-provided supervision has been distilled.

1.7 Research Hypotheses

This investigation is guided by the hypothesis that embeddings produced by pretrained Vision Foundation Models exhibit structured variation that can be operationally associated with visually apparent attentive and inattentive behavior, as defined and instantiated through VLM-based supervision. It is further hypothesized that this structure can be revealed without fine-tuning the visual backbone on task-specific data, and that synthetic data, when used as a controlled probing instrument, is sufficient to expose both the discriminative capacity and the limitations of these pretrained representations.

These hypotheses are exploratory in nature and do not assume equivalence between computational attention mechanisms and human cognitive attention, which is treated exclusively through visually observable behavioral proxies.

1.8 Contributions

This work seeks to make the following contributions:

- It introduces a synthetic classroom dataset designed explicitly as an *experimental*

instrument for controlled perturbation of attention-related visual configurations, not to be interpreted as a surrogate for real data. This enables systematic probing of representational structure under known semantic variations.

- It aggregates annotation votes for over 12,000 images, from multiple contemporary Vision–Language Models and curates the resulting image–label associations in a form that permits analysis of intra and inter-model agreement, disagreement, and the exploration of potential systematic tendencies in VLM-based supervision.
- It provides an empirical analysis of frozen embeddings from Vision Foundation Models and a convolutional baseline using deliberately constrained diagnostic models (linear probes, shallow MLPs), not focusing on peak predictive performance, but accessibility and organization of task-relevant information.
- It integrates covariance-aware and information-geometric tools—including Linear Discriminant Analysis and empirical Fisher Information, to characterize how apparent attentional contrast constrains the embedding manifold, under the task-relevant supervised objective.
- It sheds light on the perspective that visually apparent attentional contrast might induce coherent, low-rank, and anisotropic structure in modern visual representations, with task-relevant variation concentrated along a small number of dominant directions, and that convolutional baselines exhibit less efficient geometric organization.
- It suggests a representation-geometric perspective as a possible avenue for studying attention in visual models, treating embeddings as more than just intermediate features for end-to-end classification, but as geometric objects to be probed, and better understood.

1.9 Structure of the Dissertation

This dissertation is organized as follows. Chapter 2 presents the theoretical background on Transformer architectures, visual representation learning, generative models, and Fisher's contributions to the study of representation geometry. Chapter 3 reviews related work. Chapter 4 describes the construction and rationale of the synthetic dataset, the assessment by the Vision Language Models, and of the probing experiments. Chapter 5 details the experimental setup and results. Finally, Chapter 6 discusses the findings, limitations, and directions for future research.

Part I

Research Preparation

2 Theoretical Framework

This chapter introduces the theoretical concepts and analytical tools that will be used throughout the study. The discussion is organized around Visual Foundation Models and their embedding spaces, representation learning, the distinction between global and local representations in Vision Transformers, the use of latent diffusion models as experimental instruments, the role of Vision–Language Models as evaluators, and the application of Linear Discriminant Analysis and Fisher Information to the study of representation geometry.

2.1 Visual Foundation Models and Embedding Spaces

Recent advances in Computer Vision have been driven by large-scale self-supervised and weakly supervised models trained on massive image corpora. These models, commonly referred to as *Visual Foundation Models* (VFMs), aim to learn general-purpose visual representations that can be transferred to downstream tasks with minimal or no task-specific supervision.

Architectures such as Vision Transformers (ViT) (DOSOVITSKIY et al., 2020) and their derivatives, including CLIP (RADFORD et al., 2021), DINO (CARON et al., 2021) and DINOv3 (SIMÉONI et al., 2025), for instance, exemplify this paradigm. Instead of learning task-specific decision boundaries, these models learn high-dimensional embedding spaces in which semantic, geometric, and contextual regularities of images are implicitly encoded. Downstream tasks are then framed as operations on these embeddings, such as linear probing, clustering, metric learning, or lightweight nonlinear classification.

From the perspective of this work, the key assumption is not that Vision Foundation Models solve the problem of attention recognition directly. We propose that their embedding spaces already contain latent signals correlated with observable behavioral cues such as posture, gaze direction, body orientation, and interaction with objects. Vision–Language Models are accordingly treated not as explanatory models of attention, but as sources of operational supervision that instantiate these cues at scale. The research question thus shifts from *model design* to *representation interrogation*: how, and to what extent, attention-related information—defined operationally through VLM-provided assessments—is separable and exploitable within pretrained embedding spaces.

This framing justifies the adoption of frozen backbones throughout the experimental pipeline. By keeping the backbone fixed, observed differences reflect how task-relevant information is organized, accessed, and constrained within the representation space and

by the structure of the data, not attributable to end-to-end task memorization.

2.2 Global and Local Representations in Vision Transformers

Vision Transformers process images as sequences of patch embeddings, augmented by special tokens such as the classification token (`[CLS]`). Each patch embedding corresponds to a local image region (patch footprint in input space), analogous to a local receptive field in convolutional architectures, while the `[CLS]` token is intended to aggregate global information through attention mechanisms.

In practice, this architectural choice introduces a fundamental distinction between *global* and *local* representations. The `[CLS]` embedding often captures dominant semantic content and scene-level cues, whereas patch embeddings retain spatially localized information. This distinction is not merely architectural but has direct consequences for downstream tasks, depending on their specific goals.

An analogous separation between local feature extraction and global representation has long been present in convolutional neural networks through the use of *Global Average Pooling* (GAP). Introduced as an alternative to fully connected classification heads, GAP aggregates spatial feature maps into a single global descriptor by averaging activations across spatial locations, thereby producing a compact, translation-robust representation aligned with semantic content (LIN; CHEN; YAN, 2014). In this sense, GAP plays a role functionally comparable to that of the `[CLS]` token, serving as a global summary of distributed local features.

Attention is treated here as an emergent property inferred from multiple observable visual cues, such as gaze direction, limb posture, and object interaction, as opposed to a single localized feature. Under this view, the analysis deliberately focuses on global representations—namely the `[CLS]` token in Vision Transformers and Global Average Pooling in convolutional networks—on the premise that such representations are sufficient to support the identification of visually apparent attentional states.

This perspective directly informs the experimental design adopted in subsequent chapters.

2.3 Latent Diffusion Models as Experimental Instruments

Latent Diffusion Models (LDMs) (ROMBACH et al., 2022), such as Stable Diffusion and its variants, provide a mechanism for generating images by performing the diffusion process in a learned latent space, unlike previous models, that operated in pixel space.

Unlike traditional data augmentation techniques, latent diffusion models allow

for the *induction and approximation* of semantic attributes through textual prompts, random seeds, and guidance parameters. These mechanisms introduce structured but noisy variations – instead of providing strict control – that enable the systematic generation of image distributions aligned with predefined operational definitions of attention and inattention, expressed in terms of observable behaviors and contextual configurations.

2.4 Vision–Language Models

Vision–Language Models (VLMs), such as LLaVA (LIU et al., 2023), QwenVL (BAI et al., 2025), and related architectures, integrate visual encoders with large language models to enable aligned cross-modal representations and language-conditioned processing of visual inputs. These models are typically trained on large-scale image–text corpora, using combinations of contrastive, generative, and instruction-following objectives, with the aim of grounding linguistic representations in visual content.

Architecturally, VLMs combine a pretrained vision backbone—often a convolutional network or Vision Transformer—with a large language model through a learned projection or adapter mechanism. Visual features extracted from the image are embedded into a representation space compatible with the token-based processing of the language model, allowing visual information to be incorporated into the autoregressive generation of text. This design enables VLMs to perform multimodal tasks such as image captioning, visual question answering, and language-conditioned image classification.

From a representational perspective, VLM outputs do not correspond to explicit symbolic labels or deterministic semantic annotations. Instead, they reflect probabilistic associations between visual patterns and linguistic descriptions learned during pretraining. As a consequence, VLM responses are inherently contextual and sensitive to the phrasing, polarity, and framing of the textual prompt, as well as to the inductive biases of the underlying vision and language components.

Because VLMs operate by translating visual information into linguistic form, their judgments can be understood as soft, language-mediated evaluations of visual content rather than as direct measurements of underlying visual attributes. This characteristic distinguishes VLM-based assessments from traditional supervised classifiers and situates them closer to natural-language driven interfaces for interrogating visual representations.

2.5 Linear Discriminant Analysis and Fisher's Criterion

Linear Discriminant Analysis (LDA) is historically rooted in the work of Ronald A. Fisher and is therefore often referred to as *Fisher's Linear Discriminant*. Fisher originally introduced the method in the context of two-class classification problems,

seeking a projection direction w that separates the classes by maximizing the squared distance between their projected means while minimizing the sum of their projected variances (FISHER, 1936). Concretely, if m_1 and m_2 denote the class means and s_1^2 and s_2^2 the class variances after projection onto w, the *Fisher criterion* is

$$J(w) = \frac{(m_1 - m_2)^2}{s_1^2 + s_2^2}. \tag{2.1}$$

In modern notation, this can be written as

$$J(w) = \frac{w^\top S_B w}{w^\top S_W w}, \tag{2.2}$$

where S_B and S_W denote the between-class and within-class scatter matrices, respectively. This objective is a Rayleigh quotient and therefore admits a closed-form solution obtained by solving a generalized eigenvalue problem, yielding a single discriminant direction in the two-class case. Importantly, Fisher's formulation is purely criterion-based and does not invoke probabilistic assumptions about the class-conditional distributions.

Subsequently, Rao extended Fisher's discriminant analysis to the case of three and more groups, formulating classification rules under explicit probabilistic assumptions (RAO, 1948). In his treatment of multivariate normal populations with a common dispersion (covariance) matrix, Rao shows that the resulting decision surfaces are linear and coincide with Fisher's discriminant function in the two-group case, while generalizing naturally to multiple groups. When the class-conditional covariance matrices differ, the quadratic terms in the log-likelihood no longer cancel, and the corresponding decision functions become quadratic, giving rise to what is later known as Quadratic Discriminant Analysis (QDA). The shared-covariance, multiclass formulation developed in this line of work constitutes a central step in the development of what is now commonly presented in textbooks as Linear Discriminant Analysis (LDA).

2.5.1 Linear Discriminant Analysis versus Logistic Regression

Linear Discriminant Analysis (LDA) and logistic regression are both linear classification methods, but they interrogate representation spaces in different ways. Logistic regression is a discriminative model that estimates class boundaries by optimizing the conditional log-likelihood of labels given features, without modeling the structure of the feature distribution. As a result, it primarily assesses whether a linear decision boundary can be reliably supported by the embedding space under the chosen supervision objective.

By contrast, LDA explicitly incorporates second-order structure through the within-class covariance matrix. LDA evaluates whether class separation is supported by a coherent and well-conditioned variance structure, instead of focusing solely on the existence of a separating direction. In this sense, LDA probes not only linear separability, but also how class variability is organized in the representation space.

From a representation analysis perspective, this distinction is critical. A dataset may be linearly separable while exhibiting ill-conditioned class distributions. LDA is explicitly sensitive to such second-order structure through its dependence on within-class covariance, whereas logistic regression does not directly interrogate the variance structure of the representation space. For this reason, LDA provides complementary geometric information about the embedding space that goes beyond the mere existence of a linear decision boundary.

2.6 Fisher-Based Analysis of Representation Geometry

2.6.1 Empirical Fisher in Representation Space

Fisher Information provides a principled way to quantify local sensitivity of a loss function with respect to infinitesimal perturbations of its arguments. In its classical formulation, this sensitivity is defined with respect to model parameters and induces a Riemannian metric on the parameter manifold, capturing directions along which changes in the parameters most strongly affect the model's output or likelihood (AMARI; KARAKIDA; OIZUMI, 2019).

Beyond its classical use for analyzing parameter-space learning dynamics, Fisher Information has also been employed to probe the geometric organization of learned representations. In this line of work, the variable of interest is shifted from model parameters to the representation vector itself (BERARDINO et al., 2018; ACHILLE et al., 2019), treating the embedding space as the object endowed with a task-induced metric.

Similar to the true Fisher Information Matrix, the empirical Fisher is positive semi-definite. However, it offers a significant computational advantage: it can be obtained at virtually no additional cost when the gradient of the objective function is already being evaluated. Moreover, it remains applicable even in contexts where the objective function lacks an explicit probabilistic interpretation, such as reconstruction loss in autoencoders or margin-based objectives in support vector machines, where no likelihood or probability distribution is defined (MARTENS, 2020).

2.6.2 Spectral Geometry of Representation Sensitivity

The empirical Fisher matrix measures the covariance of loss gradients with respect to representation vectors z. This quantity characterizes the local geometry of the classifier's sensitivity to perturbations in embedding space. Following the interpretation in (AMARI, 1998; SCHRAUDOLPH, 2002), large eigenvalues correspond to directions of high predictive sensitivity, while small eigenvalues indicate directions to which the loss is comparatively insensitive. In this sense, the Fisher spectrum provides a geometric view of how task-related

variation is distributed across the representation (ACHILLE et al., 2019; FEATHER et al., 2025). Such spectral characterizations allow the structure of the representation space to be analyzed without direct dependence on its ambient dimensionality, in line with approaches based on effective rank and related measures (ROY; VETTERLI, 2007).

2.6.3 Spectral Summaries

Let $\{\lambda_i\}_{i=1}^{d}$ denote the eigenvalues of the empirical Fisher matrix $\mathbf{I}(z)$. Three scalar summaries capture complementary aspects of the spectrum:

1. **Top-1 Spectral Energy (E1)**

$$E_1 = \frac{\lambda_1}{\operatorname{tr}(\mathbf{I})}, \tag{2.3}$$

which quantifies the spectral concentration in the most sensitive direction. Higher values indicate stronger anisotropy in the gradient distribution (ACHILLE et al., 2019; FEATHER et al., 2025).

2. **Shannon Effective Rank (erank)**

$$\operatorname{erank} = \exp\left(-\sum_i p_i \log p_i\right), \tag{2.4}$$

where $p_i = \lambda_i / \operatorname{tr}(\mathbf{I})$ (ROY; VETTERLI, 2007). This measures the number of eigendirections that contribute substantially, providing a soft measure of dimensionality.

3. **Trace**

$$\operatorname{tr}(\mathbf{I}) = \sum_i \lambda_i = \mathbb{E}\left[\|\nabla_z \mathcal{L}(z)\|^2\right]. \tag{2.5}$$

The trace reflects the total gradient energy, representing the overall sensitivity of the loss to perturbations in representation space.

2.6.4 Practical Computation

In practice, the empirical Fisher matrix is estimated by sample averaging over a dataset $\mathcal{D} = \{(z_n, y_n)\}_{n=1}^{N}$:

$$\hat{\mathbf{I}} = \frac{1}{N} \sum_{n=1}^{N} \nabla_z \mathcal{L}(z_n)\, \nabla_z \mathcal{L}(z_n)^\top, \tag{2.5}$$

where $\nabla_z \mathcal{L}(z_n)$ denotes the gradient of the cross-entropy loss with respect to the embedding z_n for the true label y_n.

For computational efficiency, two properties are exploited.

Algorithm 1 Empirical Fisher Spectrum Computation for Representation Space (adapted)

Require: Embeddings $Z \in \mathbb{R}^{N \times d}$, labels $Y \in \{1, \dots, C\}^N$, trained linear probe f, micro-batch size B, $k_{\max}$
Ensure: Trace τ, top-k_{eff} eigenvalues $\lambda_1 \geq \cdots \geq \lambda_{k_{\text{eff}}}$, E_1, truncated erank

$\tau \leftarrow 0$
Initialize randomized SVD state for rank $k_{\max}$ on implicit operator G ▷ No need to materialize $\hat{\mathbf{I}}$
for $i = 1$ to N in steps of B **do**
    $z_{\text{batch}} \leftarrow Z[i : i + B], \quad y_{\text{batch}} \leftarrow Y[i : i + B]$
    $g_{\text{batch}} \leftarrow \nabla_z \mathcal{L}(z_{\text{batch}}, y_{\text{batch}})$ ▷ $g_{\text{batch}} \in \mathbb{R}^{B \times d}$
    $\tau \leftarrow \tau + \sum_{b=1}^{B} \|g_{\text{batch}}^{(b)}\|^2$
    Update randomized SVD with rows g_{batch} ▷ Streaming / block update
end for
$\tau \leftarrow \tau / N$ ▷ Eq. (2.6)
$k_{\text{eff}} \leftarrow \min(k_{\max}, N, d)$
Obtain top singular values $\sigma_1 \geq \cdots \geq \sigma_{k_{\text{eff}}}$ from randomized SVD of G
$\lambda_i \leftarrow \sigma_i^2 / N$ for $i = 1, \dots, k_{\text{eff}}$ ▷ since $\hat{\mathbf{I}} = \frac{1}{N} G^\top G$
$E_1 \leftarrow \lambda_1 / \tau$
$s \leftarrow \sum_{i=1}^{k_{\text{eff}}} \lambda_i$
$p_i \leftarrow \lambda_i / s$ for $i = 1, \dots, k_{\text{eff}}$
erank $\leftarrow \exp\Big(-\sum_{i=1}^{k_{\text{eff}}} p_i \log p_i\Big)$

First, the trace of the empirical Fisher can be computed directly from the gradient vectors without explicit matrix formation:

$$\operatorname{tr}(\hat{\mathbf{I}}) = \frac{1}{N} \sum_{n=1}^{N} \|\nabla_z \mathcal{L}(z_n)\|^2, \tag{2.6}$$

which avoids constructing the full $d \times d$ matrix.

Second, only the leading part of the spectrum is estimated. Let $G \in \mathbb{R}^{N \times d}$ be the gradient matrix whose n-th row is $\nabla_z \mathcal{L}(z_n)$. By construction,

$$\hat{\mathbf{I}} = \frac{1}{N} G^\top G.$$

Therefore, the eigenvalues of $\hat{\mathbf{I}}$ are given by $\lambda_i = \sigma_i^2 / N$, where $\{\sigma_i\}$ are the singular values of G. The top-k eigenvalues are obtained via randomized singular value decomposition applied to G (HALKO; MARTINSSON; TROPP, 2011), reducing computational complexity from cubic in d to $O(Ndk)$.

To accommodate memory constraints, gradients are accumulated in microbatches. Let $k_{\text{eff}} = \min(k_{\max}, N, d)$ denote the effective rank used in the truncated spectral analysis.

All computations are performed on frozen embeddings using a trained linear probe, ensuring tractability for large N and d (Algorithm 1) (HALKO; MARTINSSON; TROPP, 2011; DELATTRE; KUHN, 2023).

2.7 Synthesis

The theoretical framework developed in this chapter centers on a core premise: visually apparent attention-related configurations can be analyzed as geometric and statistical properties of embedding spaces learned by pretrained visual encoders. Vision Transformers and convolutional networks provide the representational substrate; latent diffusion models generate systematically varied visual stimuli, with text conditioning supplied by a large language model; and vision–language models serve as auxiliary tools for labeling and organizing synthetic data within the generation pipeline.

Taken together, these components define a unified conceptual framework in which pretrained visual embeddings are treated as geometric objects, synthetic data as a controlled source of variation, and supervision as a mechanism that induces structure within representation space. This synthesis provides the conceptual basis for the analyses developed in the following chapters.

3 Related Work

This chapter reviews prior work relevant to the present study, organized around five intersecting lines of inquiry: (i) attention and engagement analysis in educational settings, (ii) visual proxies and behavioral cues for attentional states, (iii) representation learning and probing of pretrained visual models, (iv) the use of synthetic data and generative models as experimental instruments, and (v) the use of Vision–Language Models as evaluators. The emphasis is on positioning the present work within existing methodological and conceptual trajectories, highlighting both continuities and deliberate departures.

3.1 Attention and Engagement Analysis in Educational Contexts

Automated analysis of attention and engagement has been investigated across multiple modalities, including physiological signals, eye tracking, and vision-based behavioral cues. Early approaches relied on electroencephalography (EEG) to derive engagement indices from neural activity (BERKA et al., 2007), as well as eye-tracking techniques to infer attentional allocation in learning contexts (GOG; SCHEITER, 2010).

Physiological sensors do not provide direct measures of these latent cognitive states, but offer indirect and instrumented proxies for attentional and engagement-related processes. Examples include heart sensors, which monitor rate variability, and electroencephalography devices, which record neural activity, or electrodermal activity (EDA) as a "sensitive index of sympathetic nervous system activity" (POH et al., 2010).

Multimodal approaches combine video, audio, and sensor data to construct a more comprehensive profile of student engagement (D'MELLO; GRAESSER, 2012). This integration, while might enable more accurate analyses, requires advanced data fusion algorithms and robust technological infrastructure.

While these methods offer high temporal resolution, their intrusiveness, concerns related to privacy and operational constraints, such as high cost, limit applicability in real classrooms.

Computer Vision based approaches have therefore gained prominence, using facial expressions, head pose, gaze direction, and body posture as observable correlates of engagement (WHITEHILL et al., 2014; BOSCH, 2016). Video analysis has become an effective approach for assessing attention and engagement in classroom settings. Such systems analyze visual data captured by cameras, predominantly relying on Convolutional Neural Networks based architectures at the time to detect relevant patterns (NEZAMI et al., 2019). In intelligent tutoring systems, (GRAFSGAARD et al., 2013) demonstrated

that facial expressions and head movements can be predictive of student frustration and engagement.

In the context of distance learning, (SHARMA et al., 2019) combine information about the movements of the eyes and head, and facial emotions to produce a concentration index with classes of engagement.

Many existing systems typically frame the task as supervised classification, training models end-to-end to predict engagement labels. However, they remain heavily dependent on annotated datasets, which are scarce, demographically biased, and legally constrained, especially when minors are involved.

In contrast to end-to-end supervised engagement detectors, a line of work has emerged that investigates pretrained visual representations as objects of analysis in their own right, emphasizing linear accessibility, geometric structure, and task-induced organization over predictive performance (ALAIN; BENGIO, 2017; RAGHU et al., 2017; KORNBLITH et al., 2019; ACHILLE et al., 2019).

3.2 Visual Proxies for Attention and Behavioral Cues

A substantial body of work has established correlations between attentional states and observable visual cues. Head orientation and gaze direction are among the most widely used indicators, often estimated via facial landmarks or 3D head pose models (MURPHY-CHUTORIAN; TRIVEDI, 2009; PATACCHIOLA; CANGELOSI, 2017; ZHANG et al., 2015). Postural cues, such as leaning forward, slouching, or resting the head on the desk, have also been associated with engagement and disengagement (D'MELLO; GRAESSER, 2012; SHARMA et al., 2019).

In affective computing and Facial Expression Recognition (FER), researchers have long investigated how facial muscle configurations relate to emotional and attentional states (EKMAN; FRIESEN, 1971; EKMAN; FRIESEN, 1978).

(MOLLAHOSSEINI et al., 2017) emphasize that most approaches in this context are based on supervised learning techniques, which require annotated image samples for training. To meet this demand, researchers have developed databases with images of individuals or actors expressing basic emotions (LYONS et al., 1998; TIAN; KANADE; COHN, 2001; LUCEY et al., 2010; PANTIC et al., 2005; GROSS et al., 2010, *apud* (MOLLAHOSSEINI et al., 2017)).

Studies indicate that posed expressions differ from spontaneous expressions in aspects such as configuration, intensity, and duration (COHN; SCHMIDT, 2004; VALSTAR et al., 2006, *ibid.*). In response to this limitation, some researchers collected natural expressions while participants watched videos, performed emotion-induction tasks in

laboratories, or interacted with computerized tutoring systems (MAVADATI et al., 2013; MCDUFF et al., 2013; SNEDDON et al., 2012; GRAFSGAARD et al., 2013, *ibid.*). Although these methods allow the capture of a large number of frames, dataset diversity remains constrained by factors such as the number of participants, head position, and environmental conditions.

Given these conditions, (MOLLAHOSSEINI et al., 2017) highlight that interest has recently increased in datasets documenting facial expressions and emotional states in real-world situations, captured from movies or the internet. These datasets are annotated using different models, such as the categorical model (DHALL et al., 2013; GOODFELLOW et al., 2015; MOLLAHOSSEINI et al., 2016, *ibid.*), the dimensional model (ZAFEIRIOU et al., 2016, *ibid.*), or the Facial Action Coding System (FACS) (BENITEZ-QUIROZ; SRINIVASAN; MARTINEZ, 2016, *ibid.*).

Nevertheless, they still present limitations, such as restricted coverage to a single affect model, a reduced number of participants, or scarcity of samples for certain emotions, such as disgust. In the context of the cited work, the authors refer to it as one among several models listed in the literature to quantify affective facial behaviors, including the categorical model, which comprises a discrete list such as the six basic emotions defined by (EKMAN; FRIESEN, 1971).

Despite these advances, two limitations persist. First, most datasets exhibit restricted demographic diversity and environmental variability. Second, many approaches implicitly assume a relatively direct mapping between visual cues and internal cognitive states.

3.3 Representation Learning and Probing of Pretrained Visual Models

The emergence of large-scale pretrained models has shifted Computer Vision from task-specific pipelines to representation-centric paradigms. Vision Transformers (ViT) (DOSOVITSKIY et al., 2020), contrastive multimodal models such as CLIP (RADFORD et al., 2021), and self-supervised frameworks such as DINO (CARON et al., 2021) and DINOv3 (SIMÉONI et al., 2025) learn high-dimensional embedding spaces in which semantic regularities seem to be implicitly encoded.

A growing body of work investigates how such representations can be probed using linear classifiers, clustering, or low-capacity adapters to assess whether specific concepts are linearly accessible (ALAIN; BENGIO, 2017; HE et al., 2020). This probing paradigm is well established in NLP and has been increasingly adopted in vision to analyze object parts, spatial layout, and semantic attributes (BAU et al., 2017; RAGHU et al., 2017).

From an information-theoretic and geometric perspective, (ACHILLE; SOATTO, 2018) proposed analyzing representations through the lens of information bottleneck theory, while Amari (AMARI, 1998) introduced the natural gradient and highlighted the role of Fisher Information as a Riemannian metric on parameter and representation spaces. Recent work has applied Fisher-based analysis to characterize task-induced geometry and anisotropy in learned representations (PAPYAN; HAN; DONOHO, 2020; SAGUN et al., 2018).

3.4 Vision–Language Models in Affective and Expression Recognition

The emergence of Vision–Language Models (VLMs), and in particular CLIP, has prompted renewed interest in their applicability to affective computing tasks, including sentiment analysis, emotion recognition, and facial expression analysis. Early investigations explored whether the semantic alignment induced by contrastive language–image pretraining could be leveraged beyond zero-shot classification, serving as a foundation for affect-related visual understanding.

(BUSTOS et al., 2023) conducted a systematic study on the use of CLIP for visual sentiment analysis, examining how sentiment-related structure manifests in the CLIP embedding space. Their work demonstrated that CLIP representations encode non-trivial affective signals and can be adapted to sentiment tasks with limited supervision. Importantly, the study highlighted both the potential and the limitations of CLIP when applied to affective domains, emphasizing that performance gains often require task-specific adaptation of the embedding space.

More recently, Vision–Language Models have been actively explored in competitive affective computing benchmarks, most notably within the 8th Workshop and Competition on Affective & Behavior Analysis in-the-Wild (ABAW), held in conjunction with the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) in 2025. At least two high-ranking submissions (second and forth places on the leader board) in the Expression Recognition (EXPR) challenge incorporated CLIP-based representations as core components of their pipelines.

(ZHOU; LING; CAI, 2025) proposed a framework combining CLIP image embeddings with sequential learning modules to model temporal dynamics in facial expression recognition. In their approach, CLIP serves as a pretrained visual backbone whose representations are further adapted through supervised learning, with temporal aggregation enabling improved performance on in-the-wild video data.

Similarly, (CABACAS-MASO et al., 2025) reported competitive performance in

the 8th ABAW challenge by integrating CLIP embeddings into a Dual-Direction Attention Mixed Feature Network (DDAMFN). Their work combined attention-based feature extraction with CLIP representations, demonstrating that Vision–Language pretraining can complement task-specific architectural inductive biases in facial expression recognition.

Collectively, these works illustrate a growing trend in which Vision–Language Models are employed as powerful pretrained backbones for affective recognition tasks. In this literature, CLIP representations are typically fine-tuned or augmented with additional layers, and success is evaluated primarily through task-level metrics such as F1 score or Concordance Correlation Coefficient. The central objective remains end-to-end optimization for benchmark performance.

While these studies were instrumental in motivating the present work—both by demonstrating the practical utility of CLIP-based representations and by highlighting their competitive performance in challenging in-the-wild settings—they largely treat pretrained representations as intermediate substrates to be optimized, but perhaps less as objects of analysis in their own right. Questions concerning the internal geometry, dimensionality, and stability of these representations prior to task-specific adaptation remain underexplored.

3.5 Synthetic Data and Generative Models as Experimental Instruments

3.5.1 Definition and Conventional Use of Synthetic Data

In the context of artificial intelligence, synthetic data are commonly defined as data generated from real data while preserving key statistical properties of the originals (EMAM, 2020). The stated objective is that analyses conducted on such data yield results comparable to those obtained with real data, with utility serving as the primary criterion for evaluating their quality.

Within this dominant framing, synthetic data are treated primarily as substitutes for real-world observations, intended to approximate empirical distributions when direct data collection is infeasible or costly. Fidelity to the original data-generating process is therefore regarded as the central measure of success.

3.5.2 Synthetic Data as an Enabling Resource

The same report argues that data synthesis can address practical constraints in AI and machine learning projects by improving data access and quality, thereby enabling studies that would otherwise be infeasible or significantly delayed. Access to data is characterized as essential not only for model training and validation, but also for evaluating

third-party technologies and testing downstream applications. From this perspective, synthetic data is positioned as an enabling resource and not merely a convenience.

This enabling role is particularly salient in settings where data access itself constitutes the primary bottleneck. In such cases, the utility of synthetic data lies less in realism than in controllability, scalability, and repeatability.

3.5.3 Legal and Regulatory Constraints

Access to real-world data for secondary analysis has become increasingly constrained by legal, institutional, and regulatory factors. Public concerns about personal data usage have intensified, and privacy regulations such as the Health Insurance Portability and Accountability Act in the United States and the General Data Protection Regulation in Europe impose strict limitations on data reuse beyond original collection purposes. These constraints often require additional consent and can introduce systematic biases, as individuals who opt in and those who decline tend to differ in relevant characteristics. As a result, data availability and representativeness are frequently compromised in practice (EMAM, 2020).

Such restrictions apply with even greater force to underage populations. With respect to the right to image of minors, Brazilian law affords reinforced protection through the Statute of the Child and Adolescent (ECA) and the General Data Protection Law (LGPD). As stated in guidance from the Brazilian Bar Association, special care is required when recording, publishing, and storing photographs, videos, and other personal data of children and adolescents, making explicit parental or guardian authorization indispensable (Ordem dos Advogados do Brasil, Seccional do Rio Grande do Norte, 2023).

Brazilian Law No. 13.709 of August 14, 2018 (LGPD) (BRASIL, 2018) further establishes strict requirements for the processing of personal and sensitive data, including specific provisions governing data related to children and adolescents. These legal constraints substantially limit the feasibility of collecting and redistributing real-world visual datasets in educational contexts.

Within this regulatory landscape, synthetic data offers the possibility of exploring diverse visual scenarios without incurring the legal and institutional barriers inherent to real-world data collection, albeit within the limitations imposed by available generative models and computational resources.

3.5.4 Data Scarcity and Rare Phenomena

A similar constraint arises in domains where the phenomena of interest are inherently rare or legally difficult to access. In the context of traffic accident analysis using CCTV footage, (SHAH et al., 2018) report that although camera-based monitoring is a promising

approach, the construction of large-scale datasets is severely limited by two factors: the low frequency of accidents, which renders passive data collection impractical, and legal restrictions on access to surveillance imagery.

Even in scenarios where sensing infrastructure is ubiquitous, usable data remains scarce. In such cases, synthetic data functions less as a replacement for real observations and more as a means of inducing controlled instances of rare or critical configurations for analysis and evaluation.

3.5.5 Biases and Dataset Composition

With respect to data quality, the literature notes that limited access to proprietary datasets often leads researchers to rely on public or open-source collections (EMAM, 2020). While such datasets can serve as useful starting points, they frequently lack the diversity and task-specific structure required by particular applications. Insufficient heterogeneity can compromise robustness, particularly in the representation of rare or marginal cases.

In facial analysis, (BUOLAMWINI; GEBRU, 2018) demonstrated that widely used benchmarks exhibit severe demographic imbalances, with a predominance of light-skinned individuals. By constructing a dataset balanced across gender and skin type, they showed that classification error rates varied markedly across phenotypic subgroups.

3.5.6 Advances in Generative Models for Image Synthesis

Synthetic data has long been used in Computer Vision, but the advent of latent diffusion models has substantially lowered the barrier to generating large volumes of high-fidelity images. (ROMBACH et al., 2022) introduced Latent Diffusion Models, which enable efficient image synthesis by operating in a compressed latent space. These models underpin widely used systems such as Stable Diffusion, making controlled image generation accessible at scale.

The availability of such models enables systematic manipulation of visual attributes through prompt design and stochastic variation, facilitating controlled experimentation that would be difficult to replicate with real-world data.

3.5.7 Epistemological Position: Synthetic Data as Instrument

Synthetic data has been widely advocated as a response to data scarcity, privacy constraints, and annotation costs. In sensitive domains such as healthcare, surveillance, and facial analysis, it is often presented as a practical means of bypassing legal and logistical barriers while approximating real-world distributions. Within this framing, fidelity to real data remains the primary objective.

An alternative epistemological stance treats synthetic images not as surrogates for natural data, but as experimental perturbations that enable the controlled induction of specific visual configurations.

This instrumental perspective aligns with recent calls to move beyond i.i.d. benchmarking and to systematically audit learned representations through controlled perturbations and counterfactual inputs, in order to expose shortcut learning and unintended decision rules (GEIRHOS et al., 2020). By systematically varying prompts and random seeds, synthetic data enables controlled exploration of sensitivity, invariances, and failure modes in pretrained embedding spaces—phenomena that are difficult or impractical to isolate using real-world data alone.

3.6 Vision–Language Models as Evaluators

Many real-world datasets are weakly labeled or entirely unlabeled. Manual annotation at scale is costly, time-consuming, and error-prone, which constrains the feasibility of supervised approaches. In this context, the generation of labeled synthetic data is commonly presented as a means of accelerating model development while enabling some control over annotation quality.

Vision–Language Models (VLMs), such as CLIP (RADFORD et al., 2021), LLaVA (LIU et al., 2023), and QwenVL (BAI et al., 2025), integrate visual encoders with large language models to enable joint aligned cross-modal representations and language-conditioned processing of visual inputs. These models have been used for image captioning, visual question answering, and zero-shot classification.

Recent work has explored using vision–language models as sources of weak supervision, dataset filters, or auxiliary validators (RADFORD et al., 2021). While such models can provide scalable semantic judgments, they are known to exhibit biases, overgeneralization, and sensitivity to prompt phrasing and spurious correlations (GEIRHOS et al., 2020; ZHOU et al., 2022). As such, their outputs invite systematic investigation as empirical artifacts for understanding model behavior.

3.7 Summary of Research Trajectories

The reviewed literature reveals a fragmentation across three major research trajectories. First, the educational data mining and affective computing communities have developed increasingly sophisticated pipelines for detecting engagement and attention, typically framed as supervised prediction problems optimized for accuracy on task-specific datasets (WHITEHILL et al., 2014; BOSCH, 2016; SHARMA et al., 2019). Second, the representation learning community has focused on the structural properties of pretrained

models, analyzing linear separability, clustering behavior, and geometric characteristics of embedding spaces, often in abstraction from concrete application domains (ALAIN; BENGIO, 2017; ACHILLE; SOATTO, 2018). Third, the synthetic data and generative modeling literature has emphasized data augmentation, privacy preservation, and dataset scaling, usually treating synthetic samples as surrogates for real data (ROMBACH et al., 2022; EMAM et al., 2020).

Part II

Proposed Approach

4 Methodology

The study of representation learning in contemporary visual models presupposes that the structure and interpretability of learned embeddings are shaped by the input signals presented to the model. In the context of this dissertation, access to suitable real-world imagery is subject to practical and methodological constraints, motivating the construction of a synthetic dataset as a controlled experimental instrument. Accordingly, this chapter details the methodology adopted for dataset construction, and describes the experimental protocols devised to probe visual representations across different learning regimes using lightweight adapters, and geometry-aware analyses.

4.1 Proposed Dataset

4.1.1 *Synthetic Data* as a Lens into Representation Spaces

Synthetic data is introduced not as a substitute for real-world observations, nor as a means of improving classifier performance, but as an experimental lens into pre-trained representation spaces. By enabling deliberate and repeatable variation of visual factors—such as posture, gaze direction, object interaction —synthetic images make it possible to probe how embeddings respond to structured semantic change.

In this role, generative models act as instruments for rendering representational sensitivities observable, rather than data factories. The emphasis is not on photorealism, but on controllability: the ability to induce specific visual configurations while limiting uncontrolled sources of variation. This controlled exposure allows the organization of representation space to be inspected under conditions that are difficult or impractical to obtain from real-world data alone.

Despite their practical effectiveness, the resulting samples are not guaranteed to be artifact-free and may exhibit generative imperfections such as missing or extra limbs, distorted hands, inconsistent geometry, or other local anomalies. From a methodological standpoint, their relevance to this work lies not in their generative novelty nor in their raw expressive capacity, but in their utility as controlled experimental instruments.

From an epistemological perspective, synthetic data generation via Latent Diffusion Models functions as a form of *experimental perturbation* applied to the visual input space. If task-relevant signals—operationally defined through Vision–Language Model supervision—are encoded in the embedding space of a Vision Foundation Model, these signals should manifest as stable geometric structure across variations induced by prompt semantics and random seeds.

4.1.2 *Scope Delimitation*: Biases and Demographic Factors

It is important to delimit the scope of the present work with respect to demographic biases and dataset balance. Although prior research has demonstrated that imbalances along axes such as ethnicity, gender, and skin tone can materially affect model behavior (BUOLAMWINI; GEBRU, 2018), the present study does not attempt to explicitly control, correct, or audit such biases through systematic demographic specification.

In particular, the synthetic image generation process does not enforce demographic balance through explicit ethnicity or related attribute specifiers in prompt design. Variation in appearance arises implicitly through stochastic factors inherent to the generative process, such as random seeds and latent sampling, not from controlled demographic constraints.

This design choice reflects the objectives of the work. The primary focus is not to evaluate fairness, demographic parity, or bias mitigation strategies, but to investigate how pretrained visual representations respond to controlled variations associated with apparent attentional configurations. Introducing explicit demographic controls would conflate this objective with additional axes of variation, complicating interpretation without directly serving the representational questions under study.

Accordingly, any demographic imbalances that may arise in the synthetic data are treated as secondary effects, not as variables of interest. The present work therefore makes no claims regarding fairness, bias reduction, or demographic representativeness.

4.1.3 Dataset Generation Pipeline Overview

The dataset design is guided by three constraints. First, the dataset must reflect a visually consistent, coherent domain, in order to reduce spurious variability unrelated to the phenomenon under investigation. Second, it must allow controlled variation of behavioral proxies commonly associated with apparent attention and inattention. Third, it must remain feasible under limited computational resources, precluding exhaustive combinatorial generation.

Synthetic image generation is adopted to address these constraints. By inducing visual attributes through prompt-based generation, synthetic data enables exploration of how pretrained visual embeddings respond to variations in posture, gaze, object interaction, and classroom context, and ultimately, to what extent they already contain signal relevant to abstract behavioral concepts, such as apparent attention, without task-specific fine-tuning.

The proposed dataset is constructed through a multi-stage pipeline designed to produce these controlled visual representations within the scope of this work. The pipeline integrates hypothesis-driven prompt generation, stochastic image realization, explicit curation, and independent semantic validation into a controlled experimental instrument

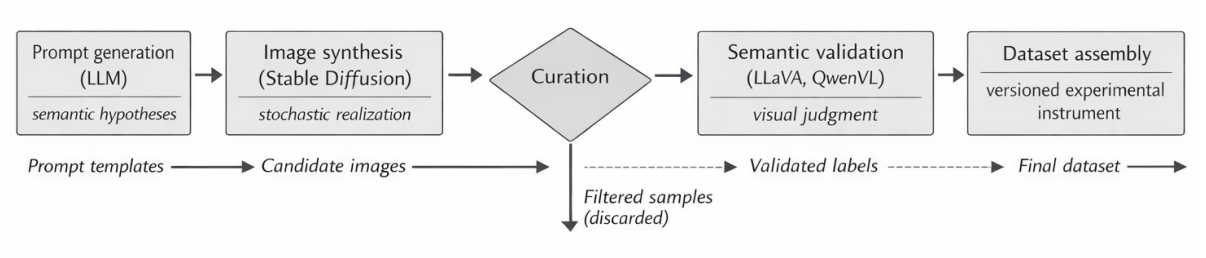


Figure 1 – Dataset Generation Pipeline

for probing task-aligned structure in pretrained visual representations, as illustrated by Figure 1.

Image synthesis is performed using a Stable Diffusion model (PODELL et al., 2023), a latent diffusion architecture for text-to-image generation conditioned on natural language prompts. In this pipeline, prompts are explicitly designed and refined as controlled generative parameters. Prompt formulations were developed through an iterative experimental process inspired by (TONG; JONES; STEINHARDT, 2024), in which a large language model was instructed to generate plausible, semantically grounded descriptions of students exhibiting visually apparent attentive and inattentive cognitive states.

Given well-documented limitations of text-to-image generation—most notably semantic drift and inconsistencies between prompt intent and visual realization—an external validation layer is introduced. Vision–Language Models are employed as judges to assign labels to the generated images, decoupling image synthesis from semantic verification. This design enforces an additional level of control over dataset integrity and seeks to ensure that label assignments reflect observable visual evidence, albeit imperfectly, and not only generative intent.

4.1.4 Prompt Taxonomy and Behavioral Proxies

The dataset is organized around a taxonomy of behavioral proxies intended to capture visually apparent manifestations of attention and inattention in classroom scenes. These proxies are neither assumed to be exhaustive nor mutually exclusive in real-world settings; they are selected because they admit consistent visual operationalization, can be reliably instantiated in static image representations, and have been previously treated as observable correlates of engagement or distraction.

After extensive experimentation, a fixed prompt template was adopted:

> *"A DSLR highly realistic photography of an elementary school student {behavior} in class."*

This template anchors the visual style, domain, and level of realism, while allowing

the behavioral component to vary, in accordance with the first and second design constraints. Fixing the non-behavioral elements of the prompt serves to reduce unintended stylistic or contextual drift across generated images, thereby isolating the effect of behavioral variation on downstream representations. The term "elementary school student" is used as a generative prior to condition visual appearance and context; the dataset targets visually plausible classroom representations, following the underlying model distribution.

Two broad categories are defined: *attentive* and *inattentive*. The dataset relies on natural-language behavioral descriptors embedded in the fixed prompt template, composed of a discrete set, in accordance with the third design constraint, instead of decomposing these categories into explicit behavioral dimensions. Each descriptor encodes a recognizable behavioral situation, which implicitly bundles multiple visual cues—such as posture, gaze, and object interaction.

Examples of attentive proxies include, but are not limited to: absorbed, actively listening, alert, concentrated, curious, disciplined. Analogous proxies are defined for apparent inattentiveness: daydreaming, interacting with unauthorized objects (e.g., mobile phones), yawning, slouching, resting the head on the desk, turning away from the instructional focus.

For completeness and reproducibility, the full taxonomy of behavioral proxies is provided in the appendixes.

4.1.5 Image Generation

As mentioned, all images in the dataset are generated using a latent diffusion model from the Stable Diffusion XL (SDXL) family, which was selected as a compromise between visual fidelity and computational feasibility. While more recent and larger generative models offer improved realism, their memory and compute requirements exceed the resources available in the present experimental setting (10 GB VRAM class devices).

Within these constraints, SDXL provides a robust and reproducible synthesis framework, is widely adopted in the research community, and supports controlled prompt-based generation with sufficient semantic stability for the purposes of this study. Importantly, the objective of the present work is not to maximize photorealism per se, but to generate semantically coherent visual instances suitable for probing the structure of pretrained visual representations.

To sample intra-concept variability, the template prompt is rendered using a fixed range of random seeds, from seed 1 to seed 100. Seeds are treated as a mechanism for stochastic variation within a semantic prompt. As such, multiple images generated from the same prompt but different seeds are considered samples from a local neighborhood of the same induced visual concept.

Image synthesis is performed at the native resolution of 1024×1024, as imposed by the architectural characteristics of the diffusion model; generation at lower resolutions induces structural degradation and semantic instability. All images are subsequently downsampled to match the input resolution of the vision backbone employed in the representation analysis.

Parameters particular to the diffusion model, such as resolution, guidance scale, number of inference steps, and scheduler configuration are held constant across all generations.

4.1.6 Automated Labeling with Vision–Language Models

Vision–Language Models are employed strictly as auxiliary instruments, serving both as organizational mechanisms for synthetic data and as indirect probes of how visual representations (obtained from Visual Transformers) interact with language-conditioned supervision (Large Language Models). Their aggregated responses are used to operationally assign synthetic images to class-specific directories based on vote consistency, thereby defining supervised contrasts for downstream analysis. They are not treated as sources of semantic ground truth, nor as authoritative claims about the true attentional state depicted in an image. Their role is procedural in nature and does not carry epistemic weight.

Concretely, each synthetically generated image is submitted to two distinct Vision–Language Models, which serve as the operational reference for instantiating the attentional categories used in this study. Using a fixed prompt that queries whether the depicted subject appears attentive or inattentive—together with its inverse formulation to mitigate prompt-induced bias—the VLMs provide scalable assessments that are pooled to assign each image to a corresponding class directory, with subdirectories reflecting the degree of intra and inter-model agreement (net votes). This procedure reflects a central methodological premise of the work: because outputs of latent diffusion models are not assumed to be semantically faithful to their prompts, Vision–Language Models are required to mediate between image generation and downstream analysis, ensuring that subsequent probing of representation geometry is grounded in a stable, operationally defined supervision.

Consider the images shown in Figures 2 and 3. Although both are generated from the same behavioral prompt, absorbed, defined [1] as “having the mind fixed on something,” the resulting images differ markedly in posture and gaze configuration. This variability reflects the stochastic nature of latent diffusion models, in which a single semantic prompt gives rise to a distribution of distinct visual realizations. Correspondingly, these visually dissimilar instances elicit divergent assessments from Vision–Language Models, as shown in

[1] <https://www.merriam-webster.com/thesaurus/absorbed>

Listings 4 and 5. Taken together, this illustrates both the nontrivial visual spread induced by prompt–seed combinations and the resulting ambiguity in operational supervision, motivating the aggregation of VLM judgments.



Figure 2 – Image for the behavior "absorbed", labeled as inconclusive by VLMs.



Figure 3 – Image for the behavior "absorbed", labeled as attentive by VLMs.

Each image is evaluated independently by these models using paired, order-inverted prompts querying apparent attentiveness (e.g., "Is the foreground student attentive or inattentive? Why?" and "Is the foreground student inattentive or attentive? Why?"). This inversion is deliberately employed to counter known ordering and prior biases in Vision–Language Models, which operate as statistical approximations over visual patches and text tokens, but cannot be considered unbiased visual judges.

The inclusion of the prompt qualifier "why" serves a methodological and auditing function. By requesting a justification alongside the binary judgment, the Vision–Language Models are encouraged to condition their responses on explicit visual attributes present in the image, such as gaze direction, posture, or object interaction. This way, we seek to avoid label tokens driven primarily by prior probabilities or prompt-induced biases. This framing increases the likelihood that the model's output reflects engagement with the visual content, as opposed to a shallow or default class preference.

From an auditing perspective, the generated rationales provide an additional signal for assessing response stability under prompt inversion and across model instances. While these explanations are not treated as faithful causal accounts or as evidence of human-like reasoning, they enable qualitative and quantitative consistency checks that help identify prompt-order effects, semantic shortcuts, or systematic biases in the labeling process.

The appendixes contain samples of votes, their aggregation, and the generated

rationales.

4.1.7 Construction of Train, Validation, and Test Splits

To ensure a controlled and auditable evaluation of representation quality, the dataset was partitioned into training, validation, and test splits according to a nested and repetition-aware protocol. The primary objective of this procedure is to isolate representation effects from sampling artifacts, data leakage, and accidental overlap between experimental conditions.

4.1.7.1 Fixed test split

A held-out test split per repetition is constructed once and kept strictly fixed across all experiments, resulting in a set of five distinct test sets. Images assigned to the test split are never reused in training or validation, regardless of training set size or repetition index. This design ensures that all reported generalization results are evaluated on the same data distribution, enabling fair comparison across models, probes, and dataset scales.

Each test split is composed of 480 images, equally distributed among the classes 'attentive' and 'inattentive'. In the largest train set, composed of 1,600 images per class, the test size amounts to 15 % of the entire set.

4.1.7.2 Nested training sets

Training splits are built in a nested fashion. Let

$$N_{\text{per}} \in \{50, 100, 150, 200, 300, 400, 800, 1200, 1600\}$$

denote the number of samples per class used for training. For each successive value of N_{per}, the training set strictly contains all samples from the smaller configuration, plus additional samples drawn without replacement. Formally, if $\mathcal{T}_N$ denotes the training set for size N, then

$$\mathcal{T}_{50} \subset \mathcal{T}_{100} \subset \mathcal{T}_{150} \subset \cdots \subset \mathcal{T}_{1600}.$$

Any reference in this work to N_{per} (e.g., $N_{\text{per}} = 50$) denotes the number of samples drawn *per class*. Since all splits are constructed to be class-balanced, the total number of images in the corresponding split is therefore $2 \times N_{\text{per}}$. As an illustration, when $N_{\text{per}} = 50$, the training split contains 70 images in total, comprising 70% of the data. In the same configuration, the validation split comprises 15% of the data, rounded up to the nearest integer, yielding 16 images. When $N_{\text{per}} = 1,600$, the training split contains 2,240 images in total, comprising 70% of the data; the validation split comprises 15% of the data, yielding 480 images.

The size of the test split constitutes an explicit exception to the split rule (70/15/15): it is fixed once and remains constant across all values of N_{per} and across all experimental repetitions.

This nested structure allows learning curves and geometric diagnostics to be interpreted as functions of sample size, instead of changing data composition.

4.1.7.3 Validation split construction

For each training configuration, a validation split, representing a proportion of 15 %, as mentioned, is drawn disjointly from both the training and test sets. The validation split is used exclusively for model selection and diagnostic analysis and is never merged with the training data. Validation sizes are kept proportional across configurations to avoid introducing scale-dependent bias.

4.1.7.4 Repetitions and seed control

To account for sampling variability, the entire splitting procedure is repeated multiple times using distinct random seeds. Each repetition generates an independent but internally consistent set of training and validation splits, while preserving the same fixed test split per repetition. Seeds are explicitly recorded and treated as part of the experimental state, enabling exact reproducibility.

4.1.7.5 Class balance and stratification

All splits are constructed to be class-balanced by design, to the allowed extent by the underlying distribution. For each value of N_{per}, the same number of samples is drawn from each class, ensuring that performance differences cannot be attributed to class imbalance, within the aforementioned constraint.

4.1.7.6 Leakage prevention and auditing

After split construction, automatic audits are performed to verify that no image appears in more than one split. In particular, intersections between training, validation, and test file identifiers are explicitly checked and enforced to be empty. Any violation immediately invalidates the split and halts the pipeline. This guarantees that all reported results are free from data leakage.

4.1.7.7 Stability and overlap diagnostics

In addition to basic leakage checks, overlap statistics across repetitions are computed and summarized. These diagnostics quantify how much of the training data is shared

between repetitions for a fixed N_{per}, providing transparency about sampling diversity and supporting robustness claims.

4.1.7.8 Rationale

This splitting protocol is intentionally conservative. By fixing the test set, nesting the training data, controlling randomness, and enforcing strict leakage checks, the experimental design ensures that observed trends in accuracy, Fisher-based metrics, or representation geometry reflect properties of the learned representations, not artifacts of data partitioning. As such, the dataset is treated not merely as a source of labels, but as an instrument for probing the structure of pretrained visual embeddings under controlled conditions.

4.1.7.9 Representative Samples for the Training Splits

In the appendix (Table 3, Table 4, Table 5, Table 6), we present the training splits for the four values of N that capture the most informative moments of the learning curves:

- N = 50 — the extreme pure few-shot case. Only 60 activities covered, 101 zero-shot in the test set, with at most 2 examples per activity. This is the starting point that defines the problem.
- N = 200 — the inflection point. Coverage grows rapidly (128 activities), and zero-shot cases drop to 32.
- N = 400 — the knee of the curve. Coverage is nearly saturated (149 activities), zero-shot cases drop to 12, and F1 begins to stabilize. This is the point where the marginal return of additional data clearly starts to diminish.
- N = 1600 — the full saturation case. Coverage is completely stabilized (identical to N = 800 and N = 1200), with only 5 zero-shot activities. It demonstrates the ceiling imposed by the distribution.

Regarding repetitions, a single repetition (e.g., Rep 1) is sufficient, as the five repetitions are consistent with one another by construction; differences across repetitions are minimal and quantitative, not qualitative.

4.2 Representation-First Experimental Methodology

This work adopts a representation-first experimental strategy to assess whether task-relevant information about visually apparent attentional states is already encoded in pretrained visual representations. The methodology asks whether such structure is present, low-dimensional, and linearly accessible in the embedding spaces of contemporary vision

encoders. Emphasis is not placed on evaluating whether a model can be trained to perform attention classification.

This probing protocol is coupled with the instrumental use of synthetic data, enabling controlled variation of visual configurations and systematic interrogation of how such variations are reflected in embedding space.

4.2.1 Frozen Backbones and Lightweight Adapters

All experiments operate on frozen visual backbones. No gradients are propagated into the encoder during training. This design choice isolates the representational properties of the pretrained models from confounding effects introduced by end-to-end optimization.

On top of the frozen embeddings, only lightweight adapters are trained. These include:

- linear classifiers,
- shallow multilayer perceptrons with limited depth,
- linear discriminant analysis (LDA).

These adapters are intentionally low-capacity and are used as diagnostic instruments; we do not frame them as competitive architectures. Their role is to assess how easily the task can be read out from the representation, not to compensate for representational deficiencies.

4.2.2 Learning Curves and Sample Efficiency

To evaluate how efficiently task-relevant information is accessed from each representation, learning curves are constructed as a function of the number of training samples per class. Performance is measured on held-out test sets using Macro-F1 and ROC-AUC.

Learning curves serve two purposes:

1. They quantify sample efficiency, revealing how quickly a representation supports reliable discrimination.
2. They distinguish between representational limitations and head-capacity effects, as gains from increased data or model complexity can be directly observed.

Consistent experimental conditions are maintained across all backbones and adapter types to ensure that observed differences arise from the embeddings themselves.

4.2.3 Linear Discriminant Analysis as a Geometric Probe

Importantly, while Fisher's original discriminant and Rao's formulation are mathematically closely related, they differ in interpretation. Fisher's approach is fundamentally discriminative, focusing on variance ratios, whereas Rao's treatment admits a generative interpretation through likelihood-based classification under Gaussian class-conditional densities with equal covariances. The present study adopts LDA in the spirit of Fisher's original criterion, emphasizing linear separability in representation space and not strict adherence to generative distributional assumptions. Under this usage, LDA is employed as a linear probe on frozen embedding spaces, serving as an analytical tool for assessing whether class-relevant information is linearly accessible in the representations produced by pretrained visual encoders.

Linear Discriminant Analysis is employed not primarily as a classifier, but as a geometric probe of class structure in embedding space. Under its standard assumptions, LDA identifies directions that maximize between-class variance while minimizing within-class scatter.

Because LDA has no trainable parameters beyond class statistics, its performance directly reflects the geometric organization of the embeddings. Strong performance under LDA therefore provides evidence that the task is already linearly structured in the representation space.

4.2.4 Empirical Fisher Information and Task-Induced Geometry

Inspired by the framework of (AMARI, 2016), we compute the empirical Fisher Information Matrix (SCHRAUDOLPH, 2002) with respect to frozen representations, instead of model parameters. Let $z \in \mathbb{R}^d$ denote an embedding vector produced by a pretrained backbone, and let $\mathcal{L}(z)$ denote the cross-entropy loss of a trained linear probe. The representation-space empirical Fisher matrix is defined as (PASCANU; BENGIO, 2013; MARTENS, 2020)

$$\mathbf{I}(z) = \mathbb{E}_{\text{data}} \left[\nabla_z \mathcal{L}(z) \, \nabla_z \mathcal{L}(z)^\top \right], \tag{4.1}$$

where the expectation is taken over the empirical data distribution.

4.2.5 Why Empirical Fisher Instead of the Hessian

While the Hessian provides an exact local second-order characterization of curvature, it is generally indefinite for neural network objectives, yielding unbounded quadratic models and unreliable updates unless heavily damped. Consequently, positive semi-definite approximations that are more robustly estimable from data are often preferred (MARTENS, 2020). The empirical Fisher Information Matrix constitutes such an alternative: it estimates

curvature through outer products of first-order gradients and is positive semi-definite by construction.

For log-likelihood losses, it coincides with the Gauss-Newton approximation to the Hessian under standard assumptions (AMARI, 1998; MARTENS, 2020). However, the objective here is not second-order optimization but geometric analysis of pretrained representations. In this context, it also characterizes first-order sensitivity to perturbations in embedding space, making it a suitable diagnostic for probing anisotropy, effective dimensionality, and task-induced structure (ACHILLE et al., 2019; DELATTRE; KUHN, 2023).

4.2.6 Unifying View: Fisher Criterion, LDA, and Information Geometry

We apply Fisher-based analysis directly to embeddings produced by frozen large pretrained visual encoders, including self-supervised models and a supervised baseline. By computing the empirical Fisher with respect to embedding vectors, we treat pretrained representations as fixed geometric spaces and examine how task supervision modulates their local sensitivity structure without altering the underlying encoder parameters.

Fisher's linear discriminant, Linear Discriminant Analysis (LDA), and the Fisher Information Matrix can be unified under a geometric framework. Fisher's original criterion identifies directions in representation space that maximize class separability relative to within-class variability, defining a task-weighted notion of discriminability. LDA formalizes this via between-class and within-class scatter matrices, yielding a subspace in which class means are maximally separated under a generalized Rayleigh quotient.

In parallel, the empirical Fisher Information Matrix computed with respect to embeddings defines a local Riemannian metric that captures the sensitivity of the loss to infinitesimal perturbations in representation space. Thus, both LDA and Fisher-based analysis characterize supervision-induced geometry: LDA emphasizes global class structure via discriminative projections, while the Fisher matrix captures local anisotropy of the loss landscape through curvature.

Together, these tools offer complementary perspectives on representation quality. LDA assesses whether class-discriminative directions are globally coherent and linearly decodable, whereas Fisher-based metrics quantify how strongly task supervision modulates local sensitivity in the embedding space. This viewpoint treats pretrained representations not merely as feature vectors, but as geometric objects shaped by the structure of downstream tasks.

To characterize the geometry induced by the discrimination objective, the empirical Fisher Information Matrix (equivalently, a Gauss–Newton approximation of the Hessian) is computed with respect to the embedding vectors.

The spectrum of this matrix describes how loss sensitivity is distributed across directions in embedding space. In particular, the concentration of eigenvalues reflects the effective dimensionality of the *task-induced loss sensitivity* under the chosen probe and operating point.

From the Fisher decomposition, the following quantities are analyzed:

- the leading eigenvalues and associated eigenvectors,
- the effective rank of the spectrum,
- the fraction of total curvature explained by the top eigen-directions.

These measures are used to assess whether the supervised discrimination objective induces a low-dimensional sensitivity structure in the embedding space.

4.2.7 Cross-Backbone Comparison

All analyses are performed consistently across multiple pretrained visual backbones, including convolutional and foundation-model architectures. This enables direct comparison of:

- sample efficiency,
- linear separability,
- task-induced spectral concentration,

under identical experimental protocols.

Such comparisons are essential to distinguish general properties of the task from architecture-specific effects.

4.2.8 Methodological Scope

The methodology deliberately avoids end-to-end fine-tuning and synthetic data augmentation at this stage. This is not a rejection of those techniques, but a principled attempt to determine whether they are necessary.

By establishing the representational sufficiency or insufficiency of frozen embeddings first, subsequent training decisions can be grounded in empirical evidence.

```
{
    "image_path": "attentive_synonyms_336_4300/absorbed_seed_1.png",
    "llava_att_inat_full_answer": "The student in the foreground appears to
        be inattentive, as they are sitting at a desk with their hands
       folded, seemingly not paying attention to the class. This could be
       due to various reasons, such as feeling unwell, being distracted by
       something else, or simply not engaged in the lesson. It is important
        for the teacher to identify the cause of the student's
       inattentiveness and address it to ensure the student's well-being
       and academic success.",
    "llava_inat_att_full_answer": "The student in the foreground appears to
        be inattentive, as they are sitting at a desk with their hands
       folded, seemingly not paying attention to the class. This could be
       due to various reasons, such as feeling unwell, being distracted by
       something else, or simply not engaged in the ongoing lesson. It is
       important for the teacher to identify the cause of the student's
       inattentiveness and address it to ensure the student's well-being
       and academic success.",
    "qwenvl_att_inat_full_answer": "The student in the foreground appears
       to be attentive. This can be inferred from their posture and focus.
       They are seated upright, with their hands resting on the desk, which
        suggests they are engaged in the activity at hand, possibly taking
       notes or working on an assignment. Their gaze is directed downward,
       indicating concentration on the task. The overall body language does
        not suggest distraction or inattention.",
    "qwenvl_inat_att_full_answer": "The student in the foreground appears
       to be attentive. This can be inferred from their posture and focus.
       They are seated upright with their hands resting on the desk, which
       suggests they are engaged in the activity at hand, likely listening
       or reading. Their gaze is directed downward, possibly at a piece of
       paper or a book, indicating concentration and attentiveness. The
       overall body language does not suggest inattention; rather, it
       conveys a sense of focus and engagement with the task.",
    "votes": {
      "llava_forward": -1,
      "llava_reverse": -1,
      "qwenvl_forward": 1,
      "qwenvl_reverse": 1
    },
    "hard_label": "inconclusive",
    "sum_attentive_votes": 2,
    "sum_inattentive_votes": -2,
    "sum_all_votes": 0
  },
```


Figure 4 – Example of a VLM-evaluation record used to derive the label for Figure 2.

```
{
  "image_path": "attentive_synonyms_336_4300/absorbed_seed_3.png",
  "llava_att_inat_full_answer": "The student in the foreground appears to
      be attentive, as she is sitting at a desk with her hands folded,
     looking at the camera. This posture suggests that she is focused and
      engaged in the activity or discussion taking place in the classroom
     . The fact that she is sitting in a classroom with other students
     also implies that she is participating in a learning environment,
     which typically requires attention and engagement.",
  "llava_inat_att_full_answer": "The student in the foreground appears to
      be attentive, as she is sitting at a desk with her hands folded,
     looking at the camera. This suggests that she is focused on the task
      at hand, possibly participating in a class or an educational
     activity. The fact that she is sitting in a classroom with other
     students also implies that she is engaged in a learning environment,
      which further supports the idea that she is attentive and actively
     participating in the class.",
  "qwenvl_att_inat_full_answer": "The student in the foreground appears
     to be attentive. This can be inferred from their posture and
     expression. They are sitting upright with their arms crossed, which
     is often a sign of engagement and focus. Additionally, their gaze
     seems directed towards something in front of them, possibly the
     teacher or a presentation, indicating that they are paying attention
      to what is happening in the classroom. The overall demeanor
     suggests that they are actively participating in the lesson.",
  "qwenvl_inat_att_full_answer": "The student in the foreground appears
     to be attentive. This can be inferred from several observations:\n\
     n1. **Posture**: The student is sitting upright with their arms
     crossed, which often indicates focus and attentiveness.\n2. **Eye
     Contact**: Although the student's gaze is directed slightly off-
     camera, it suggests they might be looking at something of interest,
     possibly related to the lesson or activity.\n3. **Environment**: The
      classroom setting, with other students engaged in activities,
     supports the idea that the student is participating in class.\n\
     nWhile it's not possible to determine the exact level of attention
     without more context, the posture and environment suggest
     attentiveness.",
  "votes": {
    "llava_forward": 1,
    "llava_reverse": 1,
    "qwenvl_forward": 1,
    "qwenvl_reverse": 1
  },
  "hard_label": "attentive",
  "sum_attentive_votes": 4,
  "sum_inattentive_votes": 0,
  "sum_all_votes": 4
},
```


Figure 5 – Example of a VLM-evaluation record used to derive the label for Figure 3.

Part III

Experimental Evaluation and Conclusions

5 Results

This chapter presents the empirical results obtained in this work. The objective is not to exhaustively benchmark architectures, but to assess, in a controlled and reproducible setting, the quality and usefulness of different visual representations when transferred to a downstream binary classification task using frozen representations and a limited number of labeled samples.

The chapter is organized into two main sections. First, we describe the dataset used in the experiments, including its construction, splits, and basic properties relevant to interpretation of the results. This section establishes the empirical context and constraints under which all subsequent evaluations were conducted.

Second, we report and analyze the experimental results obtained using frozen visual backbones coupled with lightweight classification heads. The analysis focuses on learning curves as a function of the number of labeled samples per class, comparing different model families and probing strategies. Performance is evaluated using complementary metrics that capture both class-balanced accuracy and ranking quality. We seek to identify performance regimes, saturation effects, and the marginal benefits of increased model capacity or data availability, as opposed to maximizing absolute scores.

Finally, we observe the behavior of task-induced geometry of frozen representations using the empirical Fisher Information Matrix computed with respect to the embedding vectors, in an effort to characterize how supervised task pressure induces anisotropy and low-dimensional structure in frozen embedding spaces, as revealed by the spectral properties of the empirical Fisher Information Matrix.

Throughout this chapter, results are reported as averages over multiple independent splits, with variability explicitly indicated.

5.1 Dataset Results

5.1.1 Vision Foundation Models, the Probing Paradox, and Instrumental Data for Identifiable Representation Analysis

As mentioned, the advent of Vision Foundation Models, trained on massive and heterogeneous image corpora, has fundamentally altered the role of data in Computer Vision research. For a broad class of tasks, representation learning is no longer constrained by the availability of labeled datasets: pretrained encoders released by large industrial laboratories already provide rich, transferable visual representations that no individual,

and few academic efforts can hope to rival in scale or diversity.

This shift, however, gives rise to a probing paradox. While the burden of learning general-purpose representations has been largely externalized to pretrained models, the scientific task of understanding those representations—their geometric structure, encoded invariances, and systematic failure modes—remains intrinsically data-dependent. Probing embedding spaces, assessing sensitivity to semantic variation, and characterizing inductive biases cannot be performed without deliberately structured inputs against which such properties can be made observable.

The data required for such probing need not — and arguably should not — approximate the true distribution of the real world in an uncontrolled manner. When multiple sources of variation are allowed to fluctuate simultaneously, as is typical of naturalistic data, observed structure in the embedding space cannot be unambiguously attributed to the semantic factor under investigation. In this sense, uncontrolled realism can obscure rather than illuminate representational sensitivities. What is required instead are inputs that allow systematic manipulation of visual attributes under explicit operational control. Also, when multiple sources of variation are allowed to fluctuate simultaneously, observed structure in the embedding space cannot be unambiguously attributed to the semantic factor under investigation.

What representation analysis requires instead are inputs that permit explicit, repeatable manipulation of specific visual attributes under operational control. Such control is a prerequisite for identifiability: without it, claims about low-rank structure, dominant directions of variation, or sensitivity of embeddings to semantic change are fundamentally underdetermined.

To the best of our knowledge, at the time this work was conducted, no existing dataset simultaneously satisfied the set of characteristics required for this purpose: (i) systematic variation of apparent attentional and inattentional configurations exhibited by school-aged students within a fixed classroom context; (ii) the ability to bias visual generation toward specific pose, gaze, and activity configurations while limiting confounding variation in identity and scene layout, without requiring explicit fine-grained control; and (iii) scalability sufficient to support statistical and geometric analyses across multiple pretrained models.

It is in this context that synthetic data becomes not merely convenient, but necessary. By enabling the deliberate induction of specific visual configurations through prompt design, stochastic variation, and controlled perturbations, synthetic images function not as training material, but as probes. They constitute an instrumental apparatus for interrogating pretrained encoders whose training distributions and internal organization are otherwise opaque.

From this perspective, the dataset introduced in this work is not an alternative to existing benchmarks, nor an attempt to approximate real-world image statistics. Its role is experimental: to render representation-level hypotheses testable by providing controlled, repeatable stimuli. The geometric phenomena reported in the following sections — most notably the emergence of low-rank structure and dominant Fisher directions — are meaningful precisely because the underlying inputs were designed to isolate the semantic variable of interest. This isolation does not imply that synthetic conditions faithfully reproduce real-world distributions; rather, it ensures that any observed representational structure can be traced to a known source of semantic variation, lending interpretive clarity to the findings.

It is important to emphasize that the objective of this study was not to evaluate the general quality of visual embeddings in an arbitrary setting, nor to benchmark pretrained encoders on an existing classification task. The goal was to determine whether compact, image-level representations—specifically the CLS token in Vision Transformers and the Global Average Pooling representation in convolutional networks—are sufficiently expressive to encode information associated with *apparent attentional states* as assessed by Vision–Language Models. This question is inherently context-dependent: attentional and inattentional configurations are not generic visual concepts, but emerge from the interaction between pose, gaze, activity, and classroom context. As such, answering it requires inputs in which these factors are systematically present and interpretable, which would not be found in a generic dataset optimized for unrelated objectives.

5.1.2 Dataset Composition

The dataset, prior to curation (deduplication and similarity-based filtering), consists of multiple subsets generated through distinct prompt strategies, and includes[1]:

- **4,300** images generated from synonyms of *attentive* (Figures 6-7-8);
- **2,700** images generated from *attentive* behavioral proxies, such as posture (Figures 9-10-11);
- **1,000** images generated from *inattentive* behavioral proxies (Figures 14-15-16);
- **6,900** images generated from activities that might be regarded as indicating *inattentiveness* (Figures 17-18-19).

Although an initial goal was to maintain numerical balance between attentive and inattentive categories, substantial asymmetries emerged during generation and validation. The generation-related asymmetries reflected the well-documented representational biases

[1] **4,300** images generated from synonyms of *inattentive* were discarded, due to semantic/hash collision



Figure 6 – "attentive", seed 8



Figure 7 – "curious", seed 86



Figure 8 – "engaged", seed 23



Figure 9 – "focused on the teacher with concentrated face", seed 29



Figure 10 – "listening attentively seated at desk", seed 2



Figure 11 – "looking at the teacher with focused expression", seed 20

in image generation models, particularly the tendency toward frontal and canonical viewpoints. Such biases, commonly referred to as frontal bias, consistently interfered with the faithful rendering of inattentive behaviors that require non-frontal head or body orientations. Repeated attempts to elicit behaviors such as “looking sideways” or “looking away from the instructional focus” frequently collapsed into frontal or near-frontal poses, even when structural guidance was introduced via ControlNet.

To illustrate that these limitations were not confined to smaller or earlier systems, Figures 12–13 show attempts, contemporaneous with this work, to generate an image of a student in an apparent inattentive state—seated at a desk in a canonical classroom configuration but directing gaze and posture toward the back of the room, as if engaging with a peer. Despite the use of substantially larger, more capable models, the generations consistently failed to realize the intended non-frontal attentional configuration. At the time, ChatGPT’s content restrictions prevented the generation of realistic images of minors, leading us to substitute a college student and a CCTV-style framing; even under these relaxed conditions, the desired pose reliably collapsed to undesired viewpoints. This image

seems particularly interesting, because it shows all students's desks turned to the back of the classroom, and the teacher turning his back on the students. While subsequent generations of large models have since improved markedly in image synthesis quality, these examples document a concrete representational limitation present during the development of this pipeline.

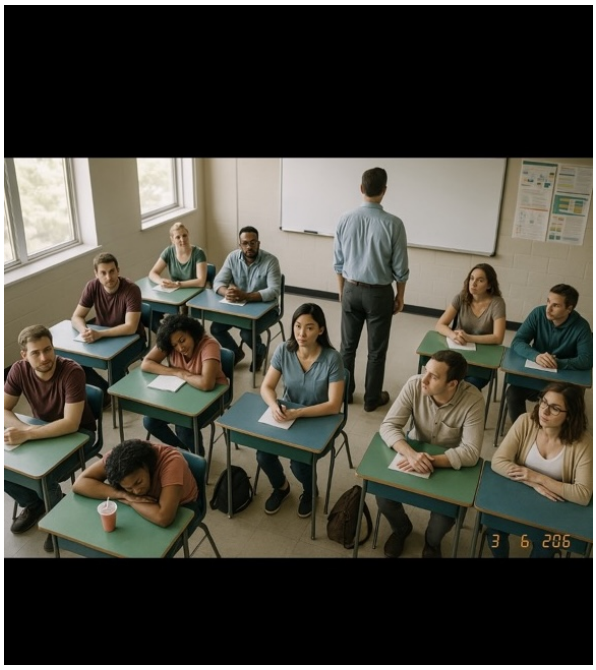


Figure 12 – Synthetic image created by ChatGPT, as CCTV-like image of college student seated, looking towards the back



Figure 13 – Synthetic image created by ChatGPT, as student seated in the classroom, looking towards the back

In parallel, additional limitations arose at the semantic level. The generation and validation pipelines rely on CLIP-based joint text–image embeddings, which are known to handle negation and antonymic concepts poorly (TONG; JONES; STEINHARDT, 2024). In practice, descriptors intended to denote inattentiveness (e.g., "not attentive") often collapsed toward embeddings associated with attentiveness, reflecting a form of semantic aliasing in the shared latent space. Distinct textual concepts, although linguistically opposed, were mapped to nearby regions of the embedding manifold, thereby undermining the reliable generation and discrimination of negative or contrastive behavioral states.

The resulting asymmetry is preserved and explicitly documented, instead of having this imbalance corrected through post hoc resampling. This imbalance is treated as an empirical observation about the interaction between language prompts and the generative model, and not as a dataset construction error.

Samples for activities under the attentive and inattentive categories, along with the respective assessments by the VLMs, are provided in the appendixes.



Figure 14 – "leaning forward with eyes half closed appearing tired", seed 3



Figure 15 – "rubbing temples with closed eyes at desk", seed 1



Figure 16 – "supporting head with both hands looking bored", seed 39



Figure 17 – "blowing a bubble with gum at the desk", seed 3



Figure 18 – "checking a phone", seed 55



Figure 19 – "having a side conversation", seed 41

5.1.3 Deduplication and Similarity-Based Filtering

Synthetic image generation might produce near-duplicate samples, particularly when multiple behavioral descriptors—despite lexical variation—collapse into highly similar visual realizations. This effect was especially pronounced for attentive-related descriptors, where distinct synonyms consistently converged to near-identical canonical depictions. Such semantic–visual collapse limits effective sample diversity even under stochastic sampling. To mitigate this effect, a deduplication stage is applied.

Images are embedded using a fixed pretrained visual encoder, and pairwise cosine similarity is computed within prompt groups. Images exceeding a similarity threshold of **0.975** are considered near-duplicates, and redundant samples are removed (Figures 20 –21).

In addition, a similarity-based filtering step is applied to remove samples in the left tail of the similarity distribution relative to their intended prompt group. These outliers (such as Figure 22) typically correspond to generations that visually diverge from the



Figure 20 – behavior ”focused”



Figure 21 – behavior ”intellectually engaged”



Figure 22 – Image generated with behavior "ready to learn" and seed 75

induced behavior or context, often due to prompt misinterpretation or generative artifacts.

The dataset, once deduplication and similarity-based filtering were performed, consists of:

- **2,340** images generated from synonyms of *attentive*;
- **2,144** images generated from *attentive* behavioral proxies, such as posture;

- **883** images generated from *inattentive* behavioral proxies;
- **6,709** images generated from activities that might be regarded as indicating *inattentiveness*.

Table 1 – Dataset composition before and after deduplication and similarity-based filtering.

Source of images	**Pre-curation**	**Post-curation**
Synonyms of *attentive*	4,300	2,340
Attentive behavioral proxies (e.g., posture)	2,700	2,144
Inattentive behavioral proxies (e.g., posture)	1,000	883
Activities indicative of *inattentiveness*	6,900	6,709
Total	**14,900**	**12,076**

5.1.4 Assessments by Vision–Language Models

Although the language model initially proposed a structured set of activities intended to operationalize inattentiveness, images generated from these behavioral proxies were disproportionately judged as attentive by the Vision-Language Models. This pattern reveals a systematic bias toward attentiveness as a default interpretation in visually ambiguous classroom scenarios. As a consequence, prompts intended to elicit inattentive configurations exhibited a higher rejection rate, not solely due to failures in image generation, but also as a result of semantic drift introduced during automated model-based evaluation.

Following aggregation of Vision–Language Model judgments, the final dataset comprises 12,076 labeled images, distributed as follows (Table 7):

- **9,525** labeled as *attentive*;
- **1,628** labeled as *inattentive*;
- **923** labeled as *inconclusive*.

5.1.5 Summary and Dataset Limitations

Despite filtering and deduplication, the dataset remains imperfect. Common residual artifacts include missing or extra limbs, distorted facial features, inconsistent object geometry. These artifacts are characteristic of contemporary diffusion-based generative models under the constraints of this work.

Such imperfections are acknowledged as inherent to the generative pipeline adopted here, and are not manually corrected. While they do not constitute realistic noise in the strict distributional sense, their presence introduces a form of uncontrolled visual variation

that is consistent with the imperfect conditions under which pretrained representations are typically deployed. The dataset therefore does not claim visual cleanliness, but controlled construction — and the distinction matters for interpreting the results.

The synthetic dataset presented in this chapter is designed as an experimental probe, not as a comprehensive benchmark for attention detection. Its construction prioritizes controllability, transparency, and feasibility over realism or completeness.

Key limitations include generative artifacts, linguistic bias in prompt interpretation, and class imbalance arising from asymmetric labeling. To prevent these imbalances from confounding downstream analyses, class proportions are explicitly controlled during the construction of the training, validation, and test splits.

5.2 Experimental Probing Results

This section presents the experimental results obtained using the dataset described in the previous section. The objective is to empirically characterize the behavior of different frozen visual representations when transferred to a downstream binary classification task, under controlled variations of training set size and classifier capacity.

Results are analyzed as a function of the number of labeled training samples per class, trying to enable the identification of distinct performance regimes ranging from low-data to larger training regimes.

The following subsections present learning curves, probe comparisons, and auxiliary analyses with the aim of identifying consistency across repetitions, performance regimes, saturation effects, and diminishing returns with respect to both data and model capacity.

Finally, we analyze the intrinsic geometry of the embedding spaces using linear probes based on Linear Discriminant Analysis to assess class separability, and information-theoretic spectral measures derived from the Fisher information, including trace and effective rank, to characterize effective dimensionality and anisotropy.

5.2.1 Evaluation Protocol

All experiments follow a unified evaluation protocol designed to isolate the properties of the visual representations from confounding factors related to joint optimization of the encoder and classifier. For each model family, the backbone is kept frozen and used solely as a feature extractor, with supervision applied only through lightweight classification heads trained on top of the extracted embeddings. This design intends to isolate the contribution of the representations themselves, independently of effects arising from backbone updating or task-driven representation adaptation.

Two types of probes are considered: a linear classifier and a shallow multilayer perceptron (MLP). This design seeks to allow the assessment of whether the class separation induced by each representation is approximately linear, or whether additional non-linear capacity provides systematic gains.

To evaluate data efficiency, training is performed using multiple set sizes of labeled samples per class. For each training size, multiple independent repetitions are executed using distinct but fixed train/validation/test splits across models and probes, seeking to ensure that comparisons across models and probes are paired and statistically meaningful. Unless otherwise stated, reported results correspond to averages over repetitions, with uncertainty indicated through confidence intervals, as shaded regions in the plots.

Performance is reported on a held-out test set using Macro-F1 as the primary metric, complemented by ROC-AUC. Macro-F1 is chosen to equally weight both classes hoping to reflect classification performance, while ROC-AUC intends to provide a threshold-independent measure of ranking quality and class separability.

Beyond absolute performance, the effect of probe capacity is analyzed through paired differences between linear and non-linear heads, non-linear probes (no drop-out vs. 20 % drop-out), as well as win-rate statistics, to try to quantify the magnitude and consistency of observed gains across repetitions.

5.2.2 Learning Curves as a Function of Data and Probe Capacity

We begin by analyzing learning curves as a function of the number of labeled samples per class, using Macro-F1 and ROC-AUC as complementary performance metrics. These curves are used as the primary empirical evidence for assessing how representation quality and probe capacity interact under frozen visual backbones.

5.2.2.1 Macro-F1 learning curves with linear probes

Figure 23 reports Macro-F1 as a function of training set size for linear probes applied to the three backbone families. Two observations emerge. First, backbone choice dominates performance across all data regimes: both foundation models consistently outperform the convolutional baseline by approximately ten percentage points. Second, performance saturates rapidly as the number of labeled samples increases, indicating strong sample efficiency and suggesting diminishing returns from additional supervision under a linear probing regime.

5.2.2.2 Effect of shallow non-linearity on Macro-F1

Figure 24 repeats the same analysis using a shallow multilayer perceptron (MLP) probe with a single hidden layer composed of 512 units. While the overall shape and

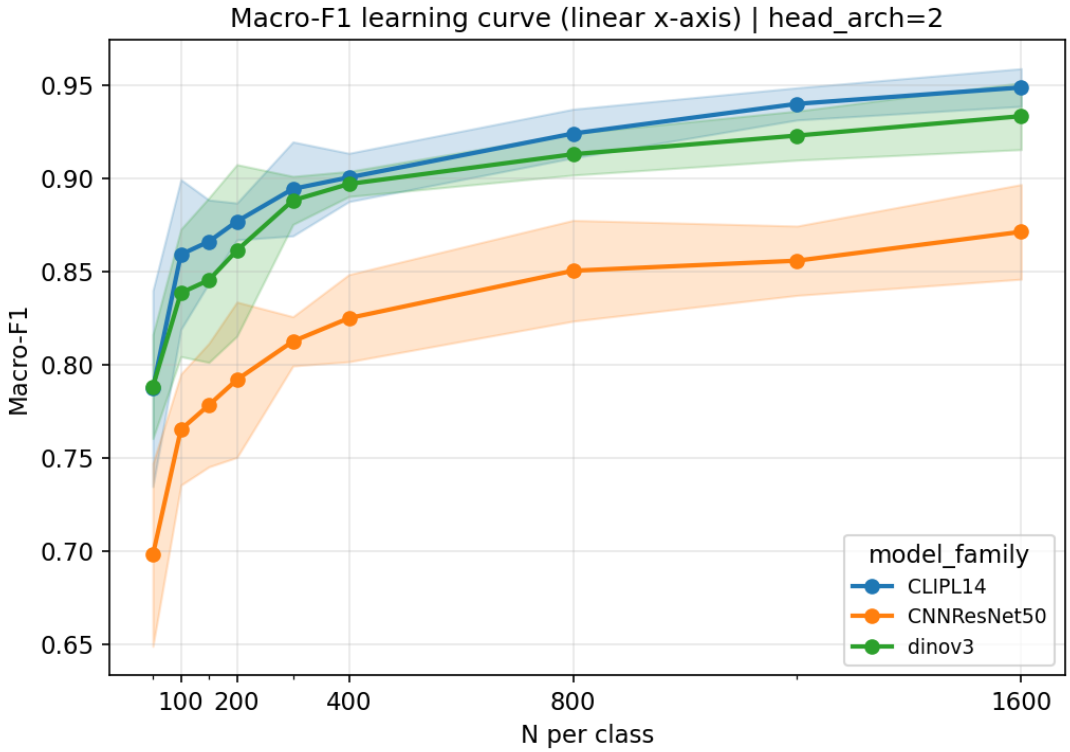


Figure 23 – Learning curve of Macro-F1 for a linear classification head, as a function of training size, comparing different frozen visual backbones.

saturation behavior of the learning curves remain consistent with those obtained using linear probes, the MLP exhibits a systematic advantage in the low-data regime, particularly at small training set sizes. This gap narrows as the number of labeled samples increases, and performance converges across probe types in the higher-data regime. Taken together, these results indicate that limited non-linear capacity can improve robustness under data scarcity, but does not substantially alter asymptotic performance or uncover additional task structure once the backbone representation is fixed.

5.2.2.3 ROC-AUC learning curves and metric consistency

To assess whether the observed trends depend on a threshold-sensitive metric, Figures 25 and 26 report ROC-AUC under the same experimental conditions. The overall qualitative pattern mirrors that observed for Macro-F1: performance is primarily determined by the choice of frozen backbone, with Vision Foundation Models consistently outperforming the convolutional baseline across data regimes.

A more fine-grained comparison, shown in Figure 28, reveals that the shallow MLP provides a measurable advantage over the linear probe in the low-data regime, particularly for smaller training set sizes. However, this advantage diminishes steadily as the number of labeled samples increases, with the performance gap approaching zero in higher-data regimes. As a result, increased probe capacity improves sample efficiency but does not

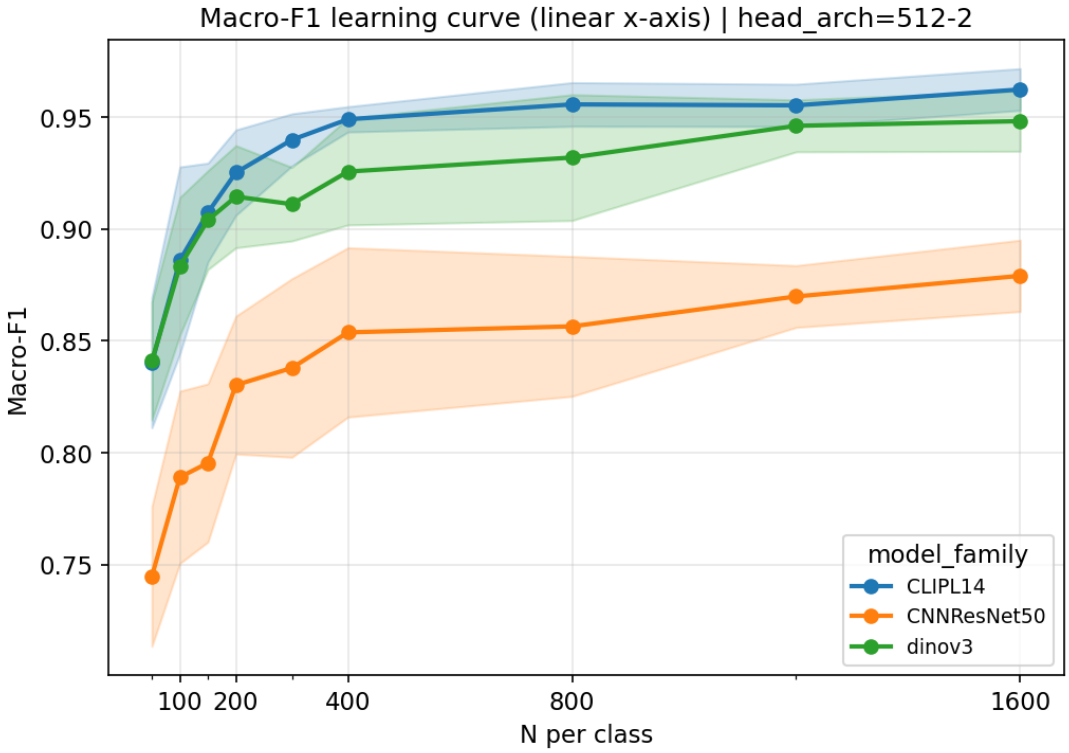


Figure 24 – Learning curve of Macro-F1 for a shallow MLP classification head, as a function of training size, comparing different frozen visual backbones.

substantially alter asymptotic performance.

The close agreement between trends observed for Macro-F1 and ROC-AUC indicates that these findings are robust to the choice of evaluation metric and do not depend on a particular decision threshold or operating point.

5.2.2.4 Paired Analysis of Probe Architectures

To complement the learning-curve analysis, we express performance in paired-difference form between probe architectures. For each training size and repetition, we compute the difference in performance between the shallow MLP (512–2) and the linear probe on matched data splits, and report the mean paired difference ($\Delta = \text{MLP} - \text{Linear}$) as a function of training set size. This representation centers the comparison on relative effects of probe capacity while factoring out variance shared across backbones, data splits, and repetitions, and allows small but systematic differences to be visualized directly.

5.2.2.4.1 MLP versus linear probe

Figure 27 reports the mean paired difference in Macro-F1. Across all backbones and training sizes, the mean gain remains small and often close to zero, with uncertainty intervals frequently overlapping zero. While positive deltas appear at low data regimes for some backbones, these effects systematically diminish as training size increases. This

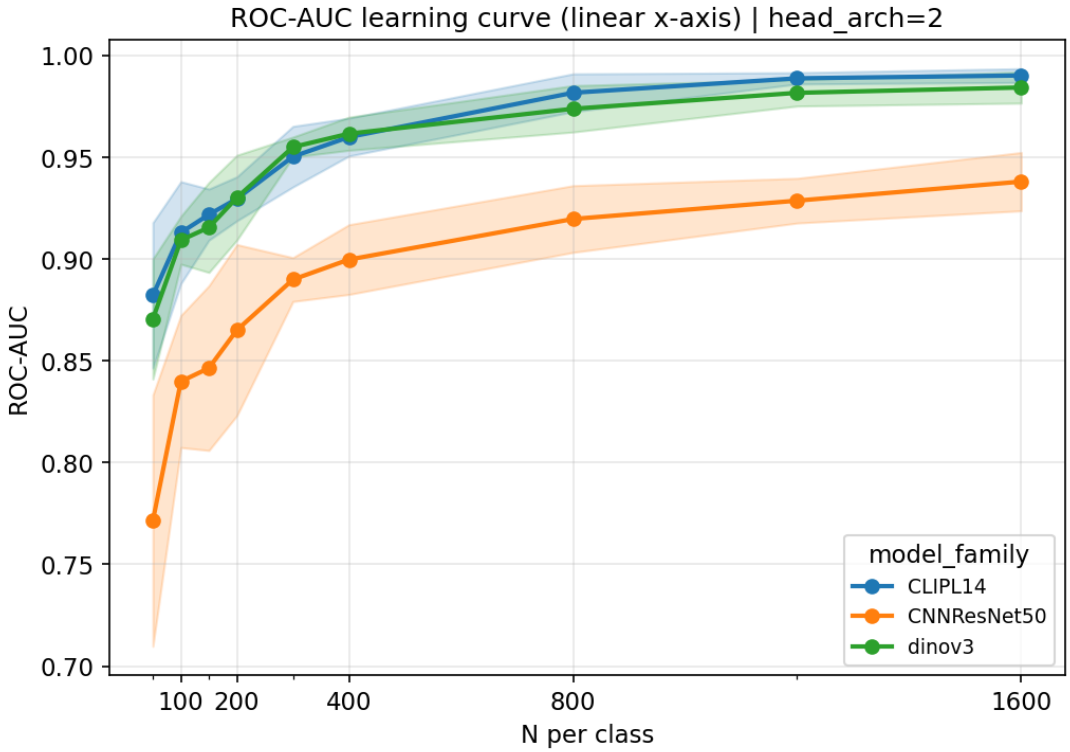


Figure 25 – Learning curve of ROC-AUC for a linear classification head, as a function of training size, comparing different frozen visual backbones.

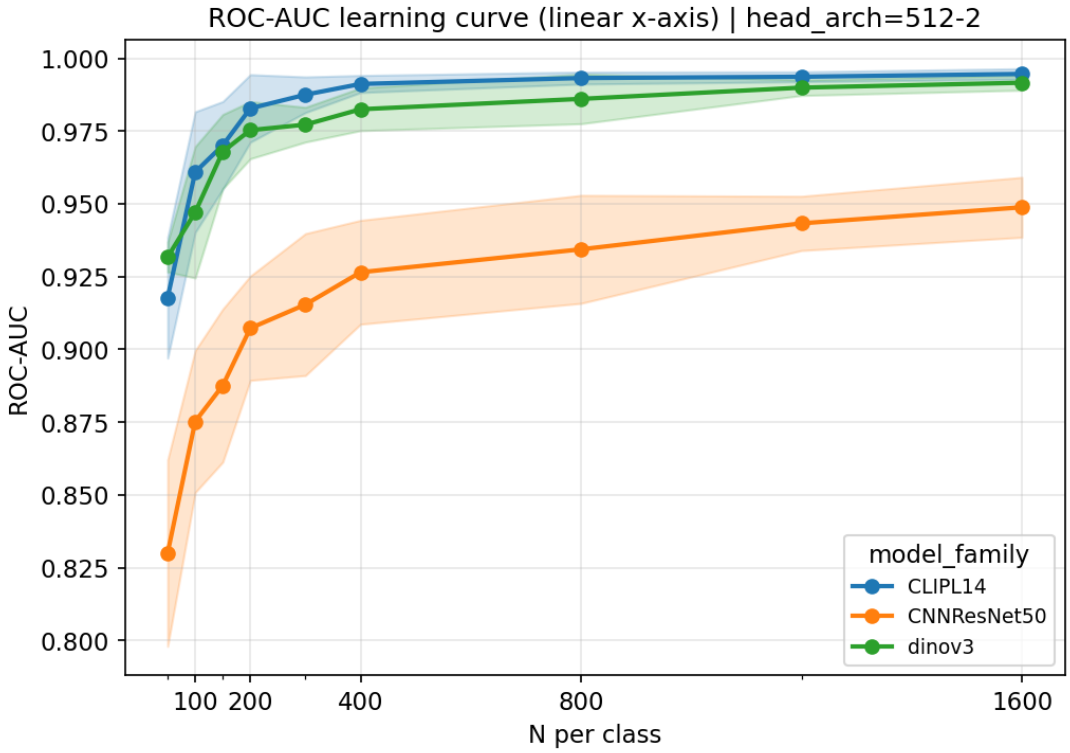


Figure 26 – Learning curve of ROC-AUC for a shallow MLP classification head, as a function of training size, comparing different frozen visual backbones.

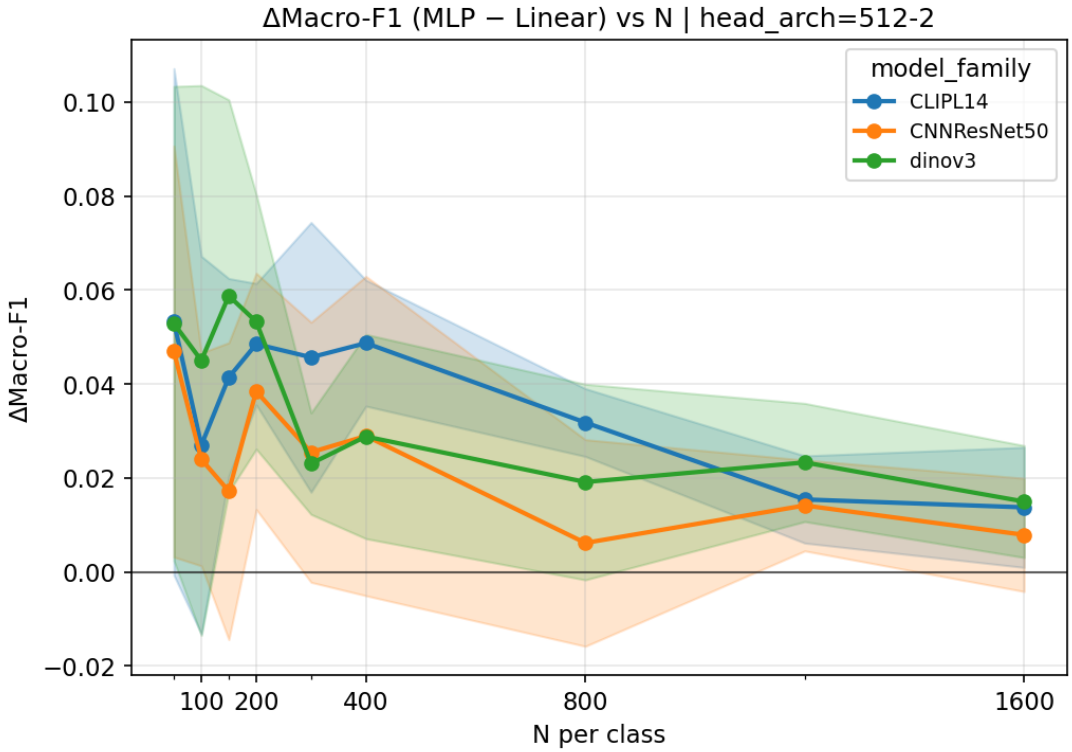


Figure 27 – Mean paired difference in Macro-F1 between an MLP classification head (512–2) and a linear probe (Δ = MLP − Linear), as a function of training size.

indicates that shallow non-linearity does not provide a robust or scalable improvement in class-balanced accuracy over linear probes.

An analogous pattern is observed for ROC-AUC in Figure 28. The paired differences remain near zero across regimes, suggesting that even when performance is evaluated through a ranking-based metric, non-linear probes fail to deliver consistent advantages over linear decision boundaries.

Finally, Figure 29 reports paired differences in test loss. While MLP heads occasionally reduce loss at very small training sizes, these benefits shrink rapidly and converge toward zero as data increases. Together, these results indicate that performance differences between linear and shallow non-linear probes largely reflect small-sample effects not from systematic representational gains. The frozen embeddings already support effective linear separation, leaving limited room for improvement through additional probe complexity.

5.2.2.4.2 Effect of dropout regularization

We further examine whether stochastic regularization at the probe level alters this behavior. The figures bellow report paired differences between MLP probes trained without dropout and the baseline MLP configuration.

Figure 30 shows the paired difference in Macro-F1. Across backbones and training

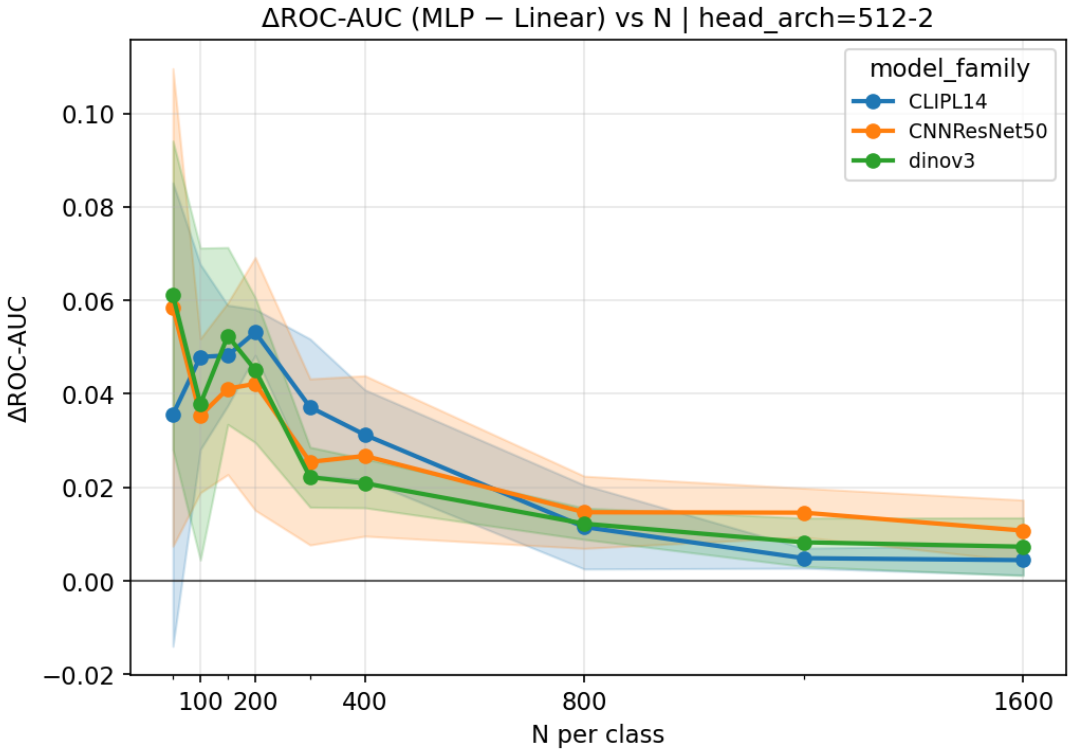


Figure 28 – Mean paired difference in ROC-AUC between an MLP classification head (512–2) and a linear probe (Δ = MLP − Linear) as a function of training size.

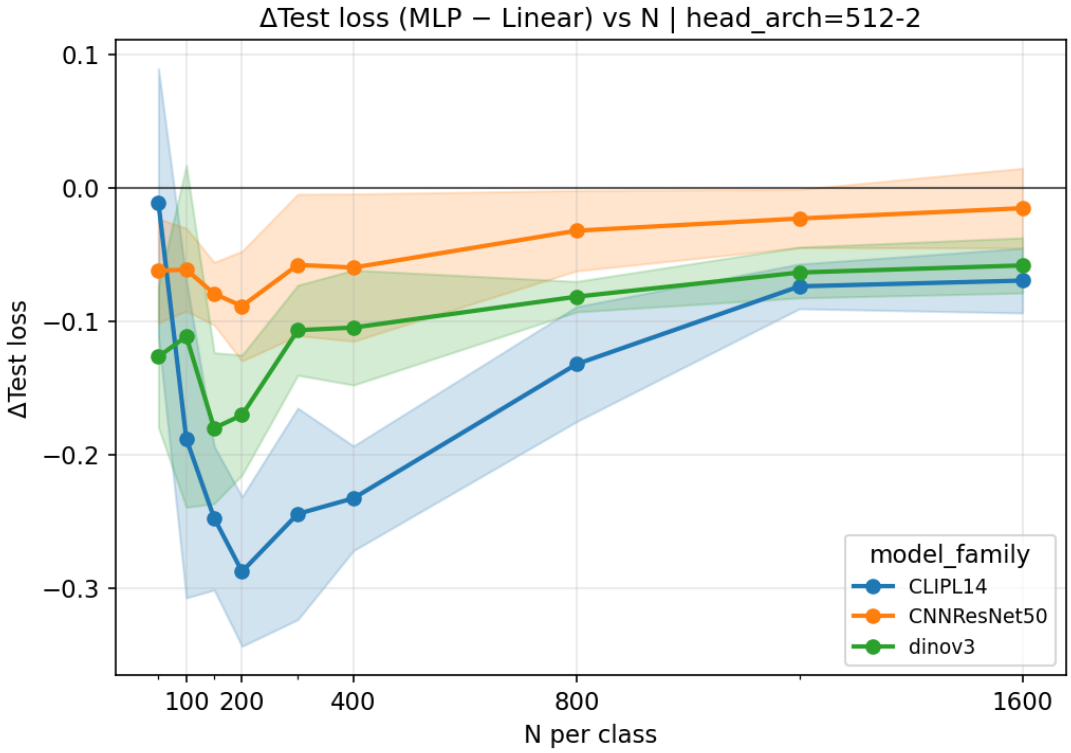


Figure 29 – Mean paired difference in test loss between an MLP classification head (512–2) and a linear probe (Δ = MLP − Linear) as a function of training size.

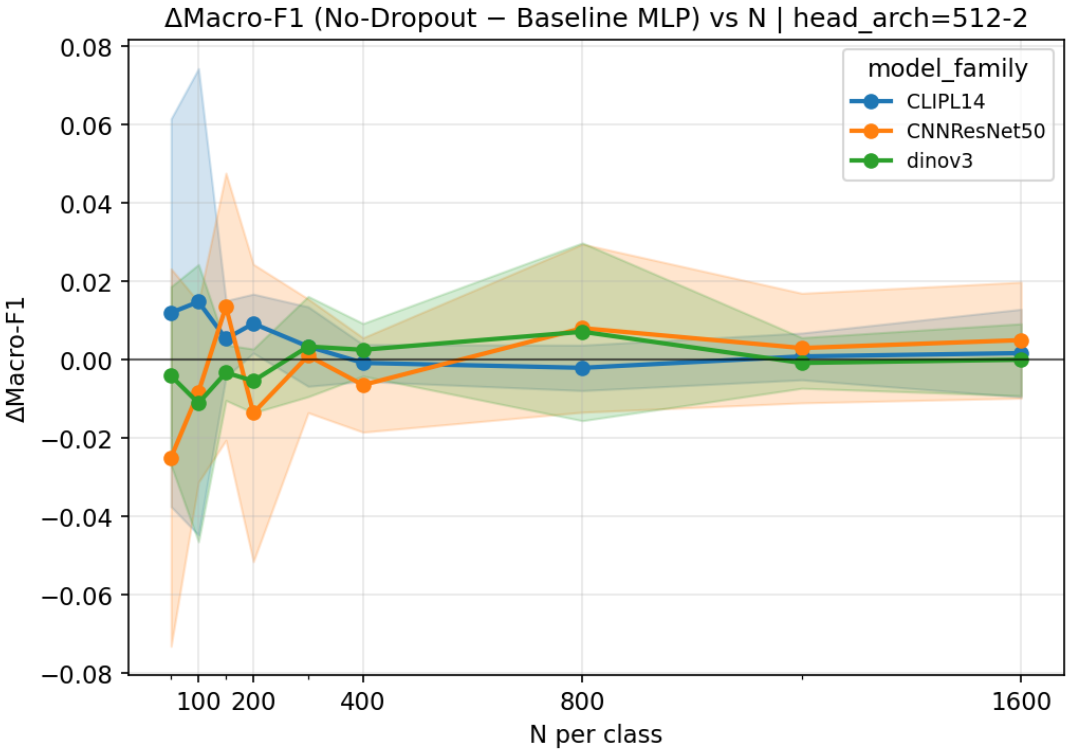


Figure 30 – Mean paired difference in Macro-F1 between an MLP classification head (512–2) with no dropout and the baseline MLP (Δ = No-Dropout − Baseline), as a function of training size.

sizes, removing dropout does not degrade performance and yields small positive or near-zero deltas. No regime exhibits systematic deterioration when dropout is removed.

A similar pattern holds for ROC-AUC in Figure 31. Paired differences remain tightly concentrated around zero, indicating that dropout does not provide measurable benefits for ranking-based performance in this setting.

Finally, Figure 32 reports paired differences in test loss. The deltas again cluster around zero, with no consistent evidence that dropout improves generalization. In some low-data regimes, dropout-free probes slightly reduce loss, but these effects are small and not persistent across training sizes.

Taken together, these results indicate that probe-level dropout has a negligible effect on downstream performance when applied to frozen representations in this task. Within the explored regimes, performance appears largely insensitive to this form of regularization, suggesting that probe capacity and regularization play a secondary role relative to the information already accessible in the frozen embeddings.

5.2.2.4.3 Win-rate analysis across repetitions

To assess whether the near-zero mean paired differences conceal consistent directional effects at the level of individual runs, we complement the paired-difference analysis

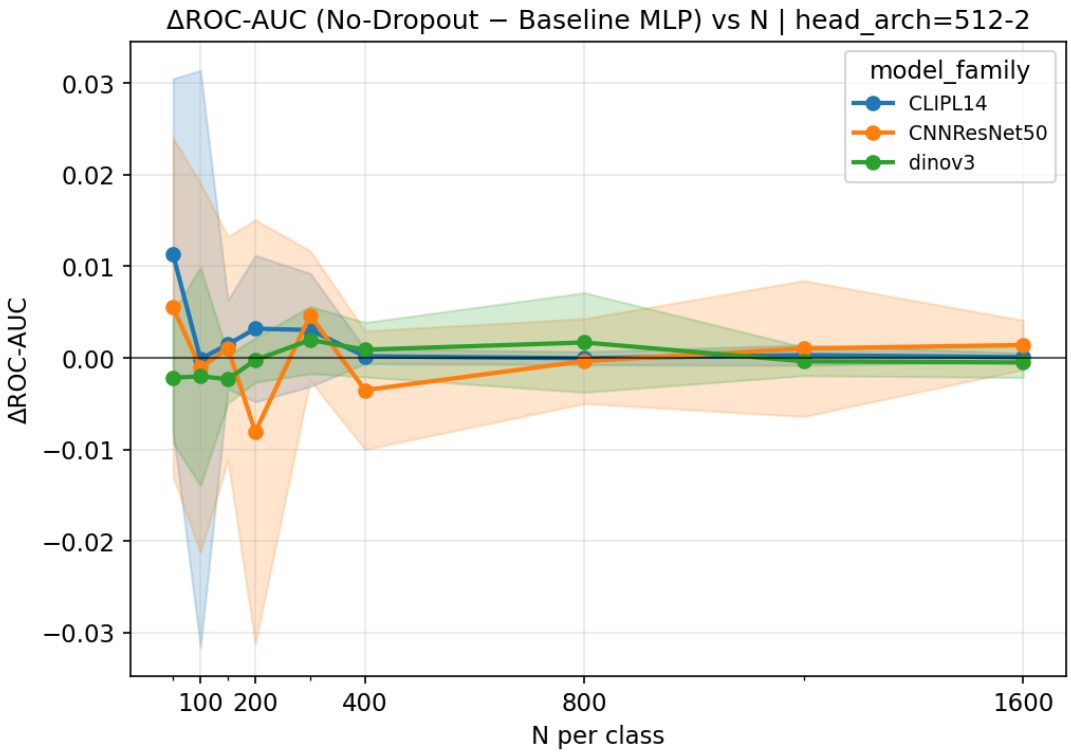


Figure 31 – Mean paired difference in ROC-AUC between an MLP classification head (512–2) with no dropout and the baseline MLP (Δ = No-Dropout − Baseline), as a function of training size.

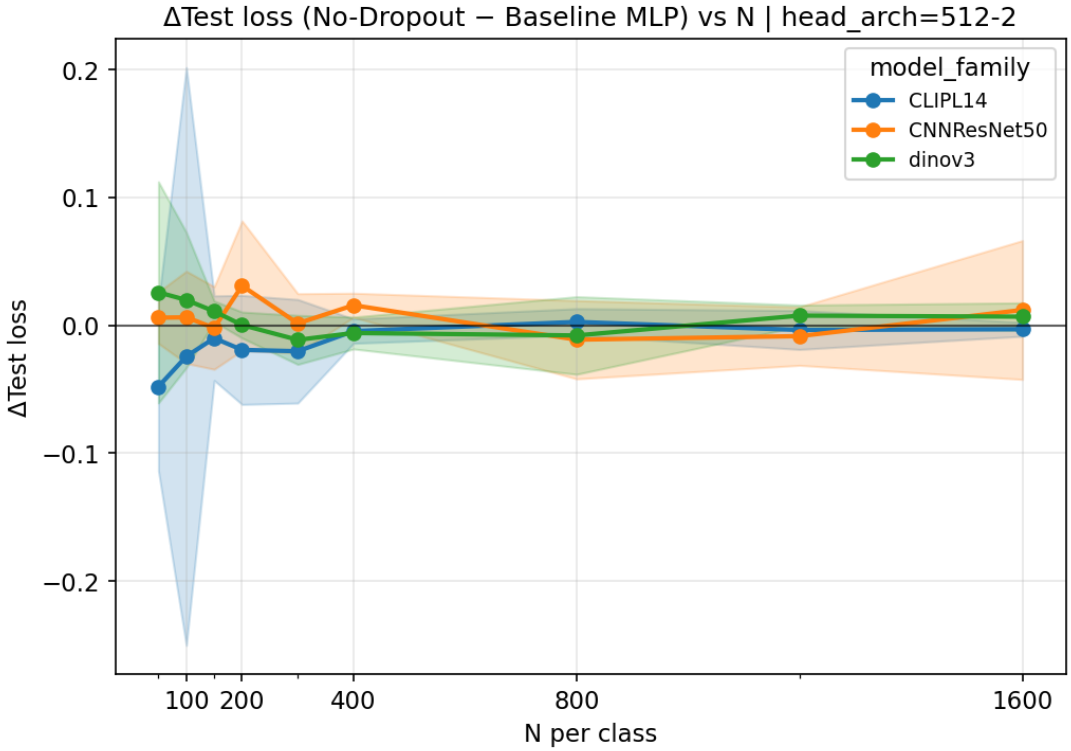


Figure 32 – Mean paired difference in test loss between an MLP classification head (512–2) with no dropout and the baseline MLP (Δ = No-Dropout − Baseline), as a function of training size.

with a win-rate analysis. The figure reports, for each training size and backbone, the empirical probability that one probe configuration achieves higher Macro-F1 than its counterpart across repeated data splits.

For the comparison between MLP and linear probes (Figure 33), win-rates reveal modest but backbone-dependent directional tendencies. For the foundation model backbones, the MLP probe more frequently outperforms the linear probe across repetitions, particularly for CLIP-L/14, where the win-rate approaches unity across most data regimes. However, these advantages remain small in magnitude and do not increase with training size, consistent with the paired-difference results.

In contrast, the convolutional baseline exhibits more variable win-rates that fluctuate across training sizes, with no stable preference for either probe configuration. Taken together, the win-rate analysis confirms that while shallow non-linearity can yield consistent but modest gains for some backbones, it does not induce a qualitatively different performance regime.

A similar pattern emerges when comparing MLP probes trained with and without dropout (Figure 34). Win-rates alternate around chance level, with no configuration exhibiting consistent superiority across repetitions. In some regimes, removing dropout leads to more frequent wins; in others, the opposite holds, indicating sensitivity to split-level variability not attributable to a robust regularization effect.

Taken together, the win-rate analysis supports the conclusions drawn from the mean paired-difference plots: differences attributable to probe architecture or probe-level regularization are small, inconsistent, and subordinate to the influence of the frozen backbone representation.

5.2.3 Closed-form Linear Probes: Fisher Linear Discriminant Analysis

5.2.3.1 LDA learning curves under frozen visual representations

The following figures report learning curves obtained using Fisher's Linear Discriminant Analysis (LDA) as a closed-form linear probe applied to frozen visual representations. Unlike the linear and MLP probes considered in previous sections, which are trained through loss minimization, LDA defines a linear decision rule directly from second-order statistics of the embeddings, without iterative optimization or task-adaptive parameter updates.

Across all backbone families, LDA exhibits learning dynamics that differ from those observed for trained discriminative probes. While performance at very small training sizes can be competitive, improvements with increasing data are irregular and, in some regimes, non-monotonic. This behavior is consistently observed in both Macro-F1 (Figure 36) and ROC-AUC (Figure 37), and contrasts with the smoother and more stable learning curves

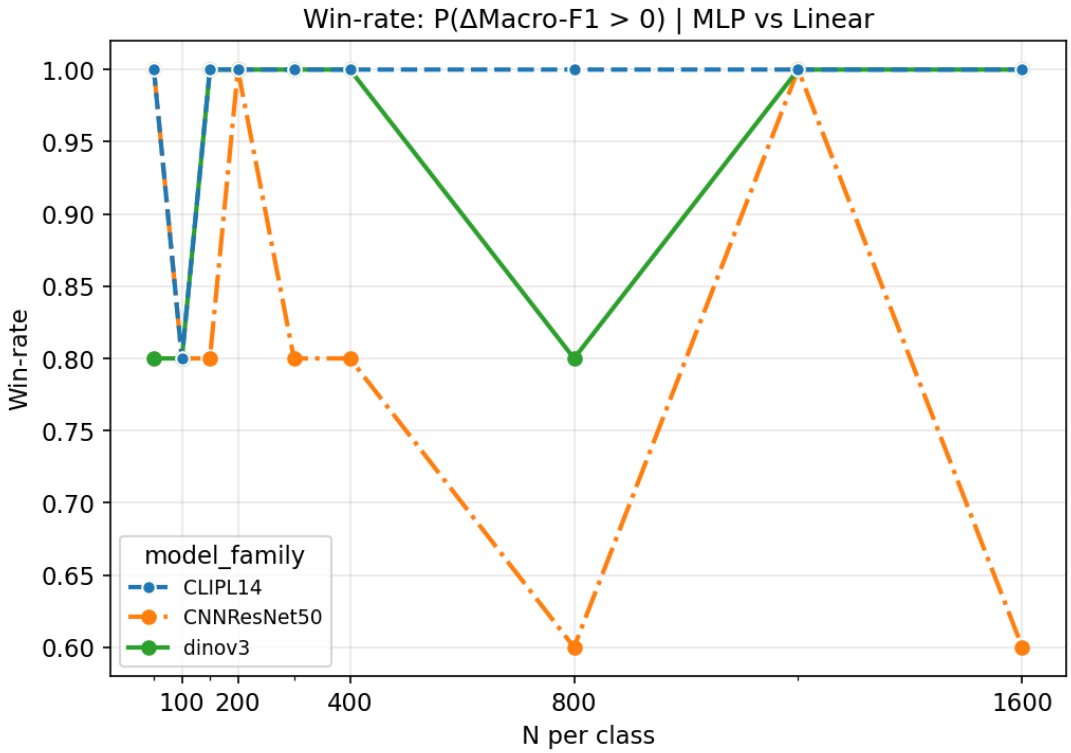


Figure 33 – Probability that an MLP classification head achieves higher Macro-F1 than a linear probe (Δ = MLP - Linear) as a function of training size, computed over repeated splits for each frozen backbone.

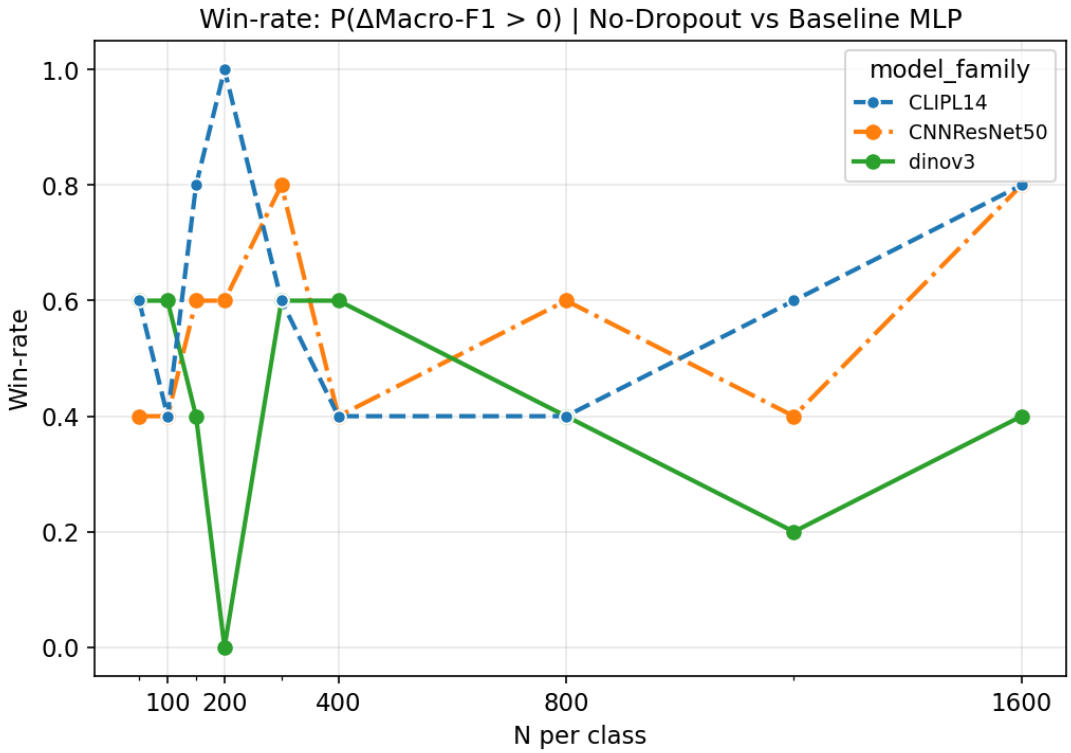


Figure 34 – Probability that an MLP classification head (512–2) with no dropout achieves higher Macro-F1 than the baseline MLP (Δ = No-Dropout - Baseline) as a function of training size, computed over repeated splits for each frozen backbone.

obtained with linear and shallow MLP heads.

These differences reflect the fixed inductive bias imposed by Fisher LDA. The method seeks directions that maximize between-class variance relative to within-class variance, but does not adapt its decision boundary in response to downstream optimization signals. As the number of labeled samples increases, this variance-based criterion does not consistently translate into improved downstream performance as additional data are provided, limiting the extent to which additional data can be systematically exploited.

Taken together, these results position Fisher LDA as a useful geometric baseline in this setting. Its behavior highlights a central theme of this chapter: effective transfer from frozen visual representations depends not only on classifier capacity, but on whether the inductive bias of the classifier aligns with the geometry of the embedding space. In contrast to trained linear probes, which adapt directly to the downstream objective, the fixed structure imposed by LDA restricts its ability to leverage increasing supervision.

To clarify the nature of the structure captured by Fisher LDA despite its irregular learning dynamics, Figure 35 visualizes the projection of test embeddings onto the one-dimensional discriminant direction. Although LDA does not yield smoothly improving learning curves as supervision increases, the resulting projection reveals a clear separation between attentive and inattentive samples along a single axis, with the decision boundary falling in a region of low sample density. This indicates that a low-dimensional discriminative structure is present in the frozen representations, even if variance-based criteria alone do not fully exploit it under increasing data.

The behavior observed for Fisher's Linear Discriminant Analysis indicates that differences between probes cannot be explained solely in terms of classifier capacity. Instead, they reflect how the inductive bias of a probe interacts with the geometric organization of the embedding space. In particular, the irregular and non-monotonic learning curves produced by LDA suggest a misalignment between variance-based separation criteria and the task-induced structure exposed by supervised optimization. This observation motivates a direct analysis of representation geometry independent of any specific probe. In the following section, we therefore examine the embedding spaces using empirical Fisher information based spectral measures, with the aim of characterizing anisotropy, effective dimensionality, and loss sensitivity, and relating these geometric properties to the empirical behavior of different probing strategies.

5.2.4 Probing Representation Geometry with Fisher Information

This section reports Fisher-based diagnostics computed in *representation space* after training lightweight probes on embeddings extracted from frozen backbones, that is, the task-induced geometry. The goal is not to re-evaluate predictive performance, but

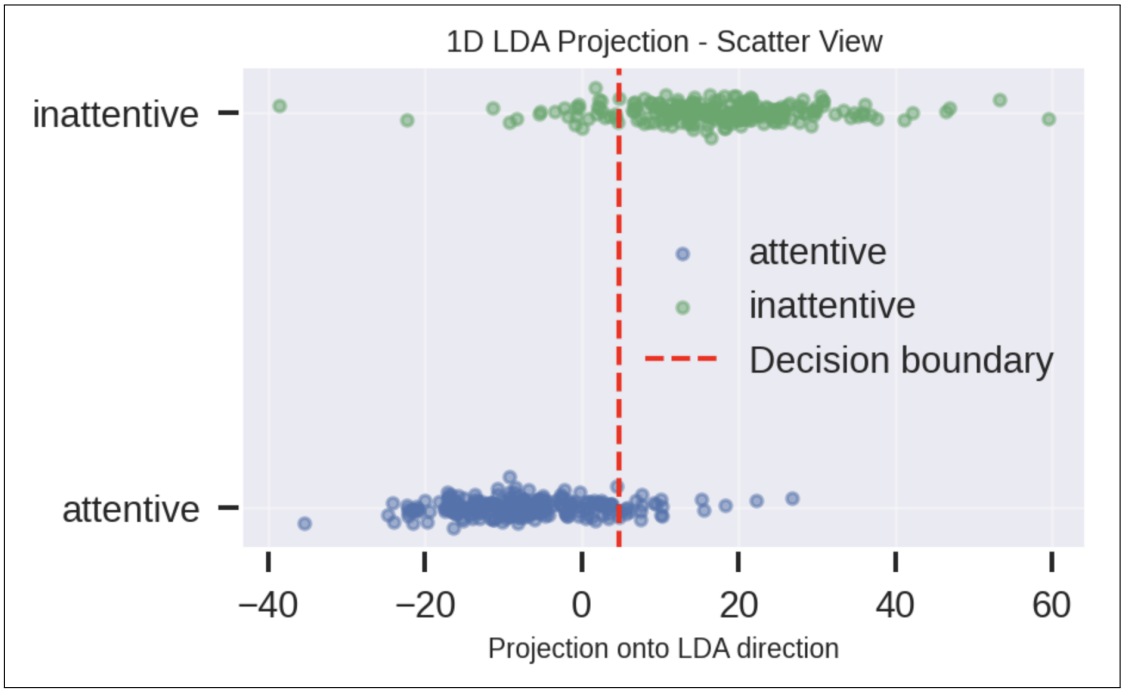


Figure 35 – Projection of test-set embeddings onto the one-dimensional Fisher discriminant direction. Despite irregular learning curves, the projection exhibits clear separation between attentive and inattentive samples, with a linear decision boundary lying in a region of low density.

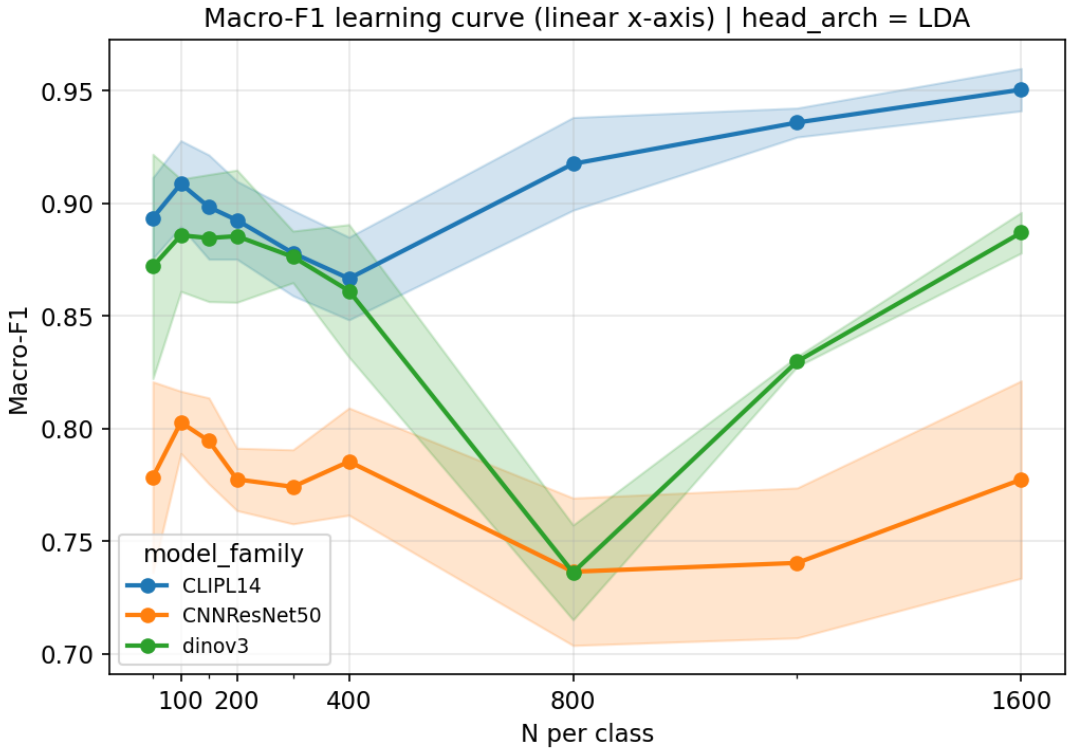


Figure 36 – Learning curve of Macro-F1 for Linear Discriminant Analysis, as a function of sample size, comparing different frozen visual backbones.

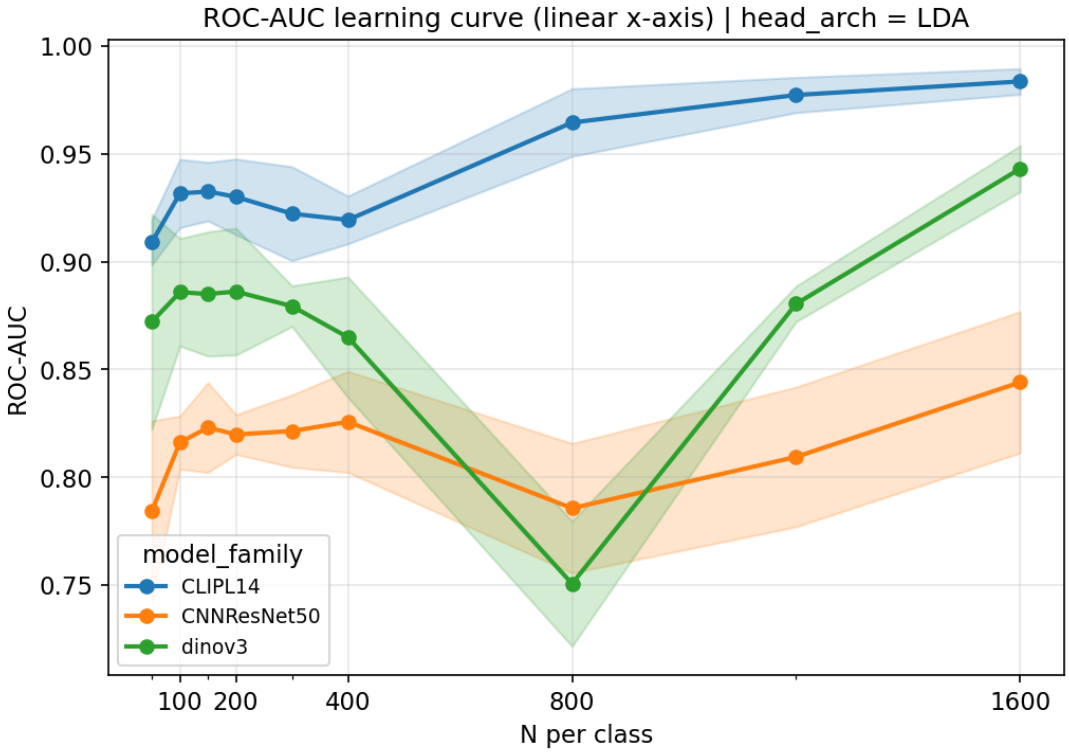


Figure 37 – Learning curve of ROC-AUC for Linear discriminant analysis, as a function of sample size, comparing different frozen visual backbones.

to quantify how the supervised objective *imprints local sensitivity* onto the embedding coordinates, and how this imprint differs across backbones and probe capacities. Complete data for train/val/test are provided in the appendixes (Tables 8–13).

5.2.4.1 Metrics and reading guide

We compute the empirical Fisher Information Matrix of the adapter's objective with respect to the embedding vectors, as a local measure of the sensitivity of a trained probe's loss to infinitesimal perturbations in representation space. The directions associated with large (steep, as opposed to flat) curvature will correspond to perturbations that strongly affect the classifier's loss, whereas directions associated with small curvature are comparatively irrelevant under the chosen probe and objective. This perspective allows us to move beyond scalar accuracy measures and examine how supervised objectives are reflected across representation dimensions.

Let $z \in \mathbb{R}^d$ denote an embedding vector and $\mathcal{L}(z)$ the loss of a trained probe. The representation-space empirical Fisher Information Matrix is defined as

$$\mathcal{I}(z) = \mathbb{E}\left[\nabla_z \mathcal{L}(z)\, \nabla_z \mathcal{L}(z)^\top\right].$$

The eigenspectrum of this matrix summarizes how loss sensitivity is distributed across directions in embedding space, capturing the magnitude and concentration of task-induced curvature.

Importantly, in binary classification settings such as the one considered here, the empirical Fisher spectrum is structurally expected to exhibit strong concentration along a single dominant direction, reflecting the effectively one-dimensional nature of the supervised signal. Accordingly, Fisher spectral analysis is not employed to infer representation quality nor evidence of semantic alignment, but to (i) confirm the expected collapse of task-induced sensitivity, and (ii) quantify its stability across data splits and adapter's architectures.

From the Fisher spectrum, we extract three complementary scalar summaries. These quantities serve as diagnostic tools for characterizing task-induced geometry under fixed probes, not as standalone indicators of semantic correctness or representational quality.

Let a trained probe produce class probabilities $p_\theta(y \mid z)$. We compute gradients with respect to z and summarize their second-order structure via (i) *trace*, (ii) *spectral concentration* $E1$, and (iii) *Shannon effective rank* (erank). Operationally:

- **trace** measures the overall magnitude of sensitivity in embedding coordinates (average squared gradient norm). Larger trace means that small perturbations in z produce larger changes in the supervised log-likelihood.
- **E1** is the fraction of total Fisher energy captured by the top eigenvalue. Values near one indicate that sensitivity is concentrated along a single dominant direction in embedding space.
- **erank** converts the normalized eigen-spectrum into an entropy-derived "effective dimensionality." Values near one indicate a near rank-1 structure; higher values indicate energy spread across multiple directions.

As $E1$/erank describe *shape* (directionality and dimensionality), trace describes *scale* (overall sensitivity). These are logically independent: the spectrum can be highly concentrated (high $E1$) at either low or high trace.

5.2.4.2 Motivation for MLP Adapter Analysis

Across all model families examined (CNN, CLIP, and DINO), the linear probe adapter exhibited remarkably consistent behavior: the effective rank of the Fisher Information Matrix was approximately equal to 1 across all data splits (train, validation, and test), across all 5 experimental repetitions, and *across all sample sizes* (ranging from 50 to 1,600 images per class). This consistency suggests that the linear probe converges to a low-dimensional solution geometry regardless of the data regime or pre-trained feature extractor.

In contrast, the shallow MLP adapter (comprising a single hidden layer with 512 units) demonstrated distinctly different characteristics. The effective rank under this

regime exhibited variation with respect to sample size and displayed different convergence properties compared to the linear probe. Given these notable differences, the subsequent analysis focuses on the behavior observed under the MLP adapter regime, as it provides insights into how additional model capacity interacts with data availability and frozen feature representations.

5.2.4.3 Top-1 spectral energy (E1): Is the task effectively one-dimensional in representation space?

Figure 38 reports the top-1 spectral energy (E1) of the Fisher spectrum as a function of the number of training samples per class. E1 measures the fraction of total Fisher energy captured by the largest eigenvalue and thus quantifies the extent to which loss sensitivity under the chosen probe and objective is concentrated along a single direction in embedding space.

In all evaluated backbones and across training sizes, the empirical Fisher spectrum with respect to embeddings collapses to an effectively rank-1 structure, confirming that the supervised attentive–inattentive objective induces a one-dimensional loss sensitivity under the chosen probes.

Geometrically, this suggests that downstream discrimination under linear or shallow probes depends predominantly on a single direction in representation space. Importantly, E1 captures concentration but not magnitude: a highly concentrated spectrum does not imply low overall sensitivity to perturbations. For this reason, E1 must be interpreted jointly with complementary Fisher-derived quantities, such as the trace and effective rank.

5.2.4.4 Shannon effective rank: How many representation dimensions meaningfully contribute to the decision?

Figure 39 shows the Shannon effective rank of the Fisher spectrum, which summarizes how many eigen-directions contribute non-negligibly to task sensitivity. While E1 reveals dominance of the leading direction, the effective rank distinguishes between strict collapse and soft dominance with residual structure.

Vision foundation models (DINOv3 and CLIP-L14) exhibit a faster and more pronounced reduction in effective rank as training data increase, converging toward nearly one-dimensional task structure. In contrast, the CNN backbone maintains a consistently higher effective rank across training sizes, which might indicate reliance on a broader set of directions to support the same task.

This pattern suggests that foundation model embeddings expose a more compact abstraction of task-relevant information, whereas CNN representations require a higher-dimensional combination of features to achieve comparable performance.

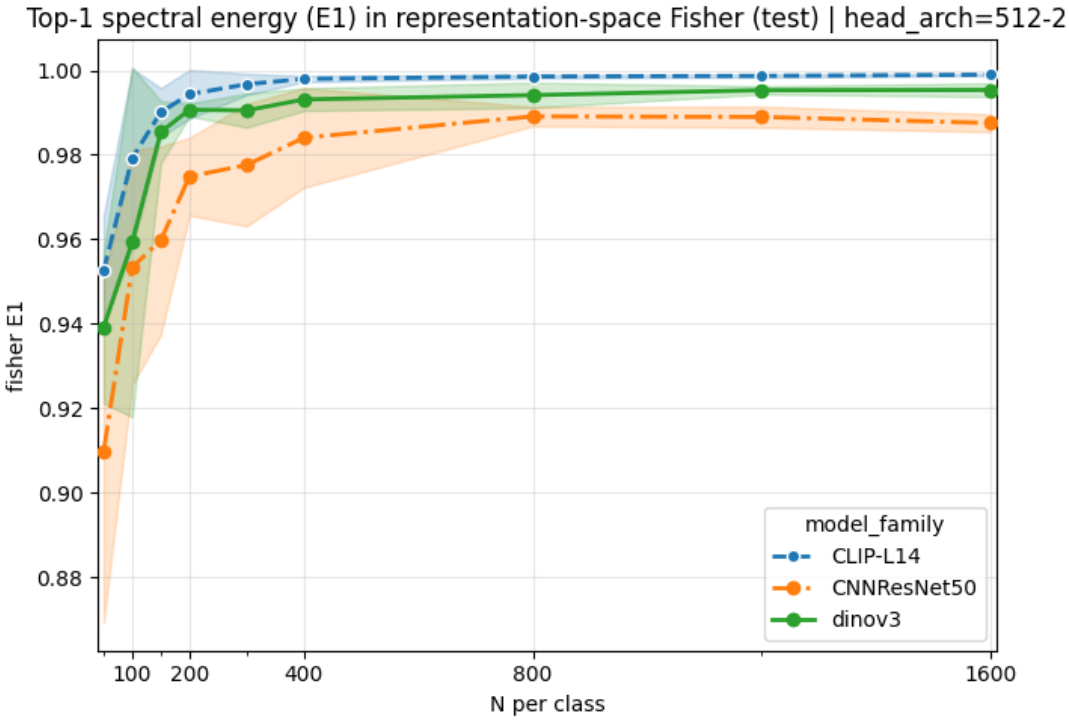


Figure 38 – Top-1 spectral energy (E1) of the representation-space Fisher Information Matrix on the test split, as a function of training set size, for different frozen visual backbones. Higher values indicate stronger concentration of Fisher information along a single dominant direction, suggesting increased anisotropy in the embedding geometry.

5.2.4.5 A dominant task direction emerges quickly and persists across backbones.

Across CLIP, DINO, and CNN embeddings, Fisher spectra become strongly anisotropic under this binary attentive–inattentive contrast, with near rank-1 behavior most clearly visible under linear probes. On the test split, linear probes already exhibit $E1 \approx 1$ at $N_{\text{per-class}} = 50$ for all three backbones (Table 8, Table 10, Table 12). This indicates that, once a linear separator is trained, the induced loss sensitivity is overwhelmingly concentrated along a single direction in embedding space.

Under the MLP probe, the same qualitative trend holds but saturation is more gradual and backbone-dependent. At $N_{\text{per-class}} = 50$, $E1$ is lower and erank is higher (e.g., CLIP MLP: $E1 \approx 0.94 - 0.97$, erank $\approx 1.27 - 1.46$; DINO MLP: $E1 \approx 0.92 - 0.95$, erank $\approx 1.39 - 1.65$; CNN MLP: $E1 \approx 0.88 - 0.95$, erank $\approx 1.43 - 2.03$ on test). As $N_{\text{per-class}}$ increases, the MLP spectra collapse as well (e.g., at $N_{\text{per-class}} = 1600$: CLIP MLP $E1 \approx 1$, erank ≈ 1.01; DINO MLP $E1 \approx 0.99 - 1$, erank $\approx 1.03 - 1.05$; CNN MLP $E1 \approx 0.99$, erank $\approx 1.08 - 1.12$). Thus, despite differences in capacity and training dynamics, the induced sensitivity ultimately concentrates into a small number of directions, often close to one.

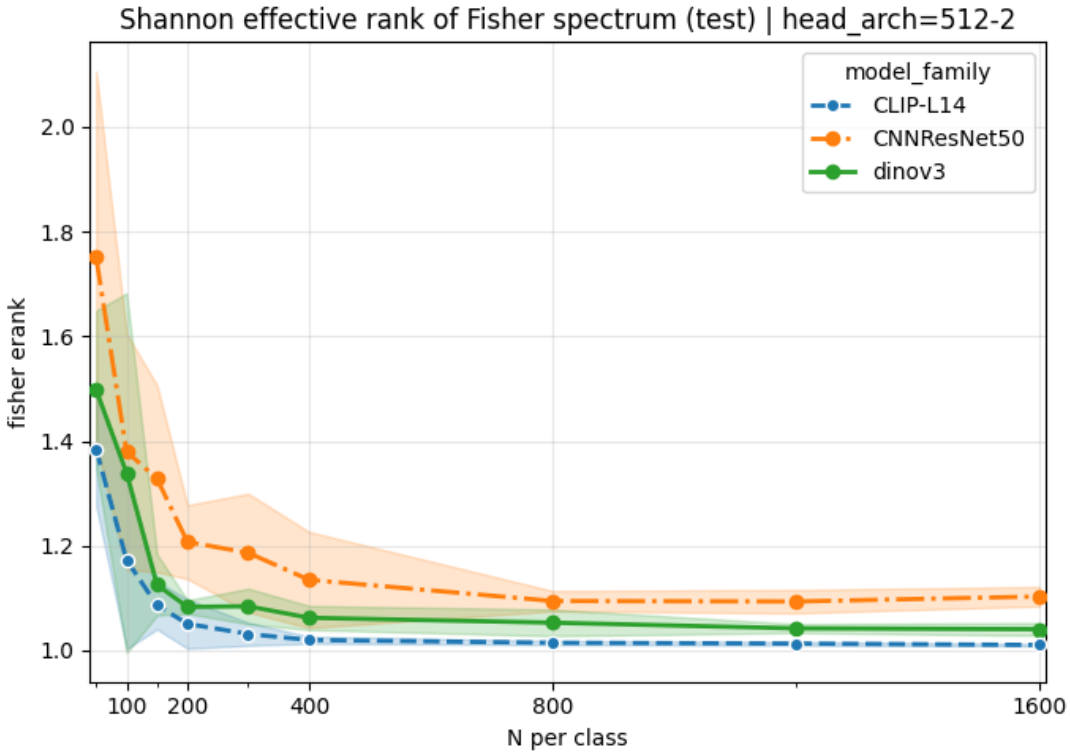


Figure 39 – Shannon effective rank of the representation-space Fisher Information spectrum on the test split, as a function of training set size, for different frozen visual backbones. The effective rank summarizes how Fisher information is distributed across spectral components, with lower values indicating greater spectral concentration.

5.2.4.6 Fisher trace: How sensitive is the classifier to perturbations in representation space?

On the interpretation of Fisher trace across backbone families. We emphasize that the Fisher trace is a scale-dependent quantity: its magnitude depends on embedding norms, dimensionality, and other factors. As a result, absolute comparisons of trace values across different backbone families must be interpreted with caution, as they may reflect differences in representation scale which might not be interpreted as intrinsic representational quality. In this work, trace is therefore used primarily as a comparative diagnostic under a fixed probing protocol, and its interpretation is guided jointly by scale-free metrics such as spectral concentration (E1) and effective rank. Within this context, Fisher trace provides meaningful insight into how sharply different representations respond to perturbations, while avoiding over-interpretation of absolute cross-family differences. We deliberately avoid additional embedding normalization in this analysis in order to probe the representation geometry as exposed natively by each frozen backbone under a fixed downstream probing protocol, by not imposing a common scale that may obscure model-specific inductive biases.

Figure 40 reports the trace of the representation-space Fisher matrix for all backbones under the MLP regime, corresponding to the total curvature magnitude aggregated

across all directions. Unlike E1 and effective rank, which characterize the shape of the spectrum, the trace quantifies overall sensitivity of the loss to perturbations in the embeddings.

An apparent separation seems to emerge, which must be taken with a grain of salt. From the figure, it seems that CNN-based representations exhibit substantially larger Fisher trace values, growing rapidly with the number of training samples, whereas DINOv3 and CLIP-L14 seem to maintain markedly lower trace throughout. We must emphasize, though, that despite the appearance of very high disparity, the upwards evolution of the metric along the *N per class* axis represents an increase of 3,313 % as opposed to 1,893 % (CLIP) and 1,578 % (DINO). As reference, the behavior of the same metric under the linear probe is reflected as a 0,4 % increase (CNN), a decrease of 22.7 % (CLIP), and a decrease of 27.6 % (DINO).

Despite the differences of scale, and abstracting the absolute values, this pattern seems to be consistent with CNN-based representations exhibiting sharper loss landscapes under the fixed probing setup considered here, in the sense that small perturbations in representation space induce larger changes in loss. Hence, the joint observation of higher trace and higher effective rank might be consistent with the lower sample efficiency and generalization observed for CNNs in the learning curves, although this relationship should be interpreted as diagnostic and not causal.

5.2.4.7 Interpreting the convergence of effective rank as a function of sample size in the MLP regime

As previously noted, under the linear probe adapter, regardless of sample size, all models collapsed to an effective rank of 1 dimension. This behavior was not followed by the results when training was performed for the MLP adapter. As the sample size increased, on average, the effective ranks went from 1.75 to 1.1 (CNN), 1.39 to 1.01 (CLIP) and 1.5 to 1.04 (DINO).

This divergent behavior between adapter architectures seems to reveal properties about the solution geometry and the interaction between model capacity and data availability. The linear probe's immediate convergence to rank ≈ 1 suggests that the frozen feature representations already encode class-discriminative information along a single dominant direction in parameter space. The pre-trained embeddings, despite their high dimensionality, possess an intrinsic low-dimensional structure for this classification task that can be accessed directly through linear transformation.

In contrast, the MLP adapter's trajectory from higher effective rank toward unity suggests that additional model capacity, when combined with limited data, leads to exploration of a higher-dimensional solution space. This phenomenon can be interpreted through the lens of data-driven regularization: with fewer samples per class (under 400,

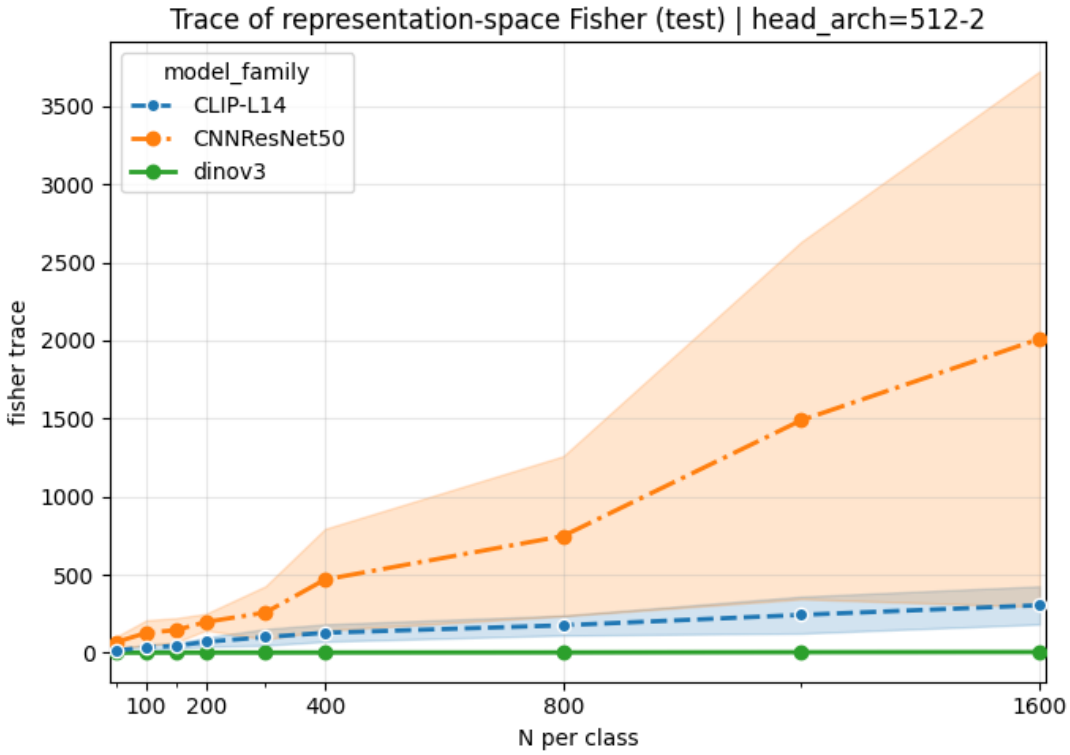


Figure 40 – Trace of the Fisher Information Matrix computed in representation space on the test split, as a function of training set size, for different frozen visual backbones. The trace measures the total Fisher information accumulated across all representation dimensions, providing a global summary of the sensitivity of the classifier to perturbations in the embedding space.

in this work), the MLP distributes its learning across multiple parameter dimensions, potentially capturing noise, not signal. As sample size increases, the abundance of data acts as an implicit regularizer, progressively constraining the solution toward the same low-dimensional manifold that the linear probe identifies immediately. The convergence of effective rank quantifies the sample complexity required for the MLP to discover the underlying geometric structure of the task.

Importantly, this geometric convergence occurs independently of the quality of the underlying features. While CLIP and DINO representations achieve higher classification performance than CNN (approximately 10 percentage points in Macro-F1), all three architectures exhibit the same rank ≈ 1 solution geometry under the linear probe. This dissociation between feature quality and solution dimensionality suggests that effective rank captures the geometric structure of learned solutions, not necessarily their discriminative power. The pre-trained features differ in how well their dominant discriminative axis aligns with true class boundaries, not in the dimensionality required to access that information.

The practical implications of these findings seem to be significant. First, the MLP's convergence pattern provides a geometric diagnostic of model-data fit that is independent of accuracy metrics. When effective rank remains elevated despite increasing sample size, it

may indicate that the adapter is exploiting additional feature dimensions or learning non-linear combinations that deviate from the dominant solution. Conversely, rapid convergence to rank ≈ 1 suggests that the frozen features are well-suited for linear separability, and additional model capacity offers minimal geometric advantage.

Second, these observations inform architecture selection as a function of data regime. In low-data scenarios (fewer than 400 samples per class, for instance), the linear probe's immediate identification of the low-dimensional solution makes it preferable to the MLP, which disperses learning across unnecessary dimensions. The MLP only becomes geometrically comparable when sufficient data is available to constrain its solution space. This suggests that for transfer learning with frozen features, model capacity should scale with data availability—not merely to prevent overfitting in the traditional sense, but to avoid geometric inefficiency in parameter space.

Finally, the uniform convergence to rank ≈ 1 across all model families validates a fundamental property of the task: given these pre-trained representations, the classification problem is intrinsically low-dimensional. This finding suggests that more complex adapter architectures are unlikely to provide geometric benefits, as the solution manifold is already one-dimensional regardless of the feature extractor. Any performance gains from deeper adapters would need to arise from non-linear feature combinations rather than from accessing higher-dimensional structure, as that structure does not appear to be present in the Fisher information geometry.

5.2.4.8 Operational use: Fisher as a representation audit

Beyond descriptive analysis, Fisher metrics can be operationalized as a lightweight *representation audit* applicable to any frozen backbone and downstream task. The audit answers three concrete questions using minimal supervision and without end-to-end fine-tuning.

5.2.4.8.1 Audit Question 1: Is the task already geometrically aligned with the representation?

This is assessed via spectral concentration. A near rank-1 Fisher spectrum (high $E1$, erank ≈ 1) under a linear probe indicates that the supervised objective aligns with an existing direction in embedding space. In such cases, downstream performance is expected to saturate quickly and to exhibit limited sensitivity to increased probe capacity.

5.2.4.8.2 Audit Question 2: Does additional capacity reveal structure or merely amplify it?

This is assessed by comparing linear and MLP probes. If increasing probe capacity substantially increases trace while leaving $E1$ and erank largely unchanged, additional capacity is amplifying sensitivity along an already identified axis, and not uncovering new

discriminative directions. Conversely, concurrent changes in $E1$/erank signal compensatory learning.

5.2.4.8.3 Audit Question 3: How energetically is the task encoded?

This is assessed via trace. Representations with comparable $E1$ but very different trace values agree on *where* the task lives in embedding space while disagreeing on *how strongly* it is expressed. This distinction provides insight into optimization stability, probe dependence, and sensitivity to perturbations, and is invisible to accuracy-based evaluation alone. The comparisons should be restricted as to insure equal footing.

5.2.4.8.4 Minimal protocol

In practice, the audit requires only: (i) a frozen backbone, (ii) a lightweight probe, (iii) gradients of the log-likelihood with respect to embeddings on held-out data. The resulting Fisher spectrum provides a compact summary of task alignment and sensitivity that is applicable in low-data regimes.

5.2.5 Summary of Findings

Across all experimental settings, downstream performance is dominated by the choice of frozen visual backbone as opposed to the capacity of the classification head. Vision foundation models consistently achieve higher Macro-F1 and ROC-AUC than the convolutional baseline across training sizes, suggesting superior sample efficiency under identical probing protocols.

When conditioning on a fixed backbone, increasing probe capacity from a linear classifier to a shallow MLP produces, at most, marginal and inconsistent gains. Paired difference and win-rate analyses show no systematic advantage for non-linear probes, suggesting that additional head expressivity does not reliably exploit structure beyond what is already linearly accessible in the embedding space.

Geometric analysis using Fisher-based diagnostics further clarifies this observation. High spectral concentration (E1), rapidly collapsing effective rank, and comparatively low Fisher trace for foundation model embeddings jointly seem to indicate that task sensitivity concentrates along a small number of stable directions, with limited curvature elsewhere. In contrast, the convolutional backbone exhibits higher effective dimensionality and larger Fisher trace, in what seems to be consistent with sharper and more fragile decision boundaries.

Taken together, these results might indicate that the observed performance differences do not arise from downstream classifier complexity, but primarily from differences

in representation geometry induced by the frozen backbones. Probe capacity modulates performance only weakly once a sufficiently structured representation is available.

5.3 Results with an Out-of-Domain Dataset

The dataset constructed in this work consists of synthetic images of K–12 students in classroom settings depicting states of apparent (in)attention. These states are operationalized through a set of observable behavioral proxies, such as gaze orientation, posture, and task engagement (see Table 3, Table 4, Table 5, Table 6). While the primary objective of the dataset is to support controlled experimentation under known generative conditions, it is important to evaluate whether the regularities learned from such synthetic data extend beyond its original domain. Despite domain adaptation being considered "one of the most challenging tasks of modern data analytics" (COURTY et al., 2016), an additional experiment was conducted using an out-of-domain (OOD) dataset.

In this context, out-of-domain refers to a shift in several fundamental aspects of the data distribution, including the use of real rather than synthetic imagery, differences in demographic characteristics of the subjects, variations in recording conditions, and distinct annotation protocols. Such shifts are known to present significant challenges for visual recognition systems, as models often learn statistical regularities tied to the training distribution rather than invariant semantic features. Consequently, evaluating performance under OOD conditions provides a more stringent assessment of the robustness and transferability of the learned representation.

As an initial exploratory step, we also tested the trained model on a very small set of real, non-synthetic classroom images collected from external sources. These images closely resembled those in the training set in terms of visual composition, posture and activities performed as in the aforementioned tables (v.g. depicting a student seated at a desk, reading a book or raising her hand). Although the results appeared qualitatively promising (with perfect scores), the test set contained only approximately ten images per class, rendering the outcome purely anecdotal. A systematic evaluation using real classroom imagery is therefore deferred to future work.

For the present experiment, we selected the DIPSER dataset (M.-CARPINTERO et al., 2025) as the external evaluation source. This dataset contains recordings of adult students engaged in learning activities similar to those performed in online education, captured through webcam footage that primarily includes the face and upper torso of the participants. Although the age group and acquisition setup differ substantially from those in our synthetic dataset, the behavioral signals relevant to attention—particularly gaze direction and upper-body posture—remain partially comparable. This makes DIPSER a reasonable candidate for assessing cross-domain generalization.

The DIPSER dataset is organized into three subject groups. For practical reasons related to storage and processing constraints, only Group 1 was downloaded and processed. This subset includes all nine experiments conducted in the dataset and comprises recordings of twenty subjects. The downloaded portion alone amounts to more than 250 GB of data and contains approximately 885,000 image frames.

According to the dataset description, "[l]abels are meticulously provided for each subject, with a dedicated file indicating the frames, levels of attention, and emotions assessed. Attention is measured on a scale from one to five, with 5 indicating maximum attention."

From the available metadata, 295 label files were processed, corresponding to 7,750 annotated entries. Among these, 4,694 frames included an explicit attention label. The distribution across attention levels was as follows:

- "1": 7
- "2": 2,093
- "3": 2,308
- "4": 335
- "5": 221

To construct a binary classification task compatible with the model trained on our dataset, the intermediate level (3) was discarded, as it represents an ambiguous or neutral state. Levels 1 and 2 were grouped as "inattentive," yielding 2,100 images, while levels 4 and 5 were grouped as "attentive," yielding 556 images.

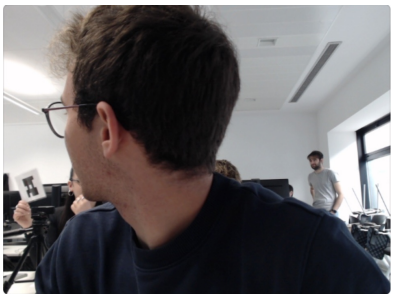


Figure 41 – labeled as attentive, level 5

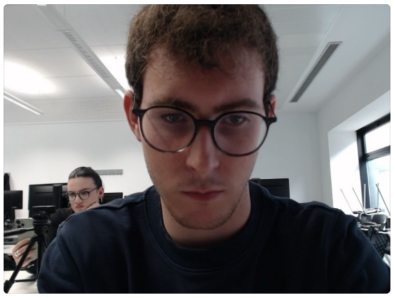


Figure 42 – labeled as inattentive, level 2

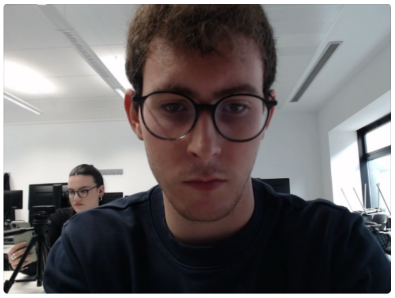


Figure 43 – labeled as attentive, level 5

All candidate images were subsequently subjected to manual verification by the author. The goal of this step was not to relabel the data but to ensure that the selected images clearly exhibited the intended behavioral proxies. From this process, 133 images per class were retained, resulting in a balanced evaluation set of 266 images. Images labeled as



Figure 44 – label from json file

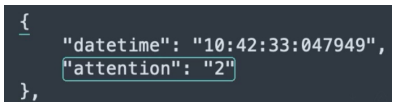


Figure 45 – label from json file

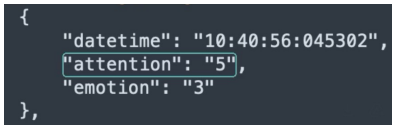


Figure 46 – label from json file

attentive generally depict subjects whose gaze is directed toward the screen or another focal object associated with the task, whereas inattentive images show subjects looking away or displaying body orientations inconsistent with task engagement.

During this inspection, several inconsistencies in the original labeling scheme became evident. In some cases, subjects assigned the highest attention level are visibly oriented away from the screen or toward the back of the room, as illustrated in Figure 41. Conversely, frames that appear visually almost identical may receive opposite labels, as shown in Figures 42 and 43. These observations highlight the subjective and context-dependent nature of the annotation process, which may rely on contextual information not directly visible in the frame itself.

Using this curated subset, we evaluated the linear probe trained on the largest available training split of the synthetic dataset developed in this work. The results are summarized below:

- Inattentive precision/recall/F1: 0.5085 / 0.9023 / 0.6504
- Attentive precision/recall/F1: 0.5667 / 0.1278 / 0.2086
- Overall/Balanced accuracy: 0.5150

These results indicate that the classifier performs close to chance level under the OOD condition. However, the confusion structure (Figure 47) reveals an important asymmetry. The model exhibits very high recall for the inattentive class (0.9023), meaning that most inattentive examples are correctly identified. In contrast, recall for the attentive class is extremely low (0.1278), indicating that the majority of attentive images are misclassified as inattentive.

This pattern suggests that the learned decision boundary primarily captures strong visual cues associated with inattentive behavior, such as pronounced head rotation, gaze deviation, or body posture inconsistent with screen engagement. These signals appear to be relatively robust across domains. By contrast, attentive behavior is visually more heterogeneous and context-dependent, making it harder to identify reliably when the training and testing distributions differ substantially.

Another factor likely contributing to the degradation in performance is the mismatch between the behavioral proxies encoded in the synthetic dataset and the annotation protocol

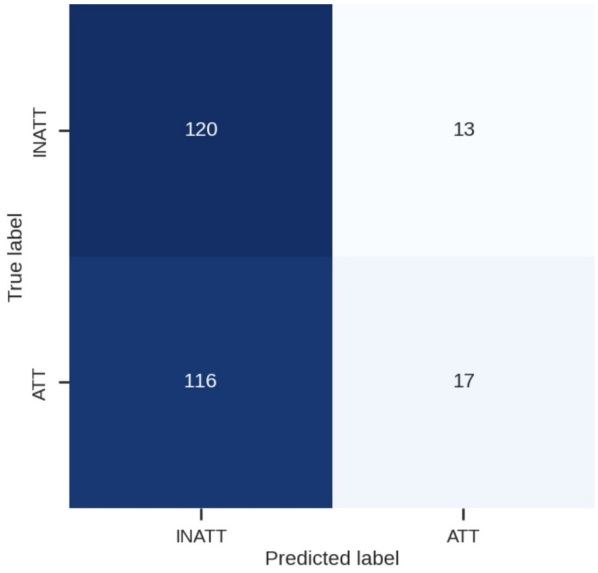


Figure 47 – Confusion Matrix (Counts)

used in DIPSER. In our dataset, attention is operationalized through clearly visible visual cues embedded in the prompt generation process. In DIPSER, however, the attention score reflects a subjective judgment that may incorporate contextual information not directly observable in a single frame. As a result, the mapping between visual appearance and attention label may be weaker or noisier.

Taken together, these results suggest that the poor performance observed on DIPSER should not be interpreted primarily as a limitation of training on synthetic data. As mentioned, in preliminary experiments involving real classroom photographs that preserved the visual structure of the synthetic dataset, the trained model achieved perfect classification scores, albeit on a very small sample. This indicates that moderate domain shift, such as the transition from synthetic to real imagery, does not necessarily impair the learned representation.

By contrast, the DIPSER dataset introduces several sources of distributional shift simultaneously. These include substantial demographic differences (adult learners rather than K–12 students), a distinct acquisition setup (webcam recordings rather than classroom scenes), different task contexts (online learning sessions), and a labeling protocol based on subjective attention scores rather than directly observable behavioral proxies. From the perspective of representation learning, these changes alter both the visual statistics of the images and the semantic meaning associated with the labels.

Consequently, the observed performance—characterized by near-chance balanced accuracy and a strong bias toward the inattentive class—should be interpreted as evidence of a severe out-of-distribution condition rather than a failure of the synthetic training paradigm itself. In other words, the experiment highlights the limits of cross-domain generalization when the target distribution diverges substantially from the visual and

semantic assumptions encoded during training.

Future work should therefore aim to quantify and disentangle the sources of distributional shift between datasets, both at the level of observable image statistics and within the representation spaces produced by modern Vision Foundation Models (VFMs). This investigation is particularly relevant in the current paradigm of Computer Vision, where pretrained encoders are widely reused as fixed feature extractors and downstream performance depends critically on the structure of their embedding spaces. Accordingly, further research should examine whether attention-related behavioral signals—such as gaze direction, head orientation, and upper-body posture—are consistently represented within these latent spaces across datasets with different visual and contextual characteristics. Controlled cross-dataset evaluations, combined with analyses of embedding geometry and representation invariances, may help determine to what extent such behavioral proxies form stable semantic directions within VFM representations, and under which conditions these directions remain transferable across domains.

6 Conclusions

This work set out to investigate whether visually apparent attentional configurations—operationalized through observable behavioral cues such as posture, gaze direction, and object interaction—are meaningfully encoded in the embedding spaces of contemporary Vision Foundation Models. The study approached attention as a semantic hypothesis about structure already present in pretrained visual representations, as opposed to framing attention as a target variable to be learned through end-to-end optimization.

The empirical evidence presented throughout this dissertation is consistent with a coherent interpretation: visually apparent attention behaves as a low-rank, readily accessible readout phenomenon in modern vision embeddings. Across models, dataset sizes, and probing configurations, attentional contrasts could be recovered with minimal adaptation, using frozen backbones and lightweight probes. Performance saturated early as the number of training samples increased, and linear or near-linear probes sufficed to extract the relevant signal. Taken together, these findings suggest that the task-relevant structure required to separate attentive from inattentive configurations is largely pre-encoded in the representation space, rather than being constructed through task-specific training.

This result has important interpretive implications. Apparent attention does not emerge in these models as a localized feature tied to a specific patch or visual primitive, but as a distributed pattern that collapses onto a small number of dominant directions when subjected to task-driven readout. In this sense, attention-related information appears organized compactly and stably within the embedding space, rather than requiring a complex, high-dimensional decision boundary. The consistency of this behavior across different probing methods suggests that the phenomenon is more likely a property of the representations themselves than an artifact of probe capacity or optimization strategy — though definitive causal attribution would require interventional analyses beyond the scope of this work.

A central contribution of this work lies in the way synthetic data and Vision-Language Models were employed to make this structure observable. Synthetic images were not introduced as substitutes for real-world data, nor as a means of improving classification performance. Instead, they functioned as controlled perturbations of the visual input space, enabling deliberate variation of attentional cues while limiting uncontrolled sources of variability. Vision–Language Models, in turn, were used strictly as auxiliary instruments to organize this synthetic data, providing operational supervision through aggregated judgments, which are not construed as authoritative semantic labels. This instrumental

configuration allowed the analysis to focus on how hypothesized attentional distinctions propagate through pretrained visual representations, without conflating representational structure with claims about ground truth attentional states.

The geometric analyses conducted in this study, including Linear Discriminant Analysis and exploratory Fisher-based diagnostics, provided complementary perspectives on these findings. While aware of the limitations inherent in such tools, their application helped characterize the anisotropy, conditioning, and effective dimensionality of task-relevant variation in embedding space. Importantly, these analyses reinforced the central readout-based conclusion rather than replacing it: they served to illuminate how and where sensitivity to attentional distinctions is concentrated, not to assert a definitive geometric theory of attention representation.

At the same time, the scope of the present work is deliberately bounded. This work makes no claim that the models studied here detect or represent attention as a latent cognitive state. The operational labels derived from Vision–Language Models are not treated as ground truth, nor are they assumed to reflect the true attentional condition of depicted subjects. The study does not propose a deployable system for attention monitoring, nor does it address fairness, demographic bias, or representativeness in a systematic way. Furthermore, the low-rank behavior observed for apparent attention should not be assumed to generalize to all semantic or behavioral attributes without empirical verification.

Within these bounds, the contribution of this dissertation is to demonstrate, through a concrete and socially relevant case study, how a representation-first experimental approach can reveal what contemporary vision models already encode. By combining controlled synthetic perturbations, auxiliary model-based organization, constrained probing, and geometric inspection, the work reframes attention not as a label to be predicted, but as a concept against which representation structure can be tested. In doing so, it illustrates how certain behavioral distinctions — at least in their visually apparent form — may already be readily accessible within pretrained embedding spaces, requiring little more than appropriate interrogation to become visible.

The findings presented here suggest that progress in understanding Vision Foundation Models may depend as much on how questions are posed to representations as on how models are trained. Attention, treated as a semantic hypothesis - not as an engineering objective - provides one such question. The broader implication is that representation analysis, when grounded in controlled experimental design and carefully bounded interpretation, can serve as a powerful lens for examining the latent structure of modern visual models, and is suggested as future work.

Several directions follow naturally from the findings reported here. First, the low-rank structure identified in attention-related embeddings invites investigation into whether analogous phenomena arise for other semantically grounded attributes that

manifest through observable visual cues — most notably emotional states, for which similar geometric regularities have been suggested in the affect recognition literature. If linearly accessible structure is a general property of semantically coherent categories in vision embeddings, rather than a peculiarity of attentional configurations, this would have broader implications for the probing paradigm as a diagnostic tool. Second, the reliance on synthetic data as an experimental probe raises the question of transferability: to what extent do the geometric regularities observed here persist when probing is conducted on naturalistic data with real subjects, where illumination, occlusion, and demographic variation introduce uncontrolled sources of noise? Third, the dominant Fisher directions identified in this work could serve as the basis for interpretability analyses, linking geometric structure in embedding space back to specific visual features through gradient-based or attention-based attribution methods — potentially clarifying which visual primitives drive the separation between attentive and inattentive configurations. Finally, extending the probing framework to multimodal and video-based representations would allow the temporal dynamics deliberately excluded from this work to be reintroduced in a controlled manner, enabling a direct assessment of what is gained — and what is already captured — by static visual representations alone.

Bibliography

ACHILLE, A. et al. *Task2Vec: Task Embedding for Meta-Learning*. 2019. Disponível em: <https://arxiv.org/abs/1902.03545>. Citado 4 vezes nas páginas 15, 16, 20, and 40.

ACHILLE, A.; SOATTO, S. Information dropout: Learning optimal representations through noise. *IEEE TPAMI*, 2018. Citado 2 vezes nas páginas 21 and 26.

ALAIN, G.; BENGIO, Y. Understanding intermediate layers using linear classifier probes. In: *International Conference on Learning Representations (ICLR)*. [S.l.: s.n.], 2017. Citado 3 vezes nas páginas 20, 21, and 26.

AMARI, S.-i. Natural gradient works efficiently in learning. *Neural Computation*, v. 10, n. 2, p. 251–276, 1998. Citado 3 vezes nas páginas 15, 21, and 40.

AMARI, S.-i. *Information Geometry and Its Applications*. [S.l.]: Springer, 2016. Citado na página 39.

AMARI, S.-i.; KARAKIDA, R.; OIZUMI, M. Fisher information and natural gradient learning in random deep networks. In: CHAUDHURI, K.; SALAKHUTDINOV, R. (Ed.). *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*. Naha, Okinawa, Japan: PMLR, 2019. (Proceedings of Machine Learning Research, v. 89), p. 694–702. Citado na página 15.

BAI, S. et al. *Qwen2.5-VL Technical Report*. 2025. Disponível em: <https://arxiv.org/abs/2502.13923>. Citado 2 vezes nas páginas 13 and 26.

BAU, D. et al. Network dissection: Quantifying interpretability of deep visual representations. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.]: IEEE, 2017. p. 6541–6549. Citado na página 21.

BENITEZ-QUIROZ, C. F.; SRINIVASAN, R.; MARTINEZ, A. M. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In: *Proceedings of IEEE International Conference on Computer Vision & Pattern Recognition (CVPR16)*. Las Vegas, NV, USA: [s.n.], 2016. Citado na página 21.

BERARDINO, A. et al. *Eigen-Distortions of Hierarchical Representations*. 2018. Disponível em: <https://arxiv.org/abs/1710.02266>. Citado na página 15.

BERKA, C. et al. Eeg correlates of task engagement and mental workload in vigilance, learning, and memory tasks. *Aviation, Space, and Environmental Medicine*, v. 78, n. 5 Suppl, p. B231–B244, May 2007. ISSN 0095-6562. PMID: 17547324. Citado na página 19.

BOSCH, N. Detecting student engagement: Human versus machine. In: *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*. New York, NY, USA: Association for Computing Machinery, 2016. (UMAP '16), p. 317–320. ISBN 9781450343688. Disponível em: <https://doi.org/10.1145/2930238.2930371>. Citado 2 vezes nas páginas 19 and 26.

BRASIL. *Lei Geral de Proteção de Dados Pessoais (LGPD)*. 2018. Lei n.º 13.709, de 14 de agosto de 2018. Disponível em: <https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/L13709compilado.htm>. Citado na página 24.

BUOLAMWINI, J.; GEBRU, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In: FRIEDLER, S. A.; WILSON, C. (Ed.). *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. PMLR, 2018. (Proceedings of Machine Learning Research, v. 81), p. 77–91. Disponível em: <https://proceedings.mlr.press/v81/buolamwini18a.html>. Citado 2 vezes nas páginas 25 and 30.

BUSTOS, C. et al. *On the use of Vision-Language models for Visual Sentiment Analysis: a study on CLIP*. 2023. Disponível em: <https://arxiv.org/abs/2310.12062>. Citado na página 22.

CABACAS-MASO, J. et al. *Enhancing Facial Expression Recognition through Dual-Direction Attention Mixed Feature Networks and CLIP: Application to 8th ABAW Challenge*. 2025. Disponível em: <https://arxiv.org/abs/2503.12260>. Citado na página 22.

CARON, M. et al. Emerging properties in self-supervised vision transformers. In: *ICCV*. [S.l.: s.n.], 2021. Citado 2 vezes nas páginas 11 and 21.

COHN, J. F.; SCHMIDT, K. L. The timing of facial motion in posed and spontaneous smiles. *International Journal of Wavelets, Multiresolution and Information Processing*, v. 2, n. 02, p. 121–132, 2004. Citado na página 20.

COURTY, N. et al. *Optimal Transport for Domain Adaptation*. 2016. Disponível em: <https://arxiv.org/abs/1507.00504>. Citado na página 77.

DELATTRE, M.; KUHN, E. Computing an empirical fisher information matrix estimate in latent variable models through stochastic approximation. *Computo*, v. 5, p. 1–25, 2023. Disponível em: <https://computo-journal.org/published-202311-delattre-fim/>. Citado 2 vezes nas páginas 17 and 40.

DHALL, A. et al. Emotion recognition in the wild challenge 2013. In: *Proceedings of the 15th ACM International Conference on Multimodal Interaction*. [S.l.]: ACM, 2013. p. 509–516. Citado na página 21.

DOSOVITSKIY, A. et al. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020. Disponível em: <https://arxiv.org/abs/2010.11929>. Citado 3 vezes nas páginas 5, 11, and 21.

D'MELLO, S.; GRAESSER, A. Dynamics of affective states during complex learning. *Learning and Instruction*, v. 22, n. 2, p. 145–157, 2012. ISSN 0959-4752. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0959475211000806>. Citado 2 vezes nas páginas 19 and 20.

EKMAN, P.; FRIESEN, W. V. Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, v. 17, n. 2, p. 124, 1971. Citado 2 vezes nas páginas 20 and 21.

EKMAN, P.; FRIESEN, W. V. *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. [S.l.]: Consulting Psychologists Press, 1978. Citado na página 20.

EMAM, K. E. *Accelerating AI with Synthetic Data*. Sebastopol, CA, USA: O'Reilly Media, Inc., 2020. Copyright © 2020 O'Reilly Media, Inc. All rights reserved. Citado 3 vezes nas páginas 23, 24, and 25.

EMAM, K. E. et al. *Practical Synthetic Data Generation*. [S.l.]: O'Reilly, 2020. Citado na página 26.

FEATHER, J. et al. *Discriminating image representations with principal distortions*. 2025. Disponível em: <https://arxiv.org/abs/2410.15433>. Citado 2 vezes nas páginas 15 and 16.

FISHER, R. A. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, v. 7, n. 2, p. 179–188, 1936. Citado na página 13.

GEIRHOS, R. et al. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, Nature Publishing Group, v. 2, n. 11, p. 665–673, 2020. Citado 2 vezes nas páginas 25 and 26.

GOG, T. V.; SCHEITER, K. Eye tracking as a tool to study and enhance multimedia learning. *Learning and Instruction*, Elsevier, v. 20, n. 2, p. 95–99, 2010. Citado na página 19.

GOODFELLOW, I. J. et al. Challenges in representation learning: A report on three machine learning contests. *Neural Networks*, v. 64, p. 59–63, 2015. Citado na página 21.

GRAFSGAARD, J. F. et al. Automatically recognizing facial expression: Predicting engagement and frustration. In: *Proceedings of EDM*. [S.l.: s.n.], 2013. p. 43–50. Citado 2 vezes nas páginas 19 and 20.

GROSS, R. et al. Multi-pie. *Image and Vision Computing*, v. 28, n. 5, p. 807–813, 2010. Citado na página 20.

HALKO, N.; MARTINSSON, P.-G.; TROPP, J. A. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, v. 53, n. 2, p. 217–288, 2011. Citado na página 17.

HE, K. et al. Momentum contrast for unsupervised visual representation learning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.]: IEEE, 2020. p. 9729–9738. Citado na página 21.

JAMES, W. *The Principles of Psychology, Volume 1*. Public Domain, 2018. Originally published in 1890. Disponível em: <https://www.gutenberg.org/ebooks/57628>. Citado na página 1.

KORNBLITH, S. et al. *Similarity of Neural Network Representations Revisited*. 2019. Disponível em: <https://arxiv.org/abs/1905.00414>. Citado na página 20.

LIN, M.; CHEN, Q.; YAN, S. *Network In Network*. 2014. Disponível em: <https://arxiv.org/abs/1312.4400>. Citado na página 12.

LIU, H. et al. *Visual Instruction Tuning*. 2023. Disponível em: <https://arxiv.org/abs/2304.08485>. Citado 3 vezes nas páginas 5, 13, and 26.

LUCEY, P. et al. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. [S.l.]: IEEE, 2010. p. 94–101. Citado na página 20.

LYONS, M. et al. Coding facial expressions with gabor wavelets. In: *Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition*. [S.l.]: IEEE, 1998. p. 200–205. Citado na página 20.

M.-CARPINTERO, L. et al. *DIPSER: A Dataset for In-Person Student Engagement Recognition in the Wild*. 2025. Disponível em: <https://arxiv.org/abs/2502.20209>. Citado na página 77.

MANCAS, M. et al. (Ed.). *From Human Attention to Computational Attention: A Multidisciplinary Approach*. New York: Springer, 2016. ISBN 978-1-4939-3433-1. Disponível em: <https://link.springer.com/book/10.1007/978-1-4939-3435-5>. Citado na página 1.

MARTENS, J. New insights and perspectives on the natural gradient method. *Journal of Machine Learning Research*, v. 21, n. 146, p. 1–76, 2020. Citado 3 vezes nas páginas 15, 39, and 40.

MAVADATI, S. M. et al. Disfa: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, v. 4, n. 2, p. 151–160, 2013. Citado na página 20.

MCDUFF, D. et al. Affectiva-mit facial expression dataset (am-fed): Naturalistic and spontaneous facial expressions collected. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. [S.l.: s.n.], 2013. p. 881–888. Citado na página 20.

MOLLAHOSSEINI, A. et al. Facial expression recognition from world wild web. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. [S.l.: s.n.], 2016. Citado na página 21.

MOLLAHOSSEINI, A. et al. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, PP, n. 99, p. 1–1, 2017. Citado na página 20.

MURPHY-CHUTORIAN, E.; TRIVEDI, M. M. Head pose estimation in computer vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 31, n. 4, p. 607–626, 2009. Citado na página 20.

NEZAMI, O. M. et al. *Automatic Recognition of Student Engagement using Deep Learning and Facial Expression*. 2019. Disponível em: <https://arxiv.org/abs/1808.02324>. Citado na página 19.

Ordem dos Advogados do Brasil, Seccional do Rio Grande do Norte. *Cartilha: A Escola e o Direito de Imagem de Crianças e Adolescentes*. 2023. Copyright © Ordem dos Advogados do Brasil, 2023. Disponível em: <https://oabrn.org.br/userfiles/files/Cartilha%20Direito%20de%20Imagem%20das%20Criancas%201_compressed.pdf>. Citado na página 24.

PANTIC, M. et al. Web-based database for facial expression analysis. In: *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*. [S.l.]: IEEE, 2005. p. 5–pp. Citado na página 20.

PAPYAN, V.; HAN, X.; DONOHO, D. L. Traces of class/cross-class structure pervade deep learning spectra. *Journal of Machine Learning Research*, v. 21, n. 32, p. 1–64, 2020. Citado na página 21.

PASCANU, R.; BENGIO, Y. Revisiting natural gradient for deep networks. *arXiv preprint arXiv:1301.3584*, 2013. Citado na página 39.

PATACCHIOLA, M.; CANGELOSI, A. Head pose estimation in the wild using convolutional neural networks and adaptive gradient methods. *Pattern Recognition*, Elsevier, v. 71, p. 132–143, nov. 2017. Citado na página 20.

PODELL, D. et al. *SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis*. 2023. Disponível em: <https://arxiv.org/abs/2307.01952>. Citado na página 31.

POH, M.-Z. et al. A wearable sensor for unobtrusive, long-term assessment of electrodermal activity. *IEEE Transactions on Biomedical Engineering*, v. 57, n. 5, p. 1243–1252, 2010. Citado na página 19.

RADFORD, A. et al. *Learning Transferable Visual Models From Natural Language Supervision*. 2021. Disponível em: <https://arxiv.org/abs/2103.00020>. Citado 3 vezes nas páginas 11, 21, and 26.

RAGHU, M. et al. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In: *Advances in Neural Information Processing Systems (NeurIPS)*. [S.l.: s.n.], 2017. p. 6076–6085. Citado 2 vezes nas páginas 20 and 21.

RAO, C. R. The utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society. Series B (Methodological)*, v. 10, n. 2, p. 159–203, 1948. Citado na página 14.

ROMBACH, R. et al. *High-Resolution Image Synthesis with Latent Diffusion Models*. 2022. Disponível em: <https://arxiv.org/abs/2112.10752>. Citado 4 vezes nas páginas 5, 12, 25, and 26.

ROY, O.; VETTERLI, M. The effective rank: A measure of effective dimensionality. In: IEEE. *Proceedings of the 15th European Signal Processing Conference*. [S.l.], 2007. p. 606–610. Citado na página 16.

RUSSAKOVSKY, O. et al. *ImageNet Large Scale Visual Recognition Challenge*. 2015. Disponível em: <https://arxiv.org/abs/1409.0575>. Citado na página 4.

SAGUN, L. et al. Empirical analysis of the hessian of over-parameterized neural networks. *Journal of Machine Learning Research*, v. 18, n. 185, p. 1–27, 2018. Citado na página 21.

SCHRAUDOLPH, N. N. Fast curvature matrix-vector products for second-order gradient descent. *Neural Computation*, MIT Press, v. 14, n. 7, p. 1723–1738, 2002. Citado 2 vezes nas páginas 15 and 39.

SHAH, A. et al. Cadp: A novel dataset for cctv traffic camera based accident analysis. *arXiv preprint arXiv:1809.05782*, 2018. Citado na página 24.

SHARMA, P. et al. Student engagement detection using emotion analysis, eye tracking and head movement with machine learning. In: *International Conference on Technology and Innovation in Learning, Teaching and Education (TECH-EDU)*. [S.l.]: Springer, 2019. p. 403–414. Citado 2 vezes nas páginas 20 and 26.

SIMÉONI, O. et al. *DINOv3*. 2025. Disponível em: <https://arxiv.org/abs/2508.10104>. Citado 2 vezes nas páginas 11 and 21.

SNEDDON, I. et al. The belfast induced natural emotion database. *IEEE Transactions on Affective Computing*, v. 3, n. 1, p. 32–41, 2012. Citado na página 20.

TIAN, Y.-I.; KANADE, T.; COHN, J. F. Recognizing action units for facial expression analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 23, n. 2, p. 97–115, 2001. Citado na página 20.

TONG, S.; JONES, E.; STEINHARDT, J. *Mass-Producing Failures of Multimodal Systems with Language Models*. 2024. Disponível em: <https://arxiv.org/abs/2306.12105>. Citado 2 vezes nas páginas 31 and 51.

VALSTAR, M. F. et al. Spontaneous vs. posed facial behavior: automatic analysis of brow actions. In: *Proceedings of the 8th international conference on Multimodal interfaces*. [S.l.]: ACM, 2006. p. 162–170. Citado na página 20.

WHITEHILL, J. et al. The faces of engagement: Automatic recognition of student engagement from facial expressions. *IEEE Transactions on Affective Computing*, v. 5, n. 1, p. 86–98, jan. 2014. Citado 2 vezes nas páginas 19 and 26.

YANG, F. et al. *Student Classroom Behavior Detection based on YOLOv7-BRA and Multi-Model Fusion*. 2024. Disponível em: <https://arxiv.org/abs/2305.07825>. Citado na página 5.

ZAFEIRIOU, S. et al. Facial affect "in-the-wild": A survey and a new database. In: *International Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Affect "in-the-wild" Workshop*. [S.l.: s.n.], 2016. Citado na página 21.

ZHANG, X. et al. Appearance-based gaze estimation in the wild. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2015. p. 4511–4520. Citado na página 20.

ZHOU, K. et al. Understanding the sensitivity of clip models to prompt engineering. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2022. Citado na página 26.

ZHOU, W.; LING, C.; CAI, Z. *Emotion Recognition with CLIP and Sequential Learning*. 2025. Disponível em: <https://arxiv.org/abs/2503.09929>. Citado na página 22.

Annex

ANNEX A – Controlled vocabularies used for synthetic image generation

Table 2 – Synthetic image groups by prompt family.

Prompt family	Images
Attentive synonyms	**4,300**
Inattentive activities	**6,900**
Attentive behavioral proxies (e.g., posture)	**2,700**
Inattentive behavioral proxies (e.g., posture)	**1,000**

A.1 Attentive synonyms (4,300 images)

- absorbed
- actively listening
- alert
- attentive
- concentrated
- curious
- disciplined
- eager
- emotionally present
- engaged
- focused
- head up
- immersed in the task
- intellectually engaged
- interested and engaged
- interested
- involved
- leaning forward
- mentally present
- motivated
- observing
- oriented toward front
- participating
- paying attention
- present
- ready to learn
- serious
- sitting upright
- thoughtful
- visibly interested
- watchful
- with a calm focus
- with a direct gaze
- with a focused face
- with a focused posture
- with a steady gaze

- with a still posture
- with an expression
- with an engaged mind
- with an intent expression
- with clear eyes
- with eyes forward
- with silent attention

A.2 Attentive behavioral proxies (2,700 images)

- focused on the teacher with concentrated face
- focused on the teacher with slight smile
- focusing on the teacher with eyes
- leaning forward with eyes on the teacher
- leaning forward with hands on desk while listening
- listening attentively seated at desk
- listening attentively standing up
- listening attentively with eyes forward
- looking at the teacher while holding a pencil
- looking at the teacher while resting both hands on desk
- looking at the teacher with focused expression
- paying attention with chin slightly lifted
- paying attention with upright posture
- raising hand seated at desk
- raising hand standing up
- reading a book seated at desk
- reading a book standing up
- reading a textbook
- reading
- sitting straight with both hands folded on desk listening
- sitting straight with one hand holding a pen listening
- turning head slightly toward the teacher with engaged look
- turning whole body toward the teacher with focused face
- watching the board with eyes wide open
- watching the board with serious expression
- watching the teacher with engaged expression
- writing in a notebook

A.3 Inattentive activities (6,900 images)

- arranging miniature animal figures on the desk
- arranging sticky notes in a row on the desk

- arranging trading cards in a fan on the desk
- assembling lego bricks at the desk
- biting a pen cap at the desk
- blowing a bubble with gum at the desk
- building a small stack of coins on the desk
- building a tiny house of cards on the desk
- checking a phone
- checking a smartwatch at the desk
- checking notifications on a fitness tracker at the desk
- chewing gum at the desk
- clicking a tally counter at the desk
- counting coins from a pocket at the desk
- dozing
- drinking from a sports bottle at the desk
- drinking from a thermos at the desk
- drinking from a water bottle during class
- eating a granola bar at the desk
- eating chips from a small bag at the desk
- eating gummy candies at the desk
- flipping through a magazine at the desk
- flipping through a picture book at the desk
- folding an origami crane at the desk
- having a side conversation
- holding a small action figure on the desk
- holding a small toy dinosaur at the desk
- hunched over
- laughing histerically and with both arms extended pointing to the ceiling
- licking a lollipop at the desk
- looking at the ceiling
- making a paper fortune teller at the desk
- peeling an orange at the desk
- picking snacks from a lunchbox at the desk
- playing a handheld game console under the desk
- playing with a marble maze at the desk
- playing with a small toy car on the desk
- pulling on hoodie drawstrings at the desk
- reading a comic book at the desk
- reading a comic book under the desk
- rearranging items in a lunchbox at the desk
- resting chin on hand
- resting head on desk
- rolling dice back and forth on the desk
- screaming from the top of his lungs
- scrolling a phone
- scrolling on a small tablet at the desk

- sipping from a juice box at the desk
- sipping from a reusable straw cup at the desk
- sitting at a desk one arm straight up other arm out to the side shouting
- sitting slumped
- solving a rubik s cube at the desk
- spinning a fidget spinner on the desk
- spinning a ring on a finger at the desk
- spinning a top on the desk
- spinning a yo yo at the desk
- squeezing a squishy toy at the desk
- squeezing a stress ball at the desk
- stacking erasers into a small tower on the desk
- stacking sticky note pads into a tower at the desk
- sticking decorative stickers into a notebook at the desk
- stretching a rubber band between fingers over the desk
- stretching putty between hands over the desk
- tearing small strips of paper at the desk
- twirling hair
- twisting a puzzle ring at the desk
- unwrapping a candy bar at the desk
- whispering to a classmate
- with face in hands

A.4 Inattentive behavioral proxies (1,000 images)

- leaning forward with eyes half closed appearing tired
- pressing forehead against desk with eyes closed
- propping chin on both hands with blank face
- resting cheek on palm with distant gaze
- resting head on folded arms on desk
- rubbing eyes looking drowsy
- rubbing temples with closed eyes at desk
- supporting head with both hands looking bored
- touching forehead with hand appearing exhausted
- yawning with mouth half covered by hand

ANNEX B – Sample Train, Test Splits Composition

Table 3 – N50 – Train→Test activity overlap, Repetition 1. *Train*: number of training examples; *Test*: occurrences in test set. A dash (–) indicates the activity is absent from that split.

Class	Activity	Train	Test
Activities present in both Train and Test			
attentive	flipping_through_a_picture_book_at_the_desk	1	7
	raising_hand_seated_at_desk	1	7
	ready_to_learn	1	6
	assembling_lego_bricks_at_the_desk	1	5
	solving_a_rubik_s_cube_at_the_desk	1	5
	arranging_sticky_notes_in_a_row_on_the_desk	1	4
	building_a_tiny_house_of_cards_on_the_desk	1	4
	eating_a_granola_bar_at_the_desk	1	4
	holding_a_small_action_figure_on_the_desk	1	4
	playing_with_a_marble_maze_at_the_desk	1	4
	sticking_decorative_stickers_into_a_notebook_at_the_desk	1	4
	actively_listening	1	3
	biting_a_pen_cap_at_the_desk	1	3
	eager	1	3
	reading_a_book_seated_at_desk	1	3
	reading_a_comic_book_under_the_desk	1	3
	spinning_a_top_on_the_desk	1	3
	stacking_erasers_into_a_small_tower_on_the_desk	1	3
	turning_whole_body_toward_the_teacher_with_focused_face	1	3
	building_a_small_stack_of_coins_on_the_desk	1	2
	focused_on_the_teacher_with_slight_smile	1	2
	leaning_forward_with_hands_on_desk_while_listening	1	2
	peeling_an_orange_at_the_desk	1	2
	pulling_on_hoodie_drawstrings_at_the_desk	1	2
	reading_a_textbook	1	2
	scrolling_on_a_small_tablet_at_the_desk	1	2
	flipping_through_a_magazine_at_the_desk	1	1
	focused_on_the_teacher_with_concentrated_face	1	1
	folding_an_origami_crane_at_the_desk	1	1
	listening_attentively_with_eyes_forward	1	1
	looking_at_the_ceiling	1	1
	reading_a_comic_book_at_the_desk	1	1

continued on next page...

...continued

Class	Activity	Train	Test
	twisting_a_puzzle_ring_at_the_desk	1	1
inattentive	rubbing_temples_with_closed_eyes_at_desk	1	15
	with_face_in_hands	2	15
	rubbing_eyes_looking_drowsy	2	14
	dozing	1	13
	leaning_forward_with_eyes_half_closed_appearing_tired	1	13
	resting_head_on_desk	2	13
	hunched_over	2	11
	pressing_forehead_against_desk_with_eyes_closed	2	10
	resting_head_on_folded_arms_on_desk	1	10
	eating_chips_from_a_small_bag_at_the_desk	1	8
	sitting_slumped	2	8
	supporting_head_with_both_hands_looking_bored	2	8
	touching_forehead_with_hand_appearing_exhausted	1	8
	yawning_with_mouth_half_covered_by_hand	1	7
	chewing_gum_at_the_desk	1	5
	resting_cheek_on_palm_with_distant_gaze	1	4
	resting_chin_on_hand	1	4
	blowing_a_bubble_with_gum_at_the_desk	1	3
	checking_a_phone	1	3
	screaming_from_the_top_of_his_lungs	1	3
	squeezing_a_stress_ball_at_the_desk	1	3
	thoughtful	1	3
	twirling_hair	1	3
	paying_attention_with_upright_posture	1	2
	propping_chin_on_both_hands_with_blank_face	1	2
	scrolling_a_phone	1	2
	spinning_a_fidget_spinner_on_the_desk	1	2
	Subtotal	67	291
Activities present in Test only (zero-shot)			
attentive	making_a_paper_fortune_teller_at_the_desk	–	6
	playing_with_a_small_toy_car_on_the_desk	–	6
	raising_hand_standing_up	–	6
	clicking_a_tally_counter_at_the_desk	–	5
	reading	–	5
	drinking_from_a_water_bottle_during_class	–	4
	eating_gummy_candies_at_the_desk	–	4
	focusing_on_the_teacher_with_eyes	–	4
	sitting_at_a_desk_one_arm_straight_up_other_arm_out_to_the_sid...	–	4
	alert	–	3
	arranging_miniature_animal_figures_on_the_desk	–	3
	leaning_forward	–	3

continued on next page...

...continued

Class	Activity	Train	Test
	oriented_toward_front	–	3
	paying_attention_with_upright_posture	–	3
	sipping_from_a_juice_box_at_the_desk	–	3
	sitting_straight_with_one_hand_holding_a_pen_listening	–	3
	turning_head_slightly_toward_the_teacher_with_engaged_look	–	3
	with_clear_eyes	–	3
	absorbed	–	2
	checking_a_smartwatch_at_the_desk	–	2
	chewing_gum_at_the_desk	–	2
	eating_chips_from_a_small_bag_at_the_desk	–	2
	immersed_in_the_task	–	2
	listening_attentively_standing_up	–	2
	picking_snacks_from_a_lunchbox_at_the_desk	–	2
	present	–	2
	propping_chin_on_both_hands_with_blank_face	–	2
	rearranging_items_in_a_lunchbox_at_the_desk	–	2
	resting_chin_on_hand	–	2
	sipping_from_a_reusable_straw_cup_at_the_desk	–	2
	sitting_upright	–	2
	spinning_a_fidget_spinner_on_the_desk	–	2
	squeezing_a_squishy_toy_at_the_desk	–	2
	squeezing_a_stress_ball_at_the_desk	–	2
	tearing_small_strips_of_paper_at_the_desk	–	2
	with_an_expression	–	2
	with_eyes_forward	–	2
	arranging_trading_cards_in_a_fan_on_the_desk	–	1
	checking_a_phone	–	1
	checking_notifications_on_a_fitness_tracker_at_the_desk	–	1
	curious	–	1
	disciplined	–	1
	drinking_from_a_thermos_at_the_desk	–	1
	emotionally_present	–	1
	focused	–	1
	having_a_side_conversation	–	1
	head_up	–	1
	holding_a_small_toy_dinosaur_at_the_desk	–	1
	intellectually_engaged	–	1
	interested	–	1
	involved	–	1
	laughing_histerically_and_with_both_arms_extended_pointing_to_...	–	1
	leaning_forward_with_eyes_on_the_teacher	–	1
	mentally_present	–	1
	participating	–	1
	paying_attention_with_chin_slightly_lifted	–	1

continued on next page...

...continued

Class	Activity	Train	Test
	playing_a_handheld_game_console_under_the_desk	–	1
	resting_cheek_on_palm_with_distant_gaze	–	1
	rolling_dice_back_and_forth_on_the_desk	–	1
	screaming_from_the_top_of_his_lungs	–	1
	spinning_a_yo_yo_at_the_desk	–	1
	stretching_putty_between_hands_over_the_desk	–	1
	thoughtful	–	1
	twirling_hair	–	1
	visibly_interested	–	1
	watchful	–	1
	watching_the_board_with_eyes_wide_open	–	1
	with_a_direct_gaze	–	1
	with_a_steady_gaze	–	1
inattentive	whispering_to_a_classmate	–	4
	listening_attentively_seated_at_desk	–	3
	observing	–	3
	stretching_a_rubber_band_between_fingers_over_the_desk	–	3
	alert	–	2
	leaning_forward_with_eyes_on_the_teacher	–	2
	leaning_forward_with_hands_on_desk_while_listening	–	2
	listening_attentively_with_eyes_forward	–	2
	mentally_present	–	2
	watching_the_board_with_eyes_wide_open	–	2
	with_silent_attention	–	2
	absorbed	–	1
	concentrated	–	1
	drinking_from_a_sports_bottle_at_the_desk	–	1
	drinking_from_a_water_bottle_during_class	–	1
	emotionally_present	–	1
	head_up	–	1
	holding_a_small_action_figure_on_the_desk	–	1
	immersed_in_the_task	–	1
	interested	–	1
	leaning_forward	–	1
	licking_a_lollipop_at_the_desk	–	1
	looking_at_the_teacher_while_resting_both_hands_on_desk	–	1
	paying_attention	–	1
	playing_a_handheld_game_console_under_the_desk	–	1
	playing_with_a_small_toy_car_on_the_desk	–	1
	present	–	1
	pulling_on_hoodie_drawstrings_at_the_desk	–	1
	serious	–	1
	sitting_straight_with_both_hands_folded_on_desk_listening	–	1

continued on next page...

...continued

Class	Activity		Train	Test
	turning_whole_body_toward_the_teacher_with_focused_face		–	1
	with_a_direct_gaze		–	1
		Subtotal	–	189
Activities present in Train only				
attentive	reading_a_book_standing_up		1	–
	writing_in_a_notebook		1	–
inattentive	watching_the_board_with_serious_expression		1	–
		Subtotal	3	–
		Total	**70**	**480**

Table 4 – N200 – Train→Test activity overlap, Repetition 1. *Train*: number of training examples; *Test*: occurrences in test set. A dash (–) indicates the activity is absent from that split.

Class	Activity	Train	Test
Activities present in both Train and Test			
attentive	flipping_through_a_picture_book_at_the_desk	2	7
	raising_hand_seated_at_desk	3	7
	making_a_paper_fortune_teller_at_the_desk	1	6
	playing_with_a_small_toy_car_on_the_desk	1	6
	raising_hand_standing_up	1	6
	ready_to_learn	2	6
	assembling_lego_bricks_at_the_desk	2	5
	clicking_a_tally_counter_at_the_desk	1	5
	reading	1	5
	solving_a_rubik_s_cube_at_the_desk	2	5
	arranging_sticky_notes_in_a_row_on_the_desk	1	4
	building_a_tiny_house_of_cards_on_the_desk	2	4
	drinking_from_a_water_bottle_during_class	1	4
	eating_a_granola_bar_at_the_desk	2	4
	eating_gummy_candies_at_the_desk	1	4
	focusing_on_the_teacher_with_eyes	1	4
	holding_a_small_action_figure_on_the_desk	2	4
	playing_with_a_marble_maze_at_the_desk	2	4
	sitting_at_a_desk_one_arm_straight_up_other_arm_out_to_the_side...	1	4
	sticking_decorative_stickers_into_a_notebook_at_the_desk	2	4
	actively_listening	3	3
	alert	1	3
	arranging_miniature_animal_figures_on_the_desk	1	3
	biting_a_pen_cap_at_the_desk	2	3
	eager	1	3
	leaning_forward	1	3
	paying_attention_with_upright_posture	1	3
	reading_a_book_seated_at_desk	2	3
	reading_a_comic_book_under_the_desk	2	3
	sipping_from_a_juice_box_at_the_desk	1	3
	sitting_straight_with_one_hand_holding_a_pen_listening	1	3
	spinning_a_top_on_the_desk	2	3
	stacking_erasers_into_a_small_tower_on_the_desk	2	3
	turning_whole_body_toward_the_teacher_with_focused_facefocused_...	1	3
	with_clear_eyes	1	3
	absorbed	1	2
	building_a_small_stack_of_coins_on_the_desk	2	2
	checking_a_smartwatch_at_the_desk	1	2
	chewing_gum_at_the_desk	1	2

continued on next page...

...continued

Class	Activity	Train	Test
	focused_on_the_teacher_with_slight_smile	3	2
	immersed_in_the_task	1	2
	leaning_forward_with_hands_on_desk_while_listening	2	2
	listening_attentively_standing_up	1	2
	peeling_an_orange_at_the_desk	2	2
	picking_snacks_from_a_lunchbox_at_the_desk	1	2
	present	1	2
	propping_chin_on_both_hands_with_blank_face	1	2
	pulling_on_hoodie_drawstrings_at_the_desk	2	2
	reading_a_textbook	2	2
	rearranging_items_in_a_lunchbox_at_the_desk	1	2
	scrolling_on_a_small_tablet_at_the_desk	2	2
	sipping_from_a_reusable_straw_cup_at_the_desk	1	2
	sitting_upright	1	2
	spinning_a_fidget_spinner_on_the_desk	1	2
	squeezing_a_squishy_toy_at_the_desk	1	2
	squeezing_a_stress_ball_at_the_desk	1	2
	tearing_small_strips_of_paper_at_the_desk	1	2
	with_an_expression	1	2
	with_eyes_forward	1	2
	arranging_trading_cards_in_a_fan_on_the_desk	1	1
	checking_notifications_on_a_fitness_tracker_at_the_desk	1	1
	curious	1	1
	emotionally_present	1	1
	flipping_through_a_magazine_at_the_desk	2	1
	focused	1	1
	focused_on_the_teacher_with_concentrated_face	2	1
	folding_an_origami_crane_at_the_desk	2	1
	having_a_side_conversation	1	1
	head_up	1	1
	holding_a_small_toy_dinosaur_at_the_desk	1	1
	intellectually_engaged	1	1
	interested	1	1
	laughing_histerically_and_with_both_arms_extended_pointing_to_t...	1	1
	leaning_forward_with_eyes_on_the_teacher	1	1
	listening_attentively_with_eyes_forward	2	1
	looking_at_the_ceiling	2	1
	paying_attention_with_chin_slightly_lifted	1	1
	reading_a_comic_book_at_the_desk	2	1
	rolling_dice_back_and_forth_on_the_desk	1	1
	spinning_a_yo_yo_at_the_desk	1	1
	stretching_putty_between_hands_over_the_desk	1	1
	twirling_hair	1	1
	twisting_a_puzzle_ring_at_the_desk	2	1

continued on next page...

...continued

Class	Activity	Train	Test
	watching_the_board_with_eyes_wide_open	1	1
	with_a_direct_gaze	1	1
inattentive	rubbing_temples_with_closed_eyes_at_desk	5	15
	with_face_in_hands	8	15
	rubbing_eyes_looking_drowsy	8	14
	dozing	4	13
	leaning_forward_with_eyes_half_closed_appearing_tired	5	13
	resting_head_on_desk	8	13
	hunched_over	5	11
	pressing_forehead_against_desk_with_eyes_closed	6	10
	resting_head_on_folded_arms_on_desk	5	10
	eating_chips_from_a_small_bag_at_the_desk	2	8
	sitting_slumped	6	8
	supporting_head_with_both_hands_looking_bored	7	8
	touching_forehead_with_hand_appearing_exhausted	5	8
	yawning_with_mouth_half_covered_by_hand	5	7
	chewing_gum_at_the_desk	1	5
	resting_cheek_on_palm_with_distant_gaze	5	4
	resting_chin_on_hand	3	4
	whispering_to_a_classmate	1	4
	blowing_a_bubble_with_gum_at_the_desk	3	3
	checking_a_phone	2	3
	listening_attentively_seated_at_desk	1	3
	observing	1	3
	screaming_from_the_top_of_his_lungs	5	3
	squeezing_a_stress_ball_at_the_desk	1	3
	stretching_a_rubber_band_between_fingers_over_the_desk	1	3
	thoughtful	2	3
	twirling_hair	1	3
	alert	1	2
	leaning_forward_with_eyes_on_the_teacher	1	2
	leaning_forward_with_hands_on_desk_while_listening	1	2
	paying_attention_with_upright_posture	1	2
	propping_chin_on_both_hands_with_blank_face	5	2
	scrolling_a_phone	1	2
	spinning_a_fidget_spinner_on_the_desk	1	2
	watching_the_board_with_eyes_wide_open	1	2
	absorbed	1	1
	concentrated	1	1
	emotionally_present	1	1
	immersed_in_the_task	1	1
	leaning_forward	1	1
	licking_a_lollipop_at_the_desk	1	1
	playing_a_handheld_game_console_under_the_desk	1	1

continued on next page...

...continued

Class	Activity		Train	Test
	pulling_on_hoodie_drawstrings_at_the_desk		1	1
	serious		1	1
		Subtotal	245	439
Activities present in Test only (zero-shot)				
attentive	oriented_toward_front		–	3
	turning_head_slightly_toward_the_teacher_with_engaged_look		–	3
	eating_chips_from_a_small_bag_at_the_desk		–	2
	resting_chin_on_hand		–	2
	checking_a_phone		–	1
	disciplined		–	1
	drinking_from_a_thermos_at_the_desk		–	1
	involved		–	1
	mentally_present		–	1
	participating		–	1
	playing_a_handheld_game_console_under_the_desk		–	1
	resting_cheek_on_palm_with_distant_gaze		–	1
	screaming_from_the_top_of_his_lungs		–	1
	thoughtful		–	1
	visibly_interested		–	1
	watchful		–	1
	with_a_steady_gaze		–	1
inattentive	listening_attentively_with_eyes_forward		–	2
	mentally_present		–	2
	with_silent_attention		–	2
	drinking_from_a_sports_bottle_at_the_desk		–	1
	drinking_from_a_water_bottle_during_class		–	1
	head_up		–	1
	holding_a_small_action_figure_on_the_desk		–	1
	interested		–	1
	looking_at_the_teacher_while_resting_both_hands_on_desk		–	1
	paying_attention		–	1
	playing_with_a_small_toy_car_on_the_desk		–	1
	present		–	1
	sitting_straight_with_both_hands_folded_on_desk_listening		–	1
	turning_whole_body_toward_the_teacher_with_focused_face		–	1
	with_a_direct_gaze		–	1
		Subtotal	–	41
Activities present in Train only				
attentive	reading_a_book_standing_up		2	–
	attentive		1	–
	concentrated		1	–

continued on next page...

...continued

Class	Activity	Train	Test
	counting_coins_from_a_pocket_at_the_desk	1	–
	dozing	1	–
	engaged	1	–
	interested_and_engaged	1	–
	licking_a_lollipop_at_the_desk	1	–
	listening_attentively_seated_at_desk	1	–
	looking_at_the_teacher_while_holding_a_pencil	1	–
	looking_at_the_teacher_while_resting_both_hands_on_desk	1	–
	looking_at_the_teacher_with_focused_expression	1	–
	motivated	1	–
	sitting_straight_with_both_hands_folded_on_desk_listening	1	–
	spinning_a_ring_on_a_finger_at_the_desk	1	–
	stacking_sticky_note_pads_into_a_tower_at_the_desk	1	–
	stretching_a_rubber_band_between_fingers_over_the_desk	1	–
	turning_whole_body_toward_the_teacher_with_focused_face	1	–
	unwrapping_a_candy_bar_at_the_desk	1	–
	with_a_still_posture	1	–
	writing_in_a_notebook	1	–
inattentive	watching_the_board_with_serious_expression	2	–
	checking_notifications_on_a_fitness_tracker_at_the_desk	1	–
	drinking_from_a_thermos_at_the_desk	1	–
	focused	1	–
	focused_on_the_teacher_with_concentrated_face	1	–
	looking_at_the_ceiling	1	–
	paying_attention_with_chin_slightly_lifted	1	–
	sipping_from_a_reusable_straw_cup_at_the_desk	1	–
	spinning_a_ring_on_a_finger_at_the_desk	1	–
	spinning_a_yo_yo_at_the_desk	1	–
	watchful	1	–
	with_a_focused_posture	1	–
	Subtotal	35	–
	Total	**280**	**480**

Table 5 – N400 – Train→Test activity overlap, Repetition 1. *Train*: number of training examples; *Test*: occurrences in test set. A dash (–) indicates the activity is absent from that split.

Class	Activity	Train	Test
Activities present in both Train and Test			
attentive	flipping_through_a_picture_book_at_the_desk	5	7
	raising_hand_seated_at_desk	6	7
	making_a_paper_fortune_teller_at_the_desk	3	6
	playing_with_a_small_toy_car_on_the_desk	2	6
	raising_hand_standing_up	3	6
	ready_to_learn	4	6
	assembling_lego_bricks_at_the_desk	4	5
	clicking_a_tally_counter_at_the_desk	2	5
	reading	2	5
	solving_a_rubik_s_cube_at_the_desk	3	5
	arranging_sticky_notes_in_a_row_on_the_desk	3	4
	building_a_tiny_house_of_cards_on_the_desk	3	4
	drinking_from_a_water_bottle_during_class	2	4
	eating_a_granola_bar_at_the_desk	4	4
	eating_gummy_candies_at_the_desk	3	4
	focusing_on_the_teacher_with_eyes	2	4
	holding_a_small_action_figure_on_the_desk	4	4
	playing_with_a_marble_maze_at_the_desk	4	4
	sitting_at_a_desk_one_arm_straight_up_other_arm_out_to_the_side...	1	4
	sticking_decorative_stickers_into_a_notebook_at_the_desk	4	4
	actively_listening	5	3
	alert	2	3
	arranging_miniature_animal_figures_on_the_desk	1	3
	biting_a_pen_cap_at_the_desk	3	3
	eager	3	3
	leaning_forward	2	3
	oriented_toward_front	1	3
	paying_attention_with_upright_posture	1	3
	reading_a_book_seated_at_desk	5	3
	reading_a_comic_book_under_the_desk	5	3
	sipping_from_a_juice_box_at_the_desk	2	3
	sitting_straight_with_one_hand_holding_a_pen_listening	2	3
	spinning_a_top_on_the_desk	4	3
	stacking_erasers_into_a_small_tower_on_the_desk	4	3
	turning_head_slightly_toward_the_teacher_with_engaged_look	1	3
	turning_whole_body_toward_the_teacher_with_focused_facefocused_...	3	3
	with_clear_eyes	1	3
	absorbed	2	2
	building_a_small_stack_of_coins_on_the_desk	4	2

continued on next page...

...continued

Class	Activity	Train	Test
	checking_a_smartwatch_at_the_desk	2	2
	chewing_gum_at_the_desk	1	2
	focused_on_the_teacher_with_slight_smile	6	2
	immersed_in_the_task	2	2
	leaning_forward_with_hands_on_desk_while_listening	4	2
	listening_attentively_standing_up	2	2
	peeling_an_orange_at_the_desk	5	2
	picking_snacks_from_a_lunchbox_at_the_desk	1	2
	present	1	2
	propping_chin_on_both_hands_with_blank_face	1	2
	pulling_on_hoodie_drawstrings_at_the_desk	4	2
	reading_a_textbook	4	2
	rearranging_items_in_a_lunchbox_at_the_desk	2	2
	resting_chin_on_hand	1	2
	scrolling_on_a_small_tablet_at_the_desk	4	2
	sipping_from_a_reusable_straw_cup_at_the_desk	2	2
	sitting_upright	3	2
	spinning_a_fidget_spinner_on_the_desk	1	2
	squeezing_a_squishy_toy_at_the_desk	2	2
	squeezing_a_stress_ball_at_the_desk	2	2
	tearing_small_strips_of_paper_at_the_desk	2	2
	with_an_expression	2	2
	with_eyes_forward	2	2
	arranging_trading_cards_in_a_fan_on_the_desk	1	1
	checking_notifications_on_a_fitness_tracker_at_the_desk	2	1
	curious	2	1
	disciplined	1	1
	drinking_from_a_thermos_at_the_desk	1	1
	emotionally_present	1	1
	flipping_through_a_magazine_at_the_desk	5	1
	focused	1	1
	focused_on_the_teacher_with_concentrated_face	3	1
	folding_an_origami_crane_at_the_desk	3	1
	having_a_side_conversation	2	1
	head_up	2	1
	holding_a_small_toy_dinosaur_at_the_desk	1	1
	intellectually_engaged	2	1
	interested	2	1
	laughing_histerically_and_with_both_arms_extended_pointing_to_t...	2	1
	leaning_forward_with_eyes_on_the_teacher	2	1
	listening_attentively_with_eyes_forward	3	1
	looking_at_the_ceiling	3	1
	mentally_present	1	1
	participating	1	1

continued on next page...

...continued

Class	Activity	Train	Test
	paying_attention_with_chin_slightly_lifted	1	1
	playing_a_handheld_game_console_under_the_desk	1	1
	reading_a_comic_book_at_the_desk	3	1
	rolling_dice_back_and_forth_on_the_desk	1	1
	spinning_a_yo_yo_at_the_desk	3	1
	stretching_putty_between_hands_over_the_desk	2	1
	thoughtful	1	1
	twirling_hair	1	1
	twisting_a_puzzle_ring_at_the_desk	3	1
	visibly_interested	1	1
	watchful	1	1
	watching_the_board_with_eyes_wide_open	2	1
	with_a_direct_gaze	1	1
	with_a_steady_gaze	1	1
inattentive	rubbing_temples_with_closed_eyes_at_desk	10	15
	with_face_in_hands	16	15
	rubbing_eyes_looking_drowsy	16	14
	dozing	9	13
	leaning_forward_with_eyes_half_closed_appearing_tired	10	13
	resting_head_on_desk	15	13
	hunched_over	10	11
	pressing_forehead_against_desk_with_eyes_closed	11	10
	resting_head_on_folded_arms_on_desk	10	10
	eating_chips_from_a_small_bag_at_the_desk	4	8
	sitting_slumped	12	8
	supporting_head_with_both_hands_looking_bored	15	8
	touching_forehead_with_hand_appearing_exhausted	10	8
	yawning_with_mouth_half_covered_by_hand	10	7
	chewing_gum_at_the_desk	3	5
	resting_cheek_on_palm_with_distant_gaze	9	4
	resting_chin_on_hand	6	4
	whispering_to_a_classmate	2	4
	blowing_a_bubble_with_gum_at_the_desk	6	3
	checking_a_phone	5	3
	listening_attentively_seated_at_desk	1	3
	observing	1	3
	screaming_from_the_top_of_his_lungs	10	3
	squeezing_a_stress_ball_at_the_desk	2	3
	stretching_a_rubber_band_between_fingers_over_the_desk	1	3
	thoughtful	4	3
	twirling_hair	2	3
	alert	2	2
	leaning_forward_with_eyes_on_the_teacher	2	2
	leaning_forward_with_hands_on_desk_while_listening	1	2

continued on next page...

...continued

Class	Activity		Train	Test
	mentally_present		1	2
	paying_attention_with_upright_posture		3	2
	propping_chin_on_both_hands_with_blank_face		9	2
	scrolling_a_phone		2	2
	spinning_a_fidget_spinner_on_the_desk		2	2
	watching_the_board_with_eyes_wide_open		1	2
	with_silent_attention		1	2
	absorbed		2	1
	concentrated		1	1
	drinking_from_a_sports_bottle_at_the_desk		1	1
	emotionally_present		1	1
	head_up		1	1
	immersed_in_the_task		1	1
	interested		1	1
	leaning_forward		1	1
	licking_a_lollipop_at_the_desk		1	1
	looking_at_the_teacher_while_resting_both_hands_on_desk		1	1
	paying_attention		1	1
	playing_a_handheld_game_console_under_the_desk		1	1
	pulling_on_hoodie_drawstrings_at_the_desk		1	1
	serious		1	1
	turning_whole_body_toward_the_teacher_with_focused_face		1	1
		Subtotal	484	466
Activities present in Test only (zero-shot)				
attentive	eating_chips_from_a_small_bag_at_the_desk		–	2
	checking_a_phone		–	1
	involved		–	1
	resting_cheek_on_palm_with_distant_gaze		–	1
	screaming_from_the_top_of_his_lungs		–	1
inattentive	listening_attentively_with_eyes_forward		–	2
	drinking_from_a_water_bottle_during_class		–	1
	holding_a_small_action_figure_on_the_desk		–	1
	playing_with_a_small_toy_car_on_the_desk		–	1
	present		–	1
	sitting_straight_with_both_hands_folded_on_desk_listening		–	1
	with_a_direct_gaze		–	1
		Subtotal	–	14
Activities present in Train only				
attentive	reading_a_book_standing_up		4	–
	writing_in_a_notebook		3	–
	attentive		2	–

continued on next page...

...continued

Class	Activity	Train	Test
	concentrated	2	–
	counting_coins_from_a_pocket_at_the_desk	2	–
	engaged	2	–
	interested_and_engaged	2	–
	licking_a_lollipop_at_the_desk	2	–
	listening_attentively_seated_at_desk	2	–
	looking_at_the_teacher_while_holding_a_pencil	2	–
	looking_at_the_teacher_while_resting_both_hands_on_desk	2	–
	spinning_a_ring_on_a_finger_at_the_desk	2	–
	stacking_sticky_note_pads_into_a_tower_at_the_desk	2	–
	dozing	1	–
	drinking_from_a_sports_bottle_at_the_desk	1	–
	hunched_over	1	–
	leaning_forward_with_eyes_half_closed_appearing_tired	1	–
	looking_at_the_teacher_with_focused_expression	1	–
	motivated	1	–
	paying_attention	1	–
	resting_head_on_folded_arms_on_desk	1	–
	sitting_straight_with_both_hands_folded_on_desk_listening	1	–
	stretching_a_rubber_band_between_fingers_over_the_desk	1	–
	turning_whole_body_toward_the_teacher_with_focused_face	1	–
	unwrapping_a_candy_bar_at_the_desk	1	–
	with_a_calm_focus	1	–
	with_a_focused_face	1	–
	with_a_focused_posture	1	–
	with_a_still_posture	1	–
	with_an_engaged_mind	1	–
inattentive	watching_the_board_with_serious_expression	3	–
	looking_at_the_ceiling	2	–
	paying_attention_with_chin_slightly_lifted	2	–
	spinning_a_ring_on_a_finger_at_the_desk	2	–
	arranging_sticky_notes_in_a_row_on_the_desk	1	–
	checking_a_smartwatch_at_the_desk	1	–
	checking_notifications_on_a_fitness_tracker_at_the_desk	1	–
	curious	1	–
	disciplined	1	–
	drinking_from_a_thermos_at_the_desk	1	–
	eating_gummy_candies_at_the_desk	1	–
	focused	1	–
	focused_on_the_teacher_with_concentrated_face	1	–
	having_a_side_conversation	1	–
	holding_a_small_toy_dinosaur_at_the_desk	1	–
	listening_attentively_standing_up	1	–
	sipping_from_a_reusable_straw_cup_at_the_desk	1	–

continued on next page...

...continued

Class	Activity	Train	Test
	spinning_a_yo_yo_at_the_desk	1	–
	stretching_putty_between_hands_over_the_desk	1	–
	unwrapping_a_candy_bar_at_the_desk	1	–
	visibly_interested	1	–
	watchful	1	–
	with_a_focused_posture	1	–
	with_a_still_posture	1	–
	with_an_engaged_mind	1	–
	Subtotal	76	–
	Total	**560**	**480**

Table 6 – N1600 – Train→Test activity overlap, Repetition 1. *Train*: number of training examples; *Test*: occurrences in test set. A dash (–) indicates the activity is absent from that split.

Class	Activity	Train	Test
Activities present in both Train and Test			
attentive	flipping_through_a_picture_book_at_the_desk	19	7
	raising_hand_seated_at_desk	23	7
	making_a_paper_fortune_teller_at_the_desk	11	6
	playing_with_a_small_toy_car_on_the_desk	10	6
	raising_hand_standing_up	11	6
	ready_to_learn	14	6
	assembling_lego_bricks_at_the_desk	15	5
	clicking_a_tally_counter_at_the_desk	7	5
	reading	8	5
	solving_a_rubik_s_cube_at_the_desk	14	5
	arranging_sticky_notes_in_a_row_on_the_desk	12	4
	building_a_tiny_house_of_cards_on_the_desk	14	4
	drinking_from_a_water_bottle_during_class	7	4
	eating_a_granola_bar_at_the_desk	15	4
	eating_gummy_candies_at_the_desk	11	4
	focusing_on_the_teacher_with_eyes	8	4
	holding_a_small_action_figure_on_the_desk	15	4
	playing_with_a_marble_maze_at_the_desk	17	4
	sitting_at_a_desk_one_arm_straight_up_other_arm_out_to_the_side...	5	4
	sticking_decorative_stickers_into_a_notebook_at_the_desk	17	4
	actively_listening	20	3
	alert	6	3
	arranging_miniature_animal_figures_on_the_desk	5	3
	biting_a_pen_cap_at_the_desk	14	3
	eager	11	3
	leaning_forward	6	3
	oriented_toward_front	3	3
	paying_attention_with_upright_posture	5	3
	reading_a_book_seated_at_desk	20	3
	reading_a_comic_book_under_the_desk	20	3
	sipping_from_a_juice_box_at_the_desk	10	3
	sitting_straight_with_one_hand_holding_a_pen_listening	10	3
	spinning_a_top_on_the_desk	16	3
	stacking_erasers_into_a_small_tower_on_the_desk	15	3
	turning_head_slightly_toward_the_teacher_with_engaged_look	3	3
	turning_whole_body_toward_the_teacher_with_focused_facefocused_...	12	3
	with_clear_eyes	5	3
	absorbed	8	2
	building_a_small_stack_of_coins_on_the_desk	17	2

continued on next page...

...continued

Class	Activity	Train	Test
	checking_a_smartwatch_at_the_desk	10	2
	chewing_gum_at_the_desk	4	2
	eating_chips_from_a_small_bag_at_the_desk	1	2
	focused_on_the_teacher_with_slight_smile	26	2
	immersed_in_the_task	6	2
	leaning_forward_with_hands_on_desk_while_listening	17	2
	listening_attentively_standing_up	8	2
	peeling_an_orange_at_the_desk	19	2
	picking_snacks_from_a_lunchbox_at_the_desk	5	2
	present	5	2
	propping_chin_on_both_hands_with_blank_face	4	2
	pulling_on_hoodie_drawstrings_at_the_desk	16	2
	reading_a_textbook	17	2
	rearranging_items_in_a_lunchbox_at_the_desk	10	2
	resting_chin_on_hand	2	2
	scrolling_on_a_small_tablet_at_the_desk	15	2
	sipping_from_a_reusable_straw_cup_at_the_desk	9	2
	sitting_upright	11	2
	spinning_a_fidget_spinner_on_the_desk	5	2
	squeezing_a_squishy_toy_at_the_desk	8	2
	squeezing_a_stress_ball_at_the_desk	7	2
	tearing_small_strips_of_paper_at_the_desk	8	2
	with_an_expression	6	2
	with_eyes_forward	9	2
	arranging_trading_cards_in_a_fan_on_the_desk	5	1
	checking_a_phone	1	1
	checking_notifications_on_a_fitness_tracker_at_the_desk	9	1
	curious	6	1
	disciplined	2	1
	drinking_from_a_thermos_at_the_desk	3	1
	emotionally_present	4	1
	flipping_through_a_magazine_at_the_desk	19	1
	focused	5	1
	focused_on_the_teacher_with_concentrated_face	12	1
	folding_an_origami_crane_at_the_desk	14	1
	having_a_side_conversation	7	1
	head_up	6	1
	holding_a_small_toy_dinosaur_at_the_desk	5	1
	intellectually_engaged	8	1
	interested	6	1
	involved	1	1
	laughing_histerically_and_with_both_arms_extended_pointing_to_t...	10	1
	leaning_forward_with_eyes_on_the_teacher	8	1
	listening_attentively_with_eyes_forward	12	1

continued on next page...

...continued

Class	Activity	Train	Test
	looking_at_the_ceiling	14	1
	mentally_present	2	1
	participating	3	1
	paying_attention_with_chin_slightly_lifted	6	1
	playing_a_handheld_game_console_under_the_desk	2	1
	reading_a_comic_book_at_the_desk	13	1
	resting_cheek_on_palm_with_distant_gaze	1	1
	rolling_dice_back_and_forth_on_the_desk	5	1
	spinning_a_yo_yo_at_the_desk	11	1
	stretching_putty_between_hands_over_the_desk	9	1
	thoughtful	2	1
	twirling_hair	5	1
	twisting_a_puzzle_ring_at_the_desk	14	1
	visibly_interested	2	1
	watchful	2	1
	watching_the_board_with_eyes_wide_open	8	1
	with_a_direct_gaze	4	1
	with_a_steady_gaze	3	1
inattentive	rubbing_temples_with_closed_eyes_at_desk	40	15
	with_face_in_hands	66	15
	rubbing_eyes_looking_drowsy	63	14
	dozing	35	13
	leaning_forward_with_eyes_half_closed_appearing_tired	38	13
	resting_head_on_desk	60	13
	hunched_over	41	11
	pressing_forehead_against_desk_with_eyes_closed	46	10
	resting_head_on_folded_arms_on_desk	41	10
	eating_chips_from_a_small_bag_at_the_desk	18	8
	sitting_slumped	48	8
	supporting_head_with_both_hands_looking_bored	59	8
	touching_forehead_with_hand_appearing_exhausted	41	8
	yawning_with_mouth_half_covered_by_hand	41	7
	chewing_gum_at_the_desk	10	5
	resting_cheek_on_palm_with_distant_gaze	36	4
	resting_chin_on_hand	26	4
	whispering_to_a_classmate	9	4
	blowing_a_bubble_with_gum_at_the_desk	23	3
	checking_a_phone	19	3
	listening_attentively_seated_at_desk	4	3
	observing	4	3
	screaming_from_the_top_of_his_lungs	41	3
	squeezing_a_stress_ball_at_the_desk	9	3
	stretching_a_rubber_band_between_fingers_over_the_desk	4	3
	thoughtful	16	3

continued on next page...

...continued

Class	Activity		Train	Test
	twirling_hair		9	3
	alert		6	2
	leaning_forward_with_eyes_on_the_teacher		8	2
	leaning_forward_with_hands_on_desk_while_listening		4	2
	listening_attentively_with_eyes_forward		1	2
	mentally_present		3	2
	paying_attention_with_upright_posture		11	2
	propping_chin_on_both_hands_with_blank_face		36	2
	scrolling_a_phone		9	2
	spinning_a_fidget_spinner_on_the_desk		10	2
	watching_the_board_with_eyes_wide_open		5	2
	with_silent_attention		3	2
	absorbed		6	1
	concentrated		3	1
	drinking_from_a_sports_bottle_at_the_desk		3	1
	emotionally_present		5	1
	head_up		2	1
	immersed_in_the_task		5	1
	interested		2	1
	leaning_forward		5	1
	licking_a_lollipop_at_the_desk		5	1
	looking_at_the_teacher_while_resting_both_hands_on_desk		2	1
	paying_attention		3	1
	playing_a_handheld_game_console_under_the_desk		5	1
	playing_with_a_small_toy_car_on_the_desk		1	1
	pulling_on_hoodie_drawstrings_at_the_desk		4	1
	serious		5	1
	turning_whole_body_toward_the_teacher_with_focused_face		2	1
	with_a_direct_gaze		1	1
		Subtotal	1938	475
Activities present in Test only (zero-shot)				
attentive	screaming_from_the_top_of_his_lungs		–	1
inattentive	drinking_from_a_water_bottle_during_class		–	1
	holding_a_small_action_figure_on_the_desk		–	1
	present		–	1
	sitting_straight_with_both_hands_folded_on_desk_listening		–	1
		Subtotal	–	5
Activities present in Train only				
attentive	reading_a_book_standing_up		15	–
	writing_in_a_notebook		12	–
	concentrated		9	–

continued on next page...

...continued

Class	Activity	Train	Test
	counting_coins_from_a_pocket_at_the_desk	9	–
	listening_attentively_seated_at_desk	9	–
	spinning_a_ring_on_a_finger_at_the_desk	9	–
	engaged	8	–
	looking_at_the_teacher_while_holding_a_pencil	7	–
	stacking_sticky_note_pads_into_a_tower_at_the_desk	7	–
	attentive	6	–
	interested_and_engaged	6	–
	licking_a_lollipop_at_the_desk	6	–
	looking_at_the_teacher_while_resting_both_hands_on_desk	6	–
	sitting_straight_with_one_hand_holding_a_pen_listening	6	–
	stretching_a_rubber_band_between_fingers_over_the_desk	6	–
	unwrapping_a_candy_bar_at_the_desk	6	–
	looking_at_the_teacher_with_focused_expression	5	–
	motivated	5	–
	turning_whole_body_toward_the_teacher_with_focused_face	5	–
	dozing	4	–
	with_a_still_posture	4	–
	drinking_from_a_sports_bottle_at_the_desk	3	–
	leaning_forward_with_eyes_half_closed_appearing_tired	3	–
	paying_attention	3	–
	resting_head_on_folded_arms_on_desk	3	–
	with_a_calm_focus	3	–
	with_an_engaged_mind	3	–
	hunched_over	2	–
	with_a_focused_face	2	–
	with_a_focused_posture	2	–
	observing	1	–
	rubbing_temples_with_closed_eyes_at_desk	1	–
	serious	1	–
	sitting_slumped	1	–
	watching_the_board_with_serious_expression	1	–
	watching_the_teacher_with_engaged_expression	1	–
	whispering_to_a_classmate	1	–
	with_an_intent_expression	1	–
	with_face_in_hands	1	–
	with_silent_attention	1	–
inattentive	watching_the_board_with_serious_expression	12	–
	paying_attention_with_chin_slightly_lifted	9	–
	looking_at_the_ceiling	6	–
	spinning_a_ring_on_a_finger_at_the_desk	6	–
	drinking_from_a_thermos_at_the_desk	5	–
	focused	5	–
	watchful	5	–

continued on next page...

...continued

Class	Activity	Train	Test
	checking_notifications_on_a_fitness_tracker_at_the_desk	4	–
	focused_on_the_teacher_with_concentrated_face	4	–
	sipping_from_a_reusable_straw_cup_at_the_desk	4	–
	spinning_a_yo_yo_at_the_desk	4	–
	with_a_focused_posture	4	–
	checking_a_smartwatch_at_the_desk	3	–
	curious	3	–
	disciplined	3	–
	eating_gummy_candies_at_the_desk	3	–
	holding_a_small_toy_dinosaur_at_the_desk	3	–
	listening_attentively_standing_up	3	–
	stretching_putty_between_hands_over_the_desk	3	–
	unwrapping_a_candy_bar_at_the_desk	3	–
	visibly_interested	3	–
	with_a_still_posture	3	–
	arranging_sticky_notes_in_a_row_on_the_desk	2	–
	having_a_side_conversation	2	–
	with_an_engaged_mind	2	–
	arranging_miniature_animal_figures_on_the_desk	1	–
	attentive	1	–
	focusing_on_the_teacher_with_eyes	1	–
	interested_and_engaged	1	–
	making_a_paper_fortune_teller_at_the_desk	1	–
	raising_hand_standing_up	1	–
	sipping_from_a_juice_box_at_the_desk	1	–
	spinning_a_top_on_the_desk	1	–
	squeezing_a_squishy_toy_at_the_desk	1	–
	turning_whole_body_toward_the_teacher_with_focused_facefocused_...	1	–
	watching_the_teacher_with_engaged_expression	1	–
	with_a_calm_focus	1	–
	with_a_focused_face	1	–
	with_eyes_forward	1	–
	Subtotal	302	–
	Total	**2240**	**480**

ANNEX C – Activities as Distributed Among Classes by Vision-Language Models

Table 7 – Activities as distributed among classes by Vision-Language Models

Class	Activity	Count
Attentive	absorbed	87
	actively_listening	93
	alert	68
	arranging_miniature_animal_figures_on_the_desk	90
	arranging_sticky_notes_in_a_row_on_the_desk	93
	arranging_trading_cards_in_a_fan_on_the_desk	88
	assembling_lego_bricks_at_the_desk	100
	attentive	52
	biting_a_pen_cap_at_the_desk	100
	blowing_a_bubble_with_gum_at_the_desk	29
	building_a_small_stack_of_coins_on_the_desk	98
	building_a_tiny_house_of_cards_on_the_desk	98
	checking_a_phone	27
	checking_a_smartwatch_at_the_desk	90
	checking_notifications_on_a_fitness_tracker_at...	84
	chewing_gum_at_the_desk	62
	clicking_a_tally_counter_at_the_desk	93
	concentrated	65
	counting_coins_from_a_pocket_at_the_desk	71
	curious	60
	disciplined	65
	dozing	32
	drinking_from_a_sports_bottle_at_the_desk	91
	drinking_from_a_thermos_at_the_desk	76
	drinking_from_a_water_bottle_during_class	96
	eager	72
	eating_a_granola_bar_at_the_desk	95
	eating_chips_from_a_small_bag_at_the_desk	33
	eating_gummy_candies_at_the_desk	75
	emotionally_present	49

Class	Activity	Count
	engaged	56
	flipping_through_a_magazine_at_the_desk	99
	flipping_through_a_picture_book_at_the_desk	92
	focused	32
	focused_on_the_teacher_with_concentrated_face	93
	focused_on_the_teacher_with_slight_smile	95
	focusing_on_the_teacher_with_eyes	55
	folding_an_origami_crane_at_the_desk	95
	having_a_side_conversation	96
	head_up	84
	holding_a_small_action_figure_on_the_desk	91
	holding_a_small_toy_dinosaur_at_the_desk	93
	hunched_over	24
	immersed_in_the_task	61
	intellectually_engaged	58
	interested	56
	interested_and_engaged	41
	involved	36
	laughing_histerically_and_with_both_arms_exten...	98
	leaning_forward	55
	leaning_forward_with_eyes_half_closed_appearing_tired	31
	leaning_forward_with_eyes_on_the_teacher	56
	leaning_forward_with_hands_on_desk_while_listening	85
	licking_a_lollipop_at_the_desk	65
	listening_attentively_seated_at_desk	72
	listening_attentively_standing_up	89
	listening_attentively_with_eyes_forward	66
	looking_at_the_ceiling	89
	looking_at_the_teacher_while_holding_a_pencil	60
	looking_at_the_teacher_while_resting_both_hands_on_desk	53
	looking_at_the_teacher_with_focused_expression	51
	making_a_paper_fortune_teller_at_the_desk	98
	mentally_present	38
	motivated	25
	observing	50
	oriented_toward_front	46
	participating	47

Class	Activity	Count
	paying_attention	25
	paying_attention_with_chin_slightly_lifted	66
	paying_attention_with_upright_posture	56
	peeling_an_orange_at_the_desk	100
	picking_snacks_from_a_lunchbox_at_the_desk	97
	playing_a_handheld_game_console_under_the_desk	65
	playing_with_a_marble_maze_at_the_desk	98
	playing_with_a_small_toy_car_on_the_desk	97
	present	41
	pressing_forehead_against_desk_with_eyes_closed	14
	propping_chin_on_both_hands_with_blank_face	46
	pulling_on_hoodie_drawstrings_at_the_desk	89
	raising_hand_seated_at_desk	97
	raising_hand_standing_up	86
	reading	59
	reading_a_book_seated_at_desk	94
	reading_a_book_standing_up	80
	reading_a_comic_book_at_the_desk	96
	reading_a_comic_book_under_the_desk	100
	reading_a_textbook	75
	ready_to_learn	91
	rearranging_items_in_a_lunchbox_at_the_desk	97
	resting_cheek_on_palm_with_distant_gaze	22
	resting_chin_on_hand	54
	resting_head_on_desk	8
	resting_head_on_folded_arms_on_desk	17
	rolling_dice_back_and_forth_on_the_desk	91
	rubbing_temples_with_closed_eyes_at_desk	6
	screaming_from_the_top_of_his_lungs	15
	scrolling_a_phone	27
	scrolling_on_a_small_tablet_at_the_desk	96
	serious	40
	sipping_from_a_juice_box_at_the_desk	88
	sipping_from_a_reusable_straw_cup_at_the_desk	90
	sitting_at_a_desk_one_arm_straight_up_other:.	86
	sitting_slumped	15
	sitting_straight_with_both_hands_folded_on_desk_listening	75

Class	Activity	Count
	sitting_straight_with_one_hand_holding_a_pen_listening	51
	sitting_upright	74
	solving_a_rubik_s_cube_at_the_desk	99
	spinning_a_fidget_spinner_on_the_desk	57
	spinning_a_ring_on_a_finger_at_the_desk	82
	spinning_a_top_on_the_desk	94
	spinning_a_yo_yo_at_the_desk	86
	squeezing_a_squishy_toy_at_the_desk	95
	squeezing_a_stress_ball_at_the_desk	67
	stacking_erasers_into_a_small_tower_on_the_desk	96
	stacking_sticky_note_pads_into_a_tower_at_th...	94
	sticking_decorative_stickers_into_a_notebook_a...	90
	stretching_a_rubber_band_between_fingers_over:.	72
	stretching_putty_between_hands_over_the_desk	86
	supporting_head_with_both_hands_looking_bored	2
	tearing_small_strips_of_paper_at_the_desk	91
	thoughtful	28
	touching_forehead_with_hand_appearing_exhausted	5
	turning_head_slightly_toward_the_teacher_with_engaged_look	67
	turning_whole_body_toward_the_teacher_with_focused_face	133
	twirling_hair	60
	twisting_a_puzzle_ring_at_the_desk	88
	unwrapping_a_candy_bar_at_the_desk	91
	visibly_interested	37
	watchful	40
	watching_the_board_with_eyes_wide_open	66
	watching_the_board_with_serious_expression	30
	watching_the_teacher_with_engaged_expression	53
	whispering_to_a_classmate	61
	with_a_calm_focus	30
	with_a_direct_gaze	38
	with_a_focused_face	38
	with_a_focused_posture	29
	with_a_steady_gaze	30
	with_a_still_posture	42
	with_an_engaged_mind	24
	with_an_expression	28

Class	Activity	Count
	with_an_intent_expression	28
	with_clear_eyes	55
	with_eyes_forward	39
	with_face_in_hands	3
	with_silent_attention	20
	writing_in_a_notebook	61
	yawning_with_mouth_half_covered_by_hand	3
Inattentive	absorbed	8
	alert	8
	arranging_miniature_animal_figures_on_the_desk	3
	arranging_sticky_notes_in_a_row_on_the_desk	3
	attentive	1
	blowing_a_bubble_with_gum_at_the_desk	31
	checking_a_phone	26
	checking_a_smartwatch_at_the_desk	4
	checking_notifications_on_a_fitness_tracker_at...	5
	chewing_gum_at_the_desk	18
	clicking_a_tally_counter_at_the_desk	2
	concentrated	5
	counting_coins_from_a_pocket_at_the_desk	1
	curious	3
	disciplined	3
	dozing	60
	drinking_from_a_sports_bottle_at_the_desk	5
	drinking_from_a_thermos_at_the_desk	6
	drinking_from_a_water_bottle_during_class	1
	eating_chips_from_a_small_bag_at_the_desk	26
	eating_gummy_candies_at_the_desk	3
	emotionally_present	6
	focused	5
	focused_on_the_teacher_with_concentrated_face	5
	focusing_on_the_teacher_with_eyes	1
	folding_an_origami_crane_at_the_desk	1
	head_up	3
	having_a_side_conversation	2
	holding_a_small_action_figure_on_the_desk	2
	holding_a_small_toy_dinosaur_at_the_desk	3

Class	Activity	Count
	hunched_over	60
	immersed_in_the_task	6
	interested	4
	interested_and_engaged	2
	leaning_forward	8
	leaning_forward_with_eyes_half_closed_appearing_tired	61
	leaning_forward_with_eyes_on_the_teacher	13
	leaning_forward_with_hands_on_desk_while_listening	7
	licking_a_lollipop_at_the_desk	8
	listening_attentively_seated_at_desk	7
	listening_attentively_standing_up	4
	listening_attentively_with_eyes_forward	4
	looking_at_the_ceiling	7
	looking_at_the_teacher_while_resting_both_hands_on_desk	3
	looking_at_the_teacher_with_focused_expression	2
	making_a_paper_fortune_teller_at_the_desk	1
	mentally_present	6
	observing	7
	paying_attention	4
	paying_attention_with_chin_slightly_lifted	9
	paying_attention_with_upright_posture	17
	playing_a_handheld_game_console_under_the_desk	6
	playing_with_a_small_toy_car_on_the_desk	2
	present	1
	pressing_forehead_against_desk_with_eyes_closed	68
	propping_chin_on_both_hands_with_blank_face	41
	pulling_on_hoodie_drawstrings_at_the_desk	6
	raising_hand_standing_up	1
	resting_cheek_on_palm_with_distant_gaze	56
	resting_chin_on_hand	38
	resting_head_on_desk	81
	resting_head_on_folded_arms_on_desk	64
	rubbing_eyes_looking_drowsy	87
	rubbing_temples_with_closed_eyes_at_desk	62
	screaming_from_the_top_of_his_lungs	53
	scrolling_a_phone	12
	serious	7

Class	Activity	Count
	sipping_from_a_juice_box_at_the_desk	1
	sipping_from_a_reusable_straw_cup_at_the_desk	4
	sitting_slumped	71
	sitting_straight_with_both_hands_folded_on_desk_listening	3
	spinning_a_fidget_spinner_on_the_desk	20
	spinning_a_ring_on_a_finger_at_the_desk	8
	spinning_a_top_on_the_desk	1
	spinning_a_yo_yo_at_the_desk	5
	squeezing_a_squishy_toy_at_the_desk	2
	squeezing_a_stress_ball_at_the_desk	13
	stacking_erasers_into_a_small_tower_on_the_desk	2
	stacking_sticky_note_pads_into_a_tower_at_th...	1
	stretching_a_rubber_band_between_fingers_over:.	10
	stretching_putty_between_hands_over_the_desk	3
	supporting_head_with_both_hands_looking_bored	80
	thoughtful	21
	touching_forehead_with_hand_appearing_exhausted	61
	turning_whole_body_toward_the_teacher_with_focused_face	4
	twirling_hair	16
	unwrapping_a_candy_bar_at_the_desk	3
	visibly_interested	4
	watchful	5
	watching_the_board_with_eyes_wide_open	7
	watching_the_board_with_serious_expression	13
	watching_the_teacher_with_engaged_expression	1
	whispering_to_a_classmate	23
	with_a_calm_focus	1
	with_a_direct_gaze	3
	with_a_focused_face	1
	with_a_focused_posture	5
	with_a_still_posture	3
	with_an_engaged_mind	2
	with_eyes_forward	1
	with_face_in_hands	91
	with_silent_attention	6
	yawning_with_mouth_half_covered_by_hand	58
Inconclusive	absorbed	4

Class	Activity	Count
	actively_listening	3
	alert	6
	arranging_miniature_animal_figures_on_the_desk	6
	arranging_sticky_notes_in_a_row_on_the_desk	1
	arranging_trading_cards_in_a_fan_on_the_desk	8
	attentive	1
	blowing_a_bubble_with_gum_at_the_desk	40
	building_a_small_stack_of_coins_on_the_desk	1
	checking_a_phone	47
	checking_a_smartwatch_at_the_desk	6
	checking_notifications_on_a_fitness_tracker_at...	6
	chewing_gum_at_the_desk	19
	clicking_a_tally_counter_at_the_desk	3
	concentrated	3
	counting_coins_from_a_pocket_at_the_desk	2
	curious	3
	disciplined	5
	dozing	7
	drinking_from_a_sports_bottle_at_the_desk	4
	drinking_from_a_thermos_at_the_desk	18
	drinking_from_a_water_bottle_during_class	2
	eating_a_granola_bar_at_the_desk	4
	eating_chips_from_a_small_bag_at_the_desk	41
	eating_gummy_candies_at_the_desk	22
	emotionally_present	2
	focused_on_the_teacher_with_concentrated_face	2
	focusing_on_the_teacher_with_eyes	2
	having_a_side_conversation	2
	head_up	4
	holding_a_small_action_figure_on_the_desk	7
	holding_a_small_toy_dinosaur_at_the_desk	4
	hunched_over	11
	immersed_in_the_task	3
	interested	5
	interested_and_engaged	3
	involved	1
	laughing_histerically_and_with_both_arms_exten...	2

Class	Activity	Count
	leaning_forward	1
	leaning_forward_with_eyes_half_closed_appearing_tired	8
	leaning_forward_with_eyes_on_the_teacher	8
	leaning_forward_with_hands_on_desk_while_listening	4
	licking_a_lollipop_at_the_desk	27
	listening_attentively_seated_at_desk	5
	listening_attentively_standing_up	6
	listening_attentively_with_eyes_forward	9
	looking_at_the_ceiling	3
	looking_at_the_teacher_while_resting_both_hands_on_desk	3
	looking_at_the_teacher_with_focused_expression	1
	making_a_paper_fortune_teller_at_the_desk	1
	mentally_present	4
	motivated	2
	observing	8
	oriented_toward_front	2
	participating	3
	paying_attention	3
	paying_attention_with_chin_slightly_lifted	8
	paying_attention_with_upright_posture	12
	picking_snacks_from_a_lunchbox_at_the_desk	2
	playing_a_handheld_game_console_under_the_desk	29
	playing_with_a_marble_maze_at_the_desk	2
	playing_with_a_small_toy_car_on_the_desk	1
	present	4
	pressing_forehead_against_desk_with_eyes_closed	6
	propping_chin_on_both_hands_with_blank_face	11
	pulling_on_hoodie_drawstrings_at_the_desk	3
	raising_hand_standing_up	2
	ready_to_learn	1
	reading	2
	reading_a_book_standing_up	1
	reading_a_comic_book_at_the_desk	1
	rearranging_items_in_a_lunchbox_at_the_desk	3
	resting_cheek_on_palm_with_distant_gaze	14
	resting_chin_on_hand	6
	resting_head_on_desk	1

Class	Activity	Count
	resting_head_on_folded_arms_on_desk	5
	rolling_dice_back_and_forth_on_the_desk	7
	rubbing_eyes_looking_drowsy	6
	rubbing_temples_with_closed_eyes_at_desk	5
	screaming_from_the_top_of_his_lungs	32
	scrolling_a_phone	29
	serious	8
	sipping_from_a_juice_box_at_the_desk	11
	sipping_from_a_reusable_straw_cup_at_the_desk	5
	sitting_at_a_desk_one_arm_straight_up_other:.	6
	sitting_slumped	7
	sitting_straight_with_both_hands_folded_on_desk_listening	5
	sitting_straight_with_one_hand_holding_a_pen_listening	1
	sitting_upright	2
	solving_a_rubik_s_cube_at_the_desk	1
	spinning_a_fidget_spinner_on_the_desk	23
	spinning_a_ring_on_a_finger_at_the_desk	7
	spinning_a_top_on_the_desk	5
	spinning_a_yo_yo_at_the_desk	6
	squeezing_a_squishy_toy_at_the_desk	2
	squeezing_a_stress_ball_at_the_desk	20
	stacking_erasers_into_a_small_tower_on_the_desk	1
	stacking_sticky_note_pads_into_a_tower_at_th...	2
	sticking_decorative_stickers_into_a_notebook_a...	5
	stretching_a_rubber_band_between_fingers_over:.	13
	stretching_putty_between_hands_over_the_desk	5
	supporting_head_with_both_hands_looking_bored	4
	thoughtful	12
	touching_forehead_with_hand_appearing_exhausted	2
	turning_head_slightly_toward_the_teacher_with_engaged_look	3
	turning_whole_body_toward_the_teacher_with_focused_face	15
	twirling_hair	23
	twisting_a_puzzle_ring_at_the_desk	1
	unwrapping_a_candy_bar_at_the_desk	6
	visibly_interested	6
	watchful	8
	watching_the_board_with_eyes_wide_open	11

Class	Activity	Count
	watching_the_board_with_serious_expression	12
	watching_the_teacher_with_engaged_expression	7
	whispering_to_a_classmate	14
	with_a_calm_focus	1
	with_a_direct_gaze	1
	with_a_focused_face	4
	with_a_focused_posture	3
	with_a_steady_gaze	1
	with_a_still_posture	3
	with_an_engaged_mind	3
	with_an_intent_expression	3
	with_face_in_hands	4
	with_silent_attention	2
	yawning_with_mouth_half_covered_by_hand	38

ANNEX D – Fisher Metrics

Table 8 – Fisher metrics for **CNN** representations on the **Linear Probe** adapter architecture

REP	N	train							val							test						
		μ_g	std	trace	E_1	erank	λ_1	λ_2	μ_g	std	trace	E_1	erank	λ_1	λ_2	μ_g	std	trace	E_1	erank	λ_1	λ_2
1	50	14.53	5.97	246.86	1.00	1.00	246.79	0.03	17.80	3.53	329.23	1.00	1.00	329.21	0.01	17.09	5.72	324.85	1.00	1.00	324.78	0.03
	100	12.51	6.54	199.36	1.00	1.00	199.29	0.04	14.40	4.53	227.99	1.00	1.00	227.94	0.03	14.91	7.07	272.26	1.00	1.00	272.16	0.05
	150	13.76	6.43	230.77	1.00	1.00	230.70	0.04	15.15	5.27	257.20	1.00	1.00	257.09	0.08	15.21	6.57	274.50	1.00	1.00	274.41	0.04
	200	11.54	7.54	190.07	1.00	1.00	189.99	0.04	13.15	7.12	223.56	1.00	1.01	223.41	0.10	13.73	8.58	262.12	1.00	1.00	261.99	0.06
	300	11.31	8.37	198.08	1.00	1.01	197.97	0.06	11.55	7.63	191.53	1.00	1.01	191.40	0.09	13.19	9.13	257.39	1.00	1.01	257.25	0.07
	400	11.21	8.73	201.73	1.00	1.01	201.62	0.05	11.79	8.42	209.84	1.00	1.01	209.70	0.09	12.89	9.51	256.54	1.00	1.01	256.40	0.07
	800	11.36	9.39	217.32	1.00	1.01	217.13	0.09	11.96	8.90	222.13	1.00	1.01	222.00	0.07	12.50	9.77	251.91	1.00	1.01	251.75	0.08
	1200	10.30	10.72	221.02	1.00	1.01	220.90	0.07	12.35	12.60	311.39	1.00	1.01	311.04	0.19	12.53	12.77	320.05	1.00	1.01	319.83	0.11
	1600	11.67	12.04	281.07	1.00	1.01	280.91	0.09	12.34	12.75	315.01	1.00	1.01	314.71	0.16	12.42	13.00	323.11	1.00	1.01	322.90	0.10
2	50	15.15	5.29	257.44	1.00	1.00	257.37	0.03	18.94	5.08	384.62	1.00	1.00	384.55	0.04	17.87	5.91	354.16	1.00	1.00	354.05	0.05
	100	14.26	5.84	237.56	1.00	1.00	237.48	0.04	17.25	6.98	346.10	1.00	1.00	345.93	0.10	16.15	6.60	304.37	1.00	1.00	304.27	0.05
	150	14.61	6.08	250.41	1.00	1.00	250.33	0.04	16.04	6.29	296.99	1.00	1.00	296.87	0.06	16.12	6.62	303.55	1.00	1.00	303.44	0.05
	200	14.97	6.07	261.02	1.00	1.00	260.94	0.04	16.22	6.18	301.09	1.00	1.00	301.01	0.04	16.07	6.46	300.10	1.00	1.00	300.00	0.05
	300	11.71	8.39	207.44	1.00	1.01	207.33	0.06	13.88	9.42	281.19	1.00	1.00	281.06	0.07	13.51	9.61	274.84	1.00	1.01	274.67	0.08
	400	12.00	8.79	221.35	1.00	1.01	221.21	0.07	14.22	9.30	288.86	1.00	1.00	288.72	0.07	13.41	9.59	271.67	1.00	1.01	271.50	0.08
	800	11.17	10.37	232.30	1.00	1.01	232.13	0.09	12.37	11.18	278.13	1.00	1.01	277.96	0.09	12.93	11.87	308.29	1.00	1.01	308.07	0.11
	1200	10.94	11.39	249.30	1.00	1.01	249.11	0.10	12.86	12.45	320.46	1.00	1.01	320.29	0.09	12.78	12.92	330.23	1.00	1.01	329.99	0.12
	1600	10.93	12.08	265.50	1.00	1.01	265.31	0.10	12.49	12.81	320.08	1.00	1.00	319.92	0.09	13.02	13.77	358.89	1.00	1.01	358.63	0.13
3	50	13.92	5.07	219.57	1.00	1.00	219.51	0.03	16.94	6.54	329.81	1.00	1.00	329.70	0.09	16.67	6.76	323.54	1.00	1.00	323.42	0.05
	100	12.71	5.99	197.48	1.00	1.00	197.41	0.03	15.82	7.58	307.74	1.00	1.00	307.64	0.06	15.25	7.84	293.91	1.00	1.01	293.75	0.08
	150	13.06	6.65	214.80	1.00	1.00	214.72	0.04	15.08	7.88	289.70	1.00	1.00	289.58	0.07	14.82	7.74	279.74	1.00	1.01	279.58	0.07
	200	11.86	7.78	201.18	1.00	1.00	201.08	0.05	12.83	9.19	249.06	1.00	1.00	248.93	0.08	13.51	9.03	264.11	1.00	1.01	263.92	0.09
	300	11.96	8.32	212.11	1.00	1.00	212.01	0.05	12.93	9.25	252.69	1.00	1.00	252.58	0.06	12.89	9.31	252.86	1.00	1.01	252.67	0.09
	400	12.41	8.53	226.68	1.00	1.01	226.56	0.06	13.23	9.60	267.30	1.00	1.00	267.17	0.07	12.89	9.16	249.93	1.00	1.01	249.76	0.08
	800	11.94	9.91	240.73	1.00	1.00	240.61	0.06	13.60	12.02	329.43	1.00	1.01	329.18	0.14	12.12	10.56	258.28	1.00	1.01	258.09	0.09
	1200	11.06	11.06	244.71	1.00	1.01	244.54	0.08	12.94	12.66	327.75	1.00	1.01	327.48	0.14	12.13	11.66	283.11	1.00	1.01	282.90	0.10
	1600	12.14	12.00	291.53	1.00	1.00	291.43	0.05	12.45	13.53	338.12	1.00	1.01	337.86	0.13	12.10	12.35	298.95	1.00	1.01	298.75	0.09
4	50	15.06	5.63	258.46	1.00	1.00	258.41	0.02	17.79	6.62	360.24	1.00	1.00	360.15	0.04	16.21	6.19	301.23	1.00	1.00	301.15	0.04
	100	13.37	6.55	221.79	1.00	1.00	221.67	0.07	14.91	7.08	272.52	1.00	1.00	272.46	0.03	14.79	7.19	270.31	1.00	1.00	270.19	0.06
	150	13.80	6.93	238.49	1.00	1.00	238.38	0.06	15.08	7.13	278.44	1.00	1.00	278.37	0.04	14.39	7.27	259.88	1.00	1.00	259.76	0.06
	200	13.01	7.44	224.65	1.00	1.00	224.53	0.06	14.57	7.87	274.15	1.00	1.00	274.07	0.04	13.59	8.04	249.31	1.00	1.01	249.15	0.07
	300	13.18	7.79	234.43	1.00	1.01	234.30	0.06	15.05	7.84	288.01	1.00	1.00	287.87	0.07	13.44	8.00	244.78	1.00	1.01	244.63	0.08
	400	12.42	8.46	225.98	1.00	1.01	225.84	0.07	13.42	8.29	248.80	1.00	1.00	248.68	0.06	12.86	8.55	238.37	1.00	1.01	238.20	0.09
	800	11.70	10.22	241.37	1.00	1.01	241.23	0.07	12.59	9.73	253.07	1.00	1.01	252.89	0.12	11.73	10.04	238.49	1.00	1.01	238.28	0.11
	1200	11.70	9.71	231.15	1.00	1.01	231.02	0.06	12.81	10.04	265.01	1.00	1.01	264.81	0.11	11.55	9.77	228.79	1.00	1.01	228.59	0.10
	1600	11.44	12.40	284.61	1.00	1.00	284.48	0.07	13.28	13.56	360.38	1.00	1.01	360.17	0.11	11.90	12.92	308.55	1.00	1.01	308.28	0.14
5	50	15.07	5.40	256.21	1.00	1.00	256.15	0.03	15.43	7.65	296.56	1.00	1.00	296.49	0.05	16.58	6.01	311.11	1.00	1.00	311.02	0.04
	100	12.81	6.56	207.02	1.00	1.00	206.96	0.03	12.81	6.75	209.68	1.00	1.00	209.61	0.04	14.50	7.61	268.08	1.00	1.01	267.94	0.07
	150	13.56	6.80	230.00	1.00	1.00	229.92	0.04	12.67	6.34	200.76	1.00	1.00	200.69	0.04	14.41	7.24	260.01	1.00	1.00	259.89	0.06
	200	13.12	7.40	226.92	1.00	1.00	226.83	0.05	12.60	7.63	216.98	1.00	1.00	216.89	0.05	13.86	8.05	257.01	1.00	1.01	256.86	0.07
	300	11.66	8.26	204.17	1.00	1.00	204.08	0.05	12.33	9.83	248.59	1.00	1.01	248.42	0.10	12.79	9.56	254.97	1.00	1.01	254.77	0.09
	400	11.52	8.84	210.73	1.00	1.00	210.64	0.05	12.33	10.35	259.05	1.00	1.01	258.82	0.13	12.42	9.66	247.65	1.00	1.01	247.47	0.09
	800	11.34	10.29	234.53	1.00	1.01	234.37	0.08	12.51	11.60	291.11	1.00	1.01	290.92	0.11	12.29	11.02	272.40	1.00	1.01	272.20	0.10
	1200	11.90	11.66	277.73	1.00	1.01	277.55	0.09	13.29	13.23	351.75	1.00	1.01	351.46	0.16	12.54	12.03	302.14	1.00	1.01	301.94	0.10
	1600	11.13	12.52	280.59	1.00	1.01	280.41	0.10	13.36	14.10	377.43	1.00	1.01	377.17	0.14	12.75	12.97	330.89	1.00	1.01	330.70	0.10

Table 9 – Fisher metrics for **CNN** representations on the **MLP** adapter architecture

REP	N	train μ_g	train std	train trace	train E_1	train erank	train λ_1	train λ_2	val μ_g	val std	val trace	val E_1	val erank	val λ_1	val λ_2	test μ_g	test std	test trace	test E_1	test erank	test λ_1	test λ_2
1	50	4.41	2.57	26.10	0.88	2.07	22.98	0.56	6.15	1.86	41.23	0.91	1.70	37.44	0.78	6.16	3.11	47.59	0.88	1.99	42.02	0.54
	100	4.67	3.38	33.25	0.92	1.69	30.54	0.40	5.78	2.69	40.65	0.94	1.49	38.20	0.38	6.45	4.23	59.53	0.92	1.63	54.94	0.47
	150	4.50	5.76	53.40	0.98	1.16	52.40	0.24	8.59	9.26	159.63	0.98	1.15	156.51	1.12	10.16	11.41	233.45	0.98	1.21	227.66	0.78
	200	5.69	6.98	81.04	0.98	1.15	79.66	0.23	8.65	8.92	154.51	0.98	1.14	151.79	0.88	9.90	11.31	225.89	0.98	1.18	221.01	0.61
	300	6.28	8.75	115.95	0.99	1.08	114.89	0.23	9.71	14.53	305.37	0.99	1.08	302.31	0.90	13.40	17.07	471.01	0.99	1.10	465.40	0.71
	400	7.86	10.01	161.95	0.99	1.10	160.15	0.30	9.64	11.36	222.14	0.99	1.12	219.07	0.59	11.56	13.57	317.79	0.99	1.13	313.15	0.54
	800	7.91	11.09	185.58	0.99	1.10	183.39	0.46	11.84	15.60	383.52	0.99	1.10	379.17	0.74	12.51	16.12	416.31	0.99	1.11	411.12	0.77
	1200	8.19	10.85	184.79	0.99	1.11	182.52	0.34	10.99	14.13	320.62	0.98	1.13	315.67	1.12	11.02	14.21	323.34	0.99	1.12	318.82	0.71
	1600	10.63	15.23	345.13	0.99	1.10	341.15	0.85	13.34	18.68	526.87	0.99	1.12	519.19	1.62	12.92	18.61	513.25	0.99	1.12	505.82	1.51
2	50	4.27	3.83	32.89	0.95	1.37	31.37	0.27	8.46	5.91	106.43	0.95	1.35	100.70	3.19	8.48	6.47	113.84	0.95	1.43	107.74	1.08
	100	5.72	5.24	60.09	0.94	1.46	56.75	0.54	7.45	5.66	87.54	0.94	1.41	82.56	1.72	7.63	6.00	94.27	0.94	1.46	88.89	0.83
	150	5.56	5.79	64.50	0.95	1.38	61.54	0.60	7.16	6.43	92.72	0.95	1.35	88.44	1.22	7.94	7.19	114.83	0.95	1.39	109.33	1.03
	200	5.69	5.70	64.87	0.97	1.28	62.70	0.34	8.83	8.33	147.33	0.97	1.25	142.79	0.84	8.59	8.35	143.53	0.96	1.29	138.29	0.98
	300	6.48	6.23	80.71	0.96	1.32	77.61	0.44	7.89	7.12	112.95	0.96	1.30	108.84	0.66	8.11	7.51	122.13	0.96	1.34	117.03	0.88
	400	6.85	9.62	139.44	0.99	1.08	138.15	0.27	18.99	23.19	898.51	0.99	1.07	890.92	1.64	16.52	21.30	726.51	0.99	1.10	718.45	1.46
	800	8.58	11.84	213.88	0.99	1.09	211.65	0.47	13.49	17.18	477.31	0.99	1.08	472.76	0.77	13.92	17.48	499.24	0.99	1.10	493.34	1.14
	1200	8.22	15.59	310.62	0.99	1.07	307.86	0.98	24.48	39.41	2152.34	0.99	1.07	2134.20	5.74	22.68	37.87	1948.47	0.99	1.09	1928.08	5.97
	1600	9.72	14.50	304.64	0.99	1.10	300.85	1.03	14.26	19.55	585.45	0.99	1.10	578.89	1.36	14.21	19.68	589.25	0.99	1.11	581.30	1.87
3	50	4.90	2.12	28.45	0.88	2.05	25.12	0.44	4.93	1.94	28.09	0.90	1.75	25.28	0.74	6.31	2.52	46.20	0.88	2.03	40.57	0.61
	100	5.10	4.80	49.02	0.96	1.34	46.99	0.36	6.11	5.45	66.96	0.95	1.34	63.80	0.96	7.49	6.39	96.88	0.95	1.40	92.14	0.76
	150	5.27	4.58	48.74	0.94	1.52	45.74	0.35	5.64	4.49	52.02	0.93	1.51	48.62	0.81	6.70	5.24	72.26	0.93	1.55	67.34	0.77
	200	5.57	6.85	78.02	0.98	1.15	76.65	0.18	8.05	10.64	177.99	0.99	1.12	175.36	0.58	9.77	11.57	229.48	0.98	1.17	224.70	0.85
	300	6.97	8.67	123.70	0.99	1.11	122.08	0.20	10.55	12.69	272.30	0.99	1.09	269.40	0.47	11.14	13.53	307.20	0.99	1.13	302.62	0.77
	400	7.78	10.21	164.75	0.99	1.10	162.87	0.27	13.59	17.53	492.05	0.99	1.11	485.88	1.10	13.05	16.22	433.39	0.99	1.12	427.53	0.93
	800	10.44	17.00	397.92	0.99	1.05	395.47	0.63	23.61	32.66	1624.08	0.99	1.07	1610.76	2.91	20.88	30.92	1391.73	0.99	1.07	1381.00	2.41
	1200	8.17	16.14	327.37	0.99	1.06	325.21	0.72	27.83	45.84	2875.79	0.99	1.08	2849.42	8.21	26.02	43.19	2542.51	0.99	1.07	2521.06	5.53
	1600	9.34	17.33	387.48	0.99	1.06	384.68	0.87	24.37	45.24	2640.39	0.99	1.09	2611.66	8.85	24.20	42.91	2426.99	0.99	1.08	2402.76	6.01
4	50	4.15	3.23	27.61	0.94	1.47	25.99	0.38	7.06	3.83	64.53	0.96	1.27	62.14	0.69	7.36	5.12	80.40	0.94	1.47	75.76	0.62
	100	4.84	5.35	52.03	0.97	1.21	50.64	0.45	8.22	7.57	124.80	0.98	1.16	122.23	0.71	9.05	8.83	159.87	0.97	1.22	155.59	0.68
	150	5.78	5.67	65.56	0.97	1.26	63.51	0.38	8.93	8.00	143.64	0.97	1.22	139.56	0.93	8.34	7.49	125.66	0.97	1.27	121.54	0.60
	200	6.03	7.11	87.01	0.98	1.14	85.62	0.24	11.21	11.82	265.36	0.99	1.12	261.65	0.82	10.16	11.18	228.24	0.98	1.15	224.07	0.76
	300	6.94	7.67	107.09	0.98	1.15	105.15	0.25	10.90	10.11	221.15	0.98	1.15	217.24	0.77	9.53	10.14	193.56	0.98	1.17	189.52	0.75
	400	6.32	9.59	131.90	0.99	1.06	131.03	0.19	16.36	22.65	780.84	0.99	1.06	775.33	1.66	15.12	22.48	733.65	0.99	1.07	727.62	1.15
	800	8.52	13.62	258.09	0.99	1.08	255.55	0.74	17.16	25.60	949.74	0.99	1.09	940.06	2.53	16.33	25.74	929.43	0.99	1.09	919.09	2.47
	1200	8.26	14.44	276.81	0.99	1.06	274.88	0.58	24.76	40.24	2232.61	0.99	1.08	2210.36	6.04	20.93	38.09	1889.18	0.99	1.09	1868.73	5.26
	1600	7.02	13.39	228.68	0.99	1.08	226.22	0.95	27.89	50.24	3301.99	0.99	1.10	3262.33	11.77	24.89	49.48	3067.73	0.99	1.10	3030.85	11.14
5	50	4.63	2.76	29.03	0.91	1.82	26.32	0.34	5.61	3.84	46.18	0.90	1.65	41.76	1.77	6.30	3.77	53.92	0.90	1.83	48.45	0.86
	100	4.73	6.44	63.82	0.99	1.12	62.89	0.15	4.87	5.81	57.46	0.98	1.13	56.45	0.44	9.77	11.38	225.04	0.98	1.19	219.86	1.22
	150	5.79	7.14	84.53	0.98	1.19	82.68	0.28	5.87	6.88	81.83	0.98	1.16	80.17	0.55	8.96	9.87	177.74	0.97	1.23	172.81	1.03
	200	6.08	6.37	77.61	0.98	1.20	75.77	0.25	7.52	7.93	119.48	0.97	1.20	116.42	0.88	8.85	8.90	157.51	0.97	1.23	153.02	0.86
	300	6.57	7.65	101.62	0.98	1.16	99.66	0.25	8.59	10.55	185.04	0.98	1.18	180.98	1.24	9.27	10.26	191.29	0.98	1.19	186.97	0.79
	400	6.65	6.90	91.72	0.97	1.24	89.06	0.33	7.77	8.19	127.49	0.96	1.28	123.03	1.15	8.11	8.05	130.51	0.97	1.26	126.29	0.79
	800	8.76	12.05	221.78	0.99	1.09	219.44	0.43	13.56	17.91	504.81	0.99	1.10	499.13	0.89	13.72	17.67	500.56	0.99	1.10	494.86	0.99
	1200	10.04	13.67	287.83	0.99	1.09	284.74	0.56	17.41	23.17	840.04	0.99	1.11	829.47	2.12	16.49	21.66	740.89	0.99	1.10	732.42	1.71
	1600	8.30	18.17	399.12	0.99	1.07	395.68	0.84	29.19	51.91	3547.03	0.99	1.10	3501.46	13.86	29.59	50.76	3452.40	0.99	1.10	3409.56	15.13

Table 10 – Fisher metrics for **CLIP** representations on the **Linear Probe** adapter architecture

REP	N	train							val							test						
		μ_g	std	trace	E_1	erank	λ_1	λ_2	μ_g	std	trace	E_1	erank	λ_1	λ_2	μ_g	std	trace	E_1	erank	λ_1	λ_2
1	50	8.63	1.99	78.38	1.00	1.00	78.34	0.03	8.69	1.36	77.39	1.00	1.00	77.36	0.03	9.22	1.89	88.63	1.00	1.00	88.60	0.02
	100	8.47	2.21	76.56	1.00	1.00	76.52	0.04	8.15	1.59	68.90	1.00	1.00	68.86	0.03	8.76	2.30	81.97	1.00	1.00	81.93	0.03
	150	8.30	2.54	75.32	1.00	1.00	75.28	0.04	8.32	2.68	76.49	1.00	1.01	76.44	0.04	8.40	2.50	76.85	1.00	1.00	76.81	0.04
	200	8.28	2.42	74.46	1.00	1.01	74.41	0.04	8.52	2.58	79.19	1.00	1.01	79.14	0.04	8.39	2.50	76.71	1.00	1.01	76.66	0.04
	300	8.11	2.59	72.51	1.00	1.01	72.46	0.04	8.02	2.58	71.05	1.00	1.01	71.00	0.04	8.23	2.62	74.61	1.00	1.01	74.56	0.04
	400	7.17	3.34	62.52	1.00	1.01	62.46	0.05	7.17	3.46	63.46	1.00	1.01	63.40	0.05	7.31	3.43	65.12	1.00	1.01	65.06	0.05
	800	5.20	5.15	53.54	1.00	1.01	53.50	0.03	5.65	5.55	62.76	1.00	1.01	62.70	0.05	5.49	5.35	58.75	1.00	1.01	58.70	0.04
	1200	4.87	5.74	56.66	1.00	1.01	56.63	0.03	5.29	6.13	65.61	1.00	1.01	65.54	0.06	5.21	5.92	62.17	1.00	1.01	62.13	0.03
	1600	4.60	6.25	60.24	1.00	1.01	60.19	0.04	4.70	6.11	59.46	1.00	1.01	59.39	0.05	5.03	6.34	65.50	1.00	1.01	65.46	0.03
2	50	9.43	1.97	92.84	1.00	1.00	92.81	0.03	8.77	1.36	78.69	1.00	1.00	78.66	0.02	9.60	1.98	96.11	1.00	1.00	96.08	0.02
	100	9.08	2.05	86.65	1.00	1.00	86.61	0.03	8.76	1.52	79.12	1.00	1.00	79.09	0.02	9.24	2.05	89.53	1.00	1.00	89.50	0.03
	150	8.83	2.21	82.78	1.00	1.00	82.74	0.03	8.51	1.72	75.40	1.00	1.00	75.37	0.03	8.92	2.22	84.43	1.00	1.00	84.39	0.03
	200	7.98	2.75	71.27	1.00	1.01	71.21	0.05	7.96	2.49	69.60	1.00	1.01	69.55	0.04	8.14	2.81	74.22	1.00	1.01	74.17	0.04
	300	7.97	2.97	72.28	1.00	1.01	72.23	0.04	8.02	2.65	71.31	1.00	1.01	71.27	0.04	8.12	3.01	74.95	1.00	1.01	74.90	0.04
	400	6.44	4.04	57.76	1.00	1.01	57.70	0.04	7.02	4.04	65.69	1.00	1.01	65.62	0.05	6.73	4.28	63.63	1.00	1.01	63.57	0.05
	800	5.34	5.57	59.59	1.00	1.01	59.54	0.04	5.23	5.08	53.20	1.00	1.01	53.15	0.04	5.72	5.88	67.32	1.00	1.01	67.27	0.04
	1200	4.73	5.70	54.77	1.00	1.01	54.73	0.04	4.85	5.69	55.92	1.00	1.01	55.89	0.03	5.46	6.62	73.63	1.00	1.01	73.58	0.04
	1600	4.90	6.83	70.69	1.00	1.01	70.64	0.04	4.52	6.09	57.57	1.00	1.00	57.54	0.03	5.30	7.26	80.71	1.00	1.01	80.66	0.04
3	50	8.46	2.38	77.25	1.00	1.00	77.20	0.04	8.03	1.36	66.38	1.00	1.01	66.34	0.04	9.36	2.47	93.79	1.00	1.00	93.75	0.03
	100	9.06	2.05	86.28	1.00	1.00	86.25	0.03	8.58	2.06	77.84	1.00	1.01	77.79	0.05	9.34	2.01	91.27	1.00	1.00	91.24	0.03
	150	8.56	2.19	78.11	1.00	1.00	78.07	0.03	8.29	2.30	73.95	1.00	1.01	73.89	0.05	8.92	2.24	84.62	1.00	1.00	84.58	0.03
	200	8.13	2.62	72.90	1.00	1.01	72.86	0.04	8.14	2.68	73.42	1.00	1.01	73.37	0.04	8.46	2.67	78.71	1.00	1.00	78.67	0.04
	300	6.58	3.57	56.08	1.00	1.01	56.02	0.05	6.82	4.06	62.89	1.00	1.01	62.82	0.06	7.00	3.86	63.96	1.00	1.01	63.89	0.06
	400	7.26	3.27	63.40	1.00	1.01	63.34	0.05	7.17	3.40	63.01	1.00	1.01	62.94	0.06	7.41	3.37	66.20	1.00	1.01	66.14	0.05
	800	5.89	4.27	52.88	1.00	1.01	52.83	0.04	6.66	4.85	67.92	1.00	1.01	67.86	0.05	6.29	4.60	60.66	1.00	1.01	60.60	0.05
	1200	4.93	5.92	59.37	1.00	1.01	59.32	0.04	5.49	6.86	77.17	1.00	1.01	77.10	0.05	5.32	6.52	70.82	1.00	1.01	70.77	0.04
	1600	5.24	6.12	64.85	1.00	1.01	64.81	0.03	5.25	6.22	66.20	1.00	1.01	66.15	0.04	5.41	6.28	68.60	1.00	1.01	68.56	0.04
4	50	9.05	1.85	85.32	1.00	1.00	85.30	0.02	8.97	2.23	85.37	1.00	1.00	85.34	0.02	9.46	1.91	93.07	1.00	1.00	93.05	0.02
	100	7.95	2.34	68.61	1.00	1.01	68.57	0.04	8.22	2.62	74.39	1.00	1.01	74.33	0.05	8.37	2.50	76.26	1.00	1.01	76.21	0.04
	150	8.82	2.10	82.19	1.00	1.00	82.15	0.03	8.81	2.12	82.08	1.00	1.00	82.04	0.04	9.02	2.12	85.87	1.00	1.00	85.83	0.03
	200	8.27	2.55	74.87	1.00	1.01	74.82	0.04	8.27	2.35	73.86	1.00	1.01	73.81	0.04	8.54	2.49	79.16	1.00	1.01	79.11	0.04
	300	6.27	4.01	55.39	1.00	1.01	55.34	0.05	6.53	4.38	61.86	1.00	1.01	61.79	0.06	6.75	4.17	63.05	1.00	1.01	62.99	0.05
	400	6.50	4.06	58.80	1.00	1.01	58.74	0.05	6.56	4.16	60.37	1.00	1.01	60.31	0.05	6.80	4.01	62.25	1.00	1.01	62.19	0.05
	800	5.30	4.86	51.71	1.00	1.01	51.65	0.04	6.07	5.20	63.87	1.00	1.01	63.80	0.05	6.04	5.07	62.19	1.00	1.01	62.13	0.05
	1200	5.03	6.00	61.25	1.00	1.01	61.21	0.03	5.46	6.42	71.00	1.00	1.01	70.94	0.05	5.44	6.18	67.73	1.00	1.01	67.68	0.04
	1600	4.56	6.68	65.44	1.00	1.01	65.39	0.04	5.13	7.08	76.46	1.00	1.01	76.40	0.05	5.21	6.94	75.30	1.00	1.01	75.25	0.04
5	50	8.70	2.40	81.45	1.00	1.00	81.41	0.03	8.42	1.63	73.52	1.00	1.00	73.48	0.03	9.43	2.37	94.56	1.00	1.00	94.52	0.03
	100	7.81	2.54	67.45	1.00	1.01	67.40	0.04	7.20	2.38	57.42	1.00	1.01	57.37	0.04	8.47	2.85	79.79	1.00	1.01	79.73	0.04
	150	7.58	2.88	65.73	1.00	1.01	65.68	0.05	7.26	2.95	61.46	1.00	1.01	61.41	0.05	7.98	3.11	73.28	1.00	1.01	73.22	0.05
	200	8.33	2.43	75.34	1.00	1.01	75.29	0.04	8.14	2.52	72.66	1.00	1.01	72.61	0.04	8.62	2.62	81.11	1.00	1.01	81.06	0.04
	300	6.96	3.48	60.53	1.00	1.01	60.47	0.05	6.59	3.51	55.76	1.00	1.01	55.71	0.04	7.26	3.62	65.74	1.00	1.01	65.68	0.05
	400	6.97	3.59	61.49	1.00	1.01	61.44	0.05	6.83	3.66	60.00	1.00	1.01	59.95	0.04	7.21	3.70	65.70	1.00	1.01	65.64	0.05
	800	5.13	4.97	51.07	1.00	1.01	51.02	0.04	5.49	5.54	60.78	1.00	1.01	60.74	0.03	5.61	5.47	61.46	1.00	1.01	61.42	0.03
	1200	5.24	6.03	63.88	1.00	1.01	63.84	0.03	5.42	6.28	68.75	1.00	1.00	68.72	0.03	5.36	6.00	64.71	1.00	1.01	64.67	0.03
	1600	4.76	6.69	67.49	1.00	1.01	67.43	0.05	5.17	6.89	74.15	1.00	1.00	74.12	0.03	5.16	6.64	70.65	1.00	1.00	70.62	0.03

Table 11 – Fisher metrics for **CLIP** representations on the **MLP** adapter architecture

REP	N	train μ_g	train std	train trace	train E_1	train erank	train λ_1	train λ_2	val μ_g	val std	val trace	val E_1	val erank	val λ_1	val λ_2	test μ_g	test std	test trace	test E_1	test erank	test λ_1	test λ_2
1	50	3.51	0.49	12.56	0.94	1.48	11.84	0.14	3.47	0.29	12.14	0.96	1.33	11.61	0.14	3.62	0.41	13.27	0.95	1.44	12.55	0.11
	100	3.35	3.81	25.72	0.98	1.12	25.34	0.13	2.63	2.07	11.18	0.99	1.08	11.06	0.06	4.04	3.98	32.22	0.99	1.12	31.77	0.11
	150	3.88	4.92	39.20	0.99	1.05	38.98	0.06	4.35	6.40	59.90	0.99	1.06	59.47	0.24	4.75	5.85	56.74	0.99	1.05	56.43	0.07
	200	3.71	6.06	50.49	1.00	1.03	50.35	0.05	4.20	7.33	71.38	0.99	1.04	71.06	0.22	4.85	7.42	78.54	1.00	1.03	78.30	0.06
	300	4.11	6.95	65.23	1.00	1.03	65.02	0.07	3.51	6.18	50.59	1.00	1.03	50.38	0.12	4.67	7.41	76.66	1.00	1.03	76.40	0.06
	400	4.46	7.28	72.91	1.00	1.03	72.69	0.07	4.51	7.88	82.41	1.00	1.04	82.04	0.17	4.72	7.29	75.51	1.00	1.03	75.28	0.05
	800	4.65	9.55	112.89	1.00	1.01	112.73	0.04	5.51	10.45	139.51	1.00	1.02	139.20	0.15	4.73	9.49	112.41	1.00	1.02	112.19	0.06
	1200	4.25	11.64	153.64	1.00	1.01	153.56	0.02	5.05	13.41	205.22	1.00	1.02	204.64	0.38	5.48	14.23	232.56	1.00	1.01	232.27	0.12
	1600	3.81	11.34	143.22	1.00	1.01	143.06	0.07	4.41	12.30	170.76	1.00	1.02	170.34	0.27	5.17	13.55	210.46	1.00	1.01	210.21	0.08
2	50	3.54	1.54	14.92	0.96	1.29	14.38	0.15	3.65	1.19	14.71	0.97	1.23	14.25	0.14	4.01	1.44	18.16	0.97	1.27	17.55	0.13
	100	3.70	2.56	20.26	0.98	1.18	19.82	0.14	2.99	1.38	10.85	0.98	1.18	10.59	0.11	3.98	2.65	22.90	0.98	1.18	22.40	0.13
	150	3.74	4.44	33.73	0.99	1.07	33.46	0.08	3.20	3.37	21.61	0.99	1.06	21.46	0.04	4.75	5.48	52.61	0.99	1.06	52.24	0.09
	200	4.15	5.49	47.32	1.00	1.03	47.16	0.04	3.98	5.31	44.07	1.00	1.03	43.90	0.04	5.46	7.60	87.63	1.00	1.04	87.29	0.10
	300	3.63	7.53	69.83	1.00	1.01	69.74	0.03	3.78	8.03	78.81	1.00	1.01	78.72	0.03	5.68	11.25	158.80	1.00	1.02	158.49	0.10
	400	3.69	8.10	79.17	1.00	1.01	79.06	0.03	5.40	11.62	164.23	1.00	1.01	164.07	0.06	5.71	12.32	184.51	1.00	1.02	184.18	0.14
	800	4.38	11.10	142.46	1.00	1.02	142.21	0.14	4.10	11.02	138.25	1.00	1.01	138.14	0.03	5.84	14.83	254.06	1.00	1.01	253.67	0.18
	1200	4.16	11.95	160.07	1.00	1.01	159.91	0.07	4.29	12.11	165.15	1.00	1.01	165.04	0.03	6.53	18.11	370.46	1.00	1.01	370.01	0.20
	1600	5.19	14.48	236.69	1.00	1.01	236.39	0.14	4.12	13.09	188.31	1.00	1.01	188.16	0.04	6.72	19.88	440.51	1.00	1.01	439.94	0.25
3	50	3.48	0.39	12.25	0.95	1.45	11.59	0.11	3.41	0.25	11.72	0.95	1.36	11.15	0.14	3.60	0.44	13.18	0.94	1.45	12.46	0.10
	100	3.46	0.69	12.48	0.95	1.39	11.89	0.11	3.45	0.77	12.51	0.94	1.44	11.81	0.15	3.57	0.73	13.29	0.95	1.41	12.63	0.10
	150	3.60	2.65	19.99	0.98	1.14	19.66	0.09	3.71	2.85	21.84	0.98	1.14	21.47	0.11	4.21	3.09	27.26	0.98	1.14	26.80	0.11
	200	3.48	5.71	44.76	1.00	1.02	44.64	0.03	5.23	7.58	84.76	1.00	1.03	84.50	0.14	5.07	7.76	85.91	1.00	1.03	85.64	0.07
	300	3.46	6.50	54.14	1.00	1.02	54.02	0.03	4.71	8.37	92.27	1.00	1.02	92.03	0.10	4.91	8.58	97.77	1.00	1.03	97.50	0.08
	400	3.35	7.83	72.47	1.00	1.01	72.38	0.02	4.50	10.79	136.52	1.00	1.02	136.27	0.11	5.13	11.53	159.37	1.00	1.02	159.12	0.09
	800	3.39	8.81	89.12	1.00	1.01	89.03	0.04	6.59	15.49	283.45	1.00	1.02	282.83	0.27	5.16	12.48	182.48	1.00	1.01	182.29	0.07
	1200	4.53	8.33	89.87	1.00	1.02	89.65	0.10	5.03	9.76	120.49	1.00	1.03	120.16	0.12	5.17	9.56	118.08	1.00	1.02	117.87	0.06
	1600	4.52	12.63	179.90	1.00	1.01	179.81	0.03	5.36	15.96	283.41	1.00	1.01	282.96	0.21	5.51	14.62	244.14	1.00	1.01	243.94	0.09
4	50	3.59	0.42	13.04	0.94	1.47	12.31	0.11	3.52	0.33	12.47	0.95	1.37	11.86	0.14	3.67	0.42	13.65	0.94	1.46	12.88	0.11
	100	3.17	4.08	26.66	0.99	1.05	26.51	0.04	3.15	3.76	24.03	0.99	1.05	23.90	0.05	4.83	6.10	60.54	0.99	1.06	60.11	0.11
	150	3.89	3.49	27.33	0.99	1.09	27.04	0.07	3.58	2.79	20.62	0.99	1.09	20.40	0.06	4.60	3.98	37.01	0.99	1.10	36.57	0.12
	200	3.91	5.46	45.18	0.99	1.04	44.97	0.06	3.08	3.92	24.85	1.00	1.03	24.75	0.03	4.79	6.54	65.70	1.00	1.04	65.42	0.06
	300	3.44	6.74	57.20	1.00	1.02	57.10	0.03	4.28	10.32	124.86	1.00	1.02	124.52	0.22	5.04	9.68	119.18	1.00	1.02	118.97	0.05
	400	4.05	8.02	80.72	1.00	1.02	80.57	0.04	5.13	11.35	155.16	1.00	1.02	154.77	0.16	5.00	9.66	118.28	1.00	1.02	118.08	0.05
	800	3.72	9.41	102.42	1.00	1.02	102.24	0.06	5.47	12.40	183.76	1.00	1.02	183.31	0.23	5.20	11.54	160.06	1.00	1.01	159.83	0.08
	1200	4.33	11.53	151.70	1.00	1.01	151.50	0.07	5.52	13.43	210.89	1.00	1.02	210.48	0.21	5.29	13.11	199.92	1.00	1.01	199.65	0.10
	1600	3.96	13.05	186.03	1.00	1.01	185.89	0.03	6.50	18.75	393.99	1.00	1.01	393.52	0.24	6.17	18.31	373.31	1.00	1.01	372.92	0.15
5	50	3.64	1.98	17.18	0.96	1.30	16.52	0.19	3.46	1.54	14.32	0.97	1.23	13.84	0.20	3.90	1.80	18.45	0.96	1.31	17.74	0.16
	100	3.52	3.84	27.15	0.99	1.08	26.88	0.09	2.82	3.56	20.64	0.99	1.06	20.49	0.06	4.35	4.54	39.55	0.99	1.10	39.10	0.10
	150	3.77	4.47	34.17	0.99	1.06	33.94	0.06	4.43	5.65	51.49	1.00	1.04	51.27	0.06	4.38	5.21	46.34	0.99	1.07	45.97	0.08
	200	3.91	3.41	26.93	0.99	1.11	26.56	0.11	4.06	3.78	30.79	0.99	1.08	30.49	0.07	4.12	3.61	29.98	0.99	1.12	29.56	0.09
	300	4.03	4.69	38.29	0.99	1.06	38.04	0.06	4.17	5.00	42.32	0.99	1.04	42.13	0.04	4.48	5.22	47.30	0.99	1.06	46.98	0.06
	400	4.09	7.76	76.97	1.00	1.02	76.85	0.03	5.15	9.26	112.31	1.00	1.01	112.14	0.05	4.96	8.62	98.96	1.00	1.02	98.77	0.04
	800	3.57	9.13	96.08	1.00	1.01	95.95	0.06	6.18	14.07	236.14	1.00	1.01	235.86	0.13	5.08	11.93	168.27	1.00	1.01	168.04	0.07
	1200	5.36	13.02	198.15	1.00	1.01	198.02	0.04	7.22	16.72	331.67	1.00	1.01	331.32	0.14	6.60	15.71	290.47	1.00	1.01	290.18	0.08
	1600	4.44	12.67	180.14	1.00	1.01	179.96	0.05	6.44	15.97	296.48	1.00	1.01	296.21	0.10	5.76	14.79	252.02	1.00	1.01	251.79	0.06

Table 12 – Fisher metrics for **DINO** representations on the **Linear Probe** adapter architecture

REP	N	train							val							test						
		μ_g	std	trace	E_1	erank	λ_1	λ_2	μ_g	std	trace	E_1	erank	λ_1	λ_2	μ_g	std	trace	E_1	erank	λ_1	λ_2
1	50	0.96	0.44	1.11	1.00	1.00	1.10	0.00	0.89	0.27	0.86	1.00	1.00	0.86	0.00	1.09	0.45	1.39	1.00	1.00	1.39	0.00
	100	0.79	0.50	0.88	1.00	1.01	0.88	0.00	0.74	0.33	0.66	1.00	1.01	0.66	0.00	0.93	0.54	1.16	1.00	1.01	1.16	0.00
	150	0.79	0.52	0.90	1.00	1.01	0.90	0.00	0.77	0.46	0.81	1.00	1.01	0.81	0.00	0.88	0.55	1.09	1.00	1.01	1.08	0.00
	200	0.67	0.54	0.74	1.00	1.01	0.74	0.00	0.64	0.53	0.69	1.00	1.01	0.69	0.00	0.77	0.62	0.98	1.00	1.01	0.98	0.00
	300	0.62	0.59	0.74	1.00	1.01	0.74	0.00	0.57	0.50	0.57	1.00	1.01	0.57	0.00	0.70	0.66	0.92	1.00	1.01	0.92	0.00
	400	0.64	0.62	0.80	1.00	1.01	0.80	0.00	0.62	0.61	0.76	1.00	1.01	0.75	0.00	0.69	0.66	0.92	1.00	1.01	0.92	0.00
	800	0.56	0.70	0.80	1.00	1.01	0.80	0.00	0.67	0.81	1.10	1.00	1.01	1.10	0.00	0.61	0.73	0.90	1.00	1.01	0.90	0.00
	1200	0.51	0.74	0.82	1.00	1.01	0.82	0.00	0.63	0.89	1.19	1.00	1.01	1.19	0.00	0.57	0.83	1.02	1.00	1.01	1.01	0.00
	1600	0.52	0.79	0.89	1.00	1.01	0.89	0.00	0.58	0.90	1.14	1.00	1.01	1.14	0.00	0.55	0.89	1.10	1.00	1.01	1.10	0.00
2	50	1.02	0.38	1.19	1.00	1.00	1.19	0.00	0.95	0.48	1.14	1.00	1.00	1.14	0.00	1.10	0.45	1.40	1.00	1.00	1.40	0.00
	100	0.87	0.47	0.97	1.00	1.01	0.97	0.00	0.86	0.45	0.94	1.00	1.00	0.94	0.00	0.94	0.56	1.20	1.00	1.01	1.20	0.00
	150	0.84	0.50	0.95	1.00	1.01	0.95	0.00	0.78	0.44	0.79	1.00	1.01	0.79	0.00	0.89	0.58	1.14	1.00	1.01	1.13	0.00
	200	1.02	0.45	1.24	1.00	1.00	1.24	0.00	0.95	0.40	1.06	1.00	1.00	1.06	0.00	1.03	0.50	1.32	1.00	1.00	1.31	0.00
	300	0.70	0.59	0.83	1.00	1.01	0.83	0.00	0.65	0.53	0.70	1.00	1.01	0.70	0.00	0.75	0.68	1.02	1.00	1.01	1.02	0.00
	400	0.66	0.61	0.82	1.00	1.01	0.81	0.00	0.69	0.62	0.86	1.00	1.01	0.86	0.00	0.72	0.71	1.02	1.00	1.01	1.02	0.00
	800	0.68	0.65	0.89	1.00	1.01	0.89	0.00	0.68	0.62	0.84	1.00	1.01	0.84	0.00	0.71	0.70	0.99	1.00	1.01	0.99	0.00
	1200	0.54	0.72	0.81	1.00	1.01	0.81	0.00	0.56	0.73	0.85	1.00	1.01	0.85	0.00	0.64	0.88	1.19	1.00	1.01	1.19	0.00
	1600	0.56	0.76	0.90	1.00	1.01	0.90	0.00	0.53	0.70	0.77	1.00	1.01	0.76	0.00	0.63	0.89	1.19	1.00	1.01	1.19	0.00
3	50	0.97	0.38	1.09	1.00	1.00	1.09	0.00	0.91	0.35	0.96	1.00	1.00	0.96	0.00	1.09	0.44	1.38	1.00	1.00	1.38	0.00
	100	0.93	0.45	1.06	1.00	1.00	1.06	0.00	0.84	0.45	0.90	1.00	1.00	0.90	0.00	1.03	0.51	1.32	1.00	1.00	1.32	0.00
	150	1.01	0.40	1.18	1.00	1.00	1.18	0.00	0.98	0.39	1.11	1.00	1.00	1.11	0.00	1.09	0.44	1.39	1.00	1.00	1.39	0.00
	200	0.75	0.54	0.86	1.00	1.01	0.86	0.00	0.72	0.54	0.82	1.00	1.01	0.82	0.00	0.84	0.59	1.05	1.00	1.01	1.05	0.00
	300	0.67	0.60	0.81	1.00	1.01	0.81	0.00	0.68	0.64	0.88	1.00	1.01	0.88	0.00	0.76	0.65	0.99	1.00	1.01	0.99	0.00
	400	0.72	0.63	0.91	1.00	1.01	0.91	0.00	0.67	0.61	0.82	1.00	1.01	0.82	0.00	0.76	0.63	0.97	1.00	1.01	0.97	0.00
	800	0.58	0.63	0.73	1.00	1.01	0.73	0.00	0.67	0.72	0.97	1.00	1.01	0.97	0.00	0.66	0.69	0.91	1.00	1.01	0.91	0.00
	1200	0.56	0.75	0.87	1.00	1.01	0.87	0.00	0.59	0.79	0.98	1.00	1.01	0.98	0.00	0.61	0.75	0.94	1.00	1.01	0.94	0.00
	1600	0.58	0.81	0.99	1.00	1.01	0.99	0.00	0.54	0.82	0.95	1.00	1.01	0.95	0.00	0.58	0.82	1.01	1.00	1.01	1.01	0.00
4	50	1.09	0.37	1.33	1.00	1.00	1.33	0.00	1.10	0.28	1.28	1.00	1.00	1.28	0.00	1.20	0.39	1.60	1.00	1.00	1.60	0.00
	100	0.84	0.49	0.95	1.00	1.01	0.95	0.00	0.93	0.49	1.10	1.00	1.01	1.10	0.00	0.98	0.55	1.26	1.00	1.01	1.26	0.00
	150	0.86	0.50	0.99	1.00	1.01	0.99	0.00	0.92	0.48	1.08	1.00	1.01	1.08	0.00	0.96	0.56	1.23	1.00	1.01	1.22	0.00
	200	0.77	0.56	0.92	1.00	1.01	0.92	0.00	0.80	0.51	0.90	1.00	1.01	0.90	0.00	0.86	0.60	1.10	1.00	1.01	1.10	0.00
	300	0.66	0.61	0.81	1.00	1.01	0.81	0.00	0.72	0.55	0.82	1.00	1.01	0.82	0.00	0.74	0.66	0.98	1.00	1.01	0.97	0.00
	400	0.66	0.63	0.84	1.00	1.01	0.84	0.00	0.69	0.60	0.84	1.00	1.01	0.84	0.00	0.71	0.66	0.93	1.00	1.01	0.93	0.00
	800	0.55	0.67	0.75	1.00	1.01	0.75	0.00	0.67	0.74	0.99	1.00	1.01	0.99	0.00	0.63	0.73	0.92	1.00	1.01	0.92	0.00
	1200	0.54	0.74	0.84	1.00	1.01	0.84	0.00	0.61	0.80	1.01	1.00	1.01	1.01	0.00	0.59	0.78	0.95	1.00	1.01	0.95	0.00
	1600	0.49	0.75	0.80	1.00	1.01	0.80	0.00	0.56	0.84	1.02	1.00	1.01	1.02	0.00	0.57	0.83	1.02	1.00	1.00	1.02	0.00
5	50	1.16	0.31	1.44	1.00	1.00	1.44	0.00	1.12	0.28	1.34	1.00	1.00	1.34	0.00	1.21	0.34	1.57	1.00	1.00	1.57	0.00
	100	0.86	0.47	0.96	1.00	1.01	0.96	0.00	0.74	0.34	0.67	1.00	1.01	0.67	0.00	0.94	0.52	1.16	1.00	1.01	1.15	0.00
	150	0.84	0.51	0.97	1.00	1.01	0.97	0.00	0.76	0.42	0.75	1.00	1.01	0.75	0.00	0.89	0.54	1.08	1.00	1.01	1.08	0.00
	200	0.83	0.53	0.97	1.00	1.01	0.97	0.00	0.73	0.46	0.74	1.00	1.01	0.74	0.00	0.84	0.58	1.04	1.00	1.01	1.04	0.00
	300	0.75	0.59	0.91	1.00	1.01	0.91	0.00	0.63	0.47	0.62	1.00	1.01	0.62	0.00	0.75	0.61	0.93	1.00	1.01	0.93	0.00
	400	0.66	0.64	0.85	1.00	1.01	0.84	0.00	0.62	0.61	0.76	1.00	1.01	0.76	0.00	0.67	0.65	0.87	1.00	1.01	0.87	0.00
	800	0.58	0.70	0.83	1.00	1.01	0.83	0.00	0.61	0.75	0.93	1.00	1.01	0.93	0.00	0.59	0.72	0.86	1.00	1.01	0.86	0.00
	1200	0.56	0.75	0.87	1.00	1.01	0.87	0.00	0.58	0.80	0.97	1.00	1.01	0.97	0.00	0.58	0.77	0.92	1.00	1.01	0.92	0.00
	1600	0.54	0.85	1.02	1.00	1.01	1.02	0.00	0.56	0.85	1.03	1.00	1.01	1.03	0.00	0.55	0.82	0.96	1.00	1.01	0.96	0.00

Table 13 – Fisher metrics for **DINO** representations on the **MLP** adapter architecture

REP	N	train μ_g	train std	train trace	train E_1	train erank	train λ_1	train λ_2	val μ_g	val std	val trace	val E_1	val erank	val λ_1	val λ_2	test μ_g	test std	test trace	test E_1	test erank	test λ_1	test λ_2
1	50	0.33	0.22	0.16	0.93	1.62	0.14	0.00	0.31	0.19	0.13	0.92	1.57	0.12	0.00	0.41	0.24	0.23	0.93	1.60	0.21	0.00
	100	0.32	0.34	0.21	0.95	1.42	0.20	0.00	0.25	0.21	0.11	0.96	1.30	0.10	0.00	0.40	0.35	0.29	0.95	1.42	0.27	0.00
	150	0.27	0.43	0.25	0.99	1.08	0.25	0.00	0.36	0.65	0.56	0.99	1.06	0.55	0.00	0.50	0.78	0.86	0.99	1.11	0.85	0.00
	200	0.30	0.50	0.34	0.99	1.07	0.34	0.00	0.32	0.62	0.48	0.99	1.06	0.48	0.00	0.52	0.84	0.98	0.99	1.10	0.97	0.00
	300	0.35	0.58	0.46	0.99	1.09	0.46	0.00	0.30	0.51	0.35	0.99	1.08	0.35	0.00	0.46	0.73	0.74	0.99	1.11	0.73	0.00
	400	0.38	0.71	0.65	0.99	1.07	0.64	0.00	0.42	0.81	0.83	0.99	1.07	0.83	0.00	0.47	0.84	0.93	0.99	1.09	0.92	0.00
	800	0.42	0.94	1.05	1.00	1.03	1.05	0.00	0.75	1.47	2.72	1.00	1.04	2.70	0.00	0.55	1.19	1.70	0.99	1.06	1.69	0.00
	1200	0.44	1.04	1.28	1.00	1.04	1.27	0.00	0.73	1.56	2.95	0.99	1.05	2.94	0.00	0.59	1.36	2.18	0.99	1.05	2.17	0.00
	1600	0.45	1.30	1.90	1.00	1.02	1.90	0.00	0.84	2.18	5.46	0.99	1.04	5.43	0.01	0.73	2.07	4.81	0.99	1.05	4.79	0.01
2	50	0.33	0.27	0.18	0.95	1.38	0.18	0.00	0.37	0.35	0.26	0.96	1.26	0.25	0.00	0.45	0.35	0.32	0.95	1.39	0.31	0.00
	100	0.34	0.38	0.26	0.98	1.18	0.25	0.00	0.38	0.49	0.39	0.98	1.13	0.38	0.00	0.49	0.61	0.61	0.97	1.22	0.59	0.00
	150	0.25	0.39	0.22	1.00	1.04	0.21	0.00	0.31	0.70	0.58	1.00	1.02	0.58	0.00	0.68	1.21	1.92	0.99	1.07	1.90	0.00
	200	0.34	0.60	0.47	0.99	1.06	0.47	0.00	0.37	0.76	0.71	0.99	1.06	0.70	0.00	0.65	1.11	1.66	0.99	1.07	1.65	0.00
	300	0.40	0.68	0.63	0.99	1.05	0.62	0.00	0.41	0.69	0.64	0.99	1.06	0.63	0.00	0.63	1.06	1.52	0.99	1.08	1.50	0.00
	400	0.47	0.85	0.94	1.00	1.04	0.94	0.00	0.54	0.97	1.24	0.99	1.05	1.23	0.00	0.72	1.28	2.16	0.99	1.05	2.15	0.00
	800	0.48	0.82	0.90	0.99	1.08	0.89	0.00	0.47	0.77	0.81	0.99	1.09	0.80	0.00	0.58	0.97	1.26	0.99	1.09	1.25	0.00
	1200	0.42	1.21	1.63	1.00	1.03	1.63	0.00	0.70	1.86	3.94	0.99	1.05	3.92	0.01	0.84	2.09	5.09	1.00	1.04	5.07	0.01
	1600	0.44	1.29	1.85	1.00	1.03	1.85	0.00	0.67	1.81	3.72	0.99	1.06	3.69	0.02	0.85	2.15	5.32	1.00	1.04	5.30	0.01
3	50	0.33	0.23	0.16	0.96	1.33	0.15	0.00	0.33	0.20	0.15	0.96	1.27	0.14	0.00	0.44	0.31	0.29	0.95	1.43	0.28	0.00
	100	0.27	0.43	0.26	0.99	1.08	0.25	0.00	0.33	0.47	0.32	0.99	1.10	0.32	0.00	0.52	0.73	0.81	0.98	1.14	0.79	0.00
	150	0.34	0.46	0.33	0.98	1.19	0.32	0.00	0.33	0.49	0.35	0.98	1.15	0.34	0.00	0.47	0.57	0.54	0.98	1.20	0.53	0.00
	200	0.31	0.58	0.43	0.99	1.06	0.42	0.00	0.48	0.83	0.92	1.00	1.04	0.92	0.00	0.54	0.88	1.06	0.99	1.08	1.05	0.00
	300	0.29	0.60	0.44	1.00	1.03	0.44	0.00	0.51	1.06	1.37	1.00	1.03	1.37	0.00	0.63	1.24	1.93	0.99	1.05	1.92	0.00
	400	0.41	0.75	0.73	0.99	1.05	0.72	0.00	0.50	0.90	1.06	0.99	1.06	1.05	0.00	0.60	1.04	1.45	0.99	1.07	1.44	0.00
	800	0.40	0.99	1.13	1.00	1.02	1.13	0.00	0.86	1.82	4.04	1.00	1.04	4.02	0.01	0.82	1.75	3.74	1.00	1.04	3.73	0.01
	1200	0.37	1.06	1.26	1.00	1.04	1.25	0.00	0.71	1.77	3.65	0.99	1.05	3.63	0.01	0.75	1.76	3.64	0.99	1.05	3.62	0.01
	1600	0.44	1.22	1.67	1.00	1.02	1.67	0.00	0.72	2.01	4.56	0.99	1.05	4.53	0.01	0.82	2.07	4.97	0.99	1.05	4.94	0.01
4	50	0.30	0.27	0.16	0.94	1.48	0.15	0.00	0.32	0.22	0.15	0.96	1.31	0.15	0.00	0.43	0.30	0.28	0.95	1.42	0.26	0.00
	100	0.29	0.48	0.31	0.99	1.10	0.31	0.00	0.28	0.34	0.19	0.99	1.10	0.19	0.00	0.48	0.67	0.69	0.99	1.13	0.68	0.00
	150	0.33	0.59	0.46	0.99	1.09	0.46	0.00	0.32	0.45	0.30	0.99	1.07	0.30	0.00	0.49	0.75	0.79	0.99	1.11	0.78	0.00
	200	0.36	0.60	0.49	0.99	1.08	0.48	0.00	0.36	0.49	0.37	0.99	1.06	0.37	0.00	0.51	0.77	0.86	0.99	1.08	0.85	0.00
	300	0.36	0.56	0.44	0.99	1.11	0.44	0.00	0.39	0.56	0.46	0.99	1.08	0.46	0.00	0.48	0.71	0.73	0.99	1.11	0.72	0.00
	400	0.37	0.77	0.73	1.00	1.04	0.73	0.00	0.49	0.93	1.11	1.00	1.04	1.10	0.00	0.56	1.07	1.46	0.99	1.05	1.45	0.00
	800	0.37	0.89	0.93	1.00	1.04	0.93	0.00	0.67	1.34	2.25	0.99	1.05	2.23	0.00	0.60	1.29	2.02	1.00	1.04	2.01	0.00
	1200	0.41	1.09	1.36	1.00	1.03	1.36	0.00	0.72	1.67	3.31	1.00	1.04	3.29	0.00	0.69	1.67	3.26	1.00	1.03	3.25	0.00
	1600	0.37	1.10	1.35	1.00	1.03	1.35	0.00	0.66	1.72	3.39	1.00	1.04	3.38	0.00	0.66	1.71	3.36	1.00	1.03	3.35	0.00
5	50	0.35	0.20	0.16	0.92	1.68	0.15	0.00	0.33	0.16	0.14	0.92	1.57	0.12	0.00	0.41	0.20	0.21	0.92	1.65	0.19	0.00
	100	0.37	0.15	0.16	0.91	1.77	0.15	0.00	0.35	0.15	0.14	0.90	1.86	0.13	0.00	0.40	0.17	0.19	0.91	1.79	0.17	0.00
	150	0.32	0.48	0.33	0.99	1.11	0.32	0.00	0.34	0.53	0.39	0.99	1.07	0.39	0.00	0.44	0.60	0.56	0.98	1.13	0.55	0.00
	200	0.34	0.61	0.49	0.99	1.06	0.49	0.00	0.36	0.75	0.69	0.99	1.04	0.69	0.00	0.48	0.82	0.91	0.99	1.08	0.90	0.00
	300	0.38	0.75	0.70	0.99	1.05	0.70	0.00	0.34	0.70	0.60	1.00	1.04	0.60	0.00	0.49	0.88	1.01	0.99	1.07	1.00	0.00
	400	0.36	0.77	0.71	1.00	1.04	0.71	0.00	0.49	1.03	1.29	1.00	1.03	1.29	0.00	0.56	1.16	1.65	0.99	1.05	1.64	0.00
	800	0.42	0.88	0.95	1.00	1.03	0.95	0.00	0.59	1.31	2.06	0.99	1.04	2.05	0.00	0.55	1.18	1.69	0.99	1.05	1.68	0.00
	1200	0.42	1.09	1.37	1.00	1.03	1.37	0.00	0.69	1.85	3.91	1.00	1.04	3.89	0.01	0.64	1.62	3.01	1.00	1.04	3.00	0.01
	1600	0.35	1.04	1.20	1.00	1.02	1.20	0.00	0.71	2.03	4.65	1.00	1.04	4.63	0.01	0.63	1.80	3.62	1.00	1.03	3.61	0.00

ANNEX E – Sample images, as labeled by Visual Language Models

Attentive samples



Figure 48 – absorbed

Votes for image under caption: absorbed (Figure 48)
Hard label: attentive, **Votes**: attentive=4, inattentive=0, **Model votes**: llava-forward: 1, llava-reverse: 1, qwenvl-forward: 1, qwenvl-reverse: 1
LLaVA forward (att→inat): The student in the foreground is attentive, as she is sitting at her desk with a notebook and a pencil, ready to take notes or complete her homework. This suggests that she is focused on her studies and actively engaged in the learning process. The presence of other students in the classroom also indicates that the environment is conducive to learning and concentration.
LLaVA reverse (inat→att): The student in the foreground is attentive, as they are actively writing in their notebook. This suggests that they are engaged in the learning process and focused on their work. The presence of a teacher and other students in the classroom also indicates that the student is participating in a structured learning environment, which can further contribute to their attentiveness.
Qwen-VL forward (att→inat): The student in the foreground appears to be attentive. This can be inferred from several observations:
1. **Posture**: The student is sitting upright, which is generally associated with attentiveness.
2. **Focus**: The student's gaze is directed downward towards the notebook, suggesting they are engaged in writing or reading, which indicates focus and attention.

3. **Hand Position**: The student's hand is holding a pen, poised as if ready to write, further indicating active participation.
These cues suggest that the student is actively involved in the task at hand, which aligns with being attentive.
Qwen-VL reverse (inat→att): The student in the foreground appears to be attentive. This can be inferred from several observations:
1. **Posture**: The student is sitting upright, which is generally associated with attentiveness.
2. **Focus**: The student's gaze is directed downward towards the notebook, suggesting they are engaged in writing or reading, which indicates focus and attentiveness.
3. **Hand Position**: The student's hand is holding a pen, poised over the notebook, further indicating that they are actively participating in an activity that requires attention.
These cues suggest that the student is likely paying close attention to the task at hand.



Figure 49 – actively listening

Votes for image under caption: actively listening (Figure 49)
Hard label: attentive, **Votes**: attentive=4, inattentive=0, **Model votes**: llava-forward: 1, llava-reverse: 1, qwenvl-forward: 1, qwenvl-reverse: 1
LLaVA forward (att→inat): The student in the foreground is attentive, as he is smiling and looking at the camera. This suggests that he is engaged and focused on the situation, possibly enjoying the moment or participating in a school event. The presence of other students in the background, along with the school setting, further supports the idea that the student is attentive and actively participating in the educational environment.
LLaVA reverse (inat→att): The student in the foreground is attentive, as he is smiling and looking at the camera. This suggests that he is engaged in the learning process and enjoying his time in the classroom. The presence of a book and a pencil on the desk also indicates that he is actively participating in the educational activities.
Qwen-VL forward (att→inat): The student in the foreground appears to be attentive. This can be inferred from several observations:

1. **Posture and Focus**: The student is sitting upright with their hands on the notebook, which suggests they are engaged and focused on the task at hand.
2. **Expression**: The student has a smile, which could indicate enjoyment or satisfaction with the activity, suggesting attentiveness.
3. **Environment**: The classroom setting, with other students also appearing to be engaged, supports the idea that the environment is conducive to learning.
These cues collectively suggest that the student is attentive rather than inattentive.
Qwen-VL reverse (inat→att): The student in the foreground appears to be attentive. This can be inferred from several observations:
1. **Posture and Focus**: The student is sitting upright with their hands on the notebook, which suggests they are engaged and focused on the task at hand.
2. **Expression**: The student has a smile, which could indicate enjoyment or satisfaction with the activity, suggesting attentiveness.
3. **Environment**: The classroom setting is conducive to learning, with other students also appearing to be engaged in their work.
These cues collectively suggest that the student is attentive rather than inattentive.



Figure 50 – attentive

Votes for image under caption: attentive (Figure 50)
Hard label: attentive, **Votes**: attentive=4, inattentive=0, **Model votes**: llava-forward: 1, llava-reverse: 1, qwenvl-forward: 1, qwenvl-reverse: 1
LLaVA forward (att→inat): The student in the foreground appears to be attentive, as they are looking at the paper while holding a pencil. This suggests that they are focused on their work and actively engaged in the task at hand. The presence of other students in the background might also indicate that the student is paying attention to their surroundings and the activities of their peers, which could be a part of their learning experience.
LLaVA reverse (inat→att): The student in the foreground appears to be attentive, as they are looking at the paper while holding a pencil. This suggests that they are focused on their work and actively engaged in the task at

hand. The presence of other students in the background might also indicate that the student is part of a group or class, which could further emphasize the importance of their attention and engagement in the learning process.
Qwen-VL forward (att→inat): The student in the foreground appears to be attentive. This can be inferred from their posture and facial expression, which suggest focus and engagement. They are holding a pen, indicating they might be taking notes or working on an assignment, and their gaze is directed forward, likely towards the teacher or the front of the classroom where instruction is taking place. The overall demeanor suggests concentration rather than inattention.
Qwen-VL reverse (inat→att): The student in the foreground appears to be attentive. This can be inferred from several observations:
1. **Posture and Focus**: The student is seated upright, which is generally associated with attentiveness.
2. **Hand Position**: The student's hand is holding a pen, suggesting they are engaged in an activity that requires focus, such as writing or taking notes.
3. **Expression**: Although it's difficult to discern emotions clearly in black and white, the student's expression does not appear distracted or disengaged.
These cues suggest that the student is likely paying attention to something happening in the classroom, possibly listening to a teacher or participating in an activity.



Figure 51 – concentrated

Votes for image under caption: concentrated (Figure 51)
Hard label: attentive, **Votes**: attentive=4, inattentive=0, **Model votes**: llava-forward: 1, llava-reverse: 1, qwenvl-forward: 1, qwenvl-reverse: 1
LLaVA forward (att→inat): The student in the foreground appears to be attentive, as they are focused on writing in their notebook. The image shows a young girl sitting at a desk, writing with a pencil. The fact that she is writing in a notebook suggests that she is engaged in the task and paying attention to her work. This is in contrast to the other students in the background who are not actively engaged in writing or studying.
LLaVA reverse (inat→att): The student in the foreground appears to be attentive, as they are focused on writing in their notebook. The image shows a young girl sitting at a desk, writing with a pencil. The fact that she

is writing in a notebook suggests that she is engaged in the task and paying attention to her work. The presence of other students in the background also indicates that the girl is not the only one in the classroom, and she is likely following the same routine as the others.
Qwen-VL forward (att→inat): The student in the foreground appears to be attentive. This can be inferred from several observations:
1. **Posture and Focus**: The student is sitting upright, which is generally associated with attentiveness.
2. **Hand Position**: The student's hand is holding a pencil, poised over an open book, suggesting they are actively engaged in writing or taking notes.
3. **Eye Contact**: Although it's difficult to see the eyes clearly due to the angle, the overall posture suggests that the student is looking at the book, indicating focus.
These cues suggest that the student is likely paying attention to their work during class.
Qwen-VL reverse (inat→att): The student in the foreground appears to be attentive. This can be inferred from several observations:
1. **Posture and Focus**: The student is seated upright, which is generally associated with attentiveness.
2. **Hand Position**: The student's hand is holding a pencil, poised over an open book, suggesting they are actively engaged in writing or taking notes.
3. **Eye Contact**: Although it's difficult to see the eyes clearly due to the angle, the overall posture suggests that the student is looking at the book, indicating focus on the task.
These cues collectively suggest that the student is attentive and focused on their work.



Figure 52 – curious

Votes for image under caption: curious (Figure 52)
Hard label: attentive, **Votes**: attentive=4, inattentive=0, **Model votes**: llava-forward: 1, llava-reverse: 1, qwenvl-forward: 1, qwenvl-reverse: 1
LLaVA forward (att→inat): The student in the foreground appears to be attentive, as they are sitting at a desk with a book and a pencil, which suggests they are actively engaged in learning or studying. The presence of a backpack also indicates that the student is prepared for a day of school or academic activities. The fact that the

student is wearing glasses might also suggest that they are focused on their work or studies, as they need to see clearly to read or write effectively.
LLaVA reverse (inat→att): The student in the foreground appears to be attentive, as they are sitting at a desk with a book and a pencil, which suggests they are actively engaged in learning or studying. The presence of a backpack also indicates that the student is prepared for school or an educational activity. The fact that the student is wearing glasses might also suggest that they are focused on their work or the task at hand.
Qwen-VL forward (att→inat): The student in the foreground appears to be attentive. This can be inferred from their posture and gaze direction. They are sitting upright, which is often associated with engagement and focus. Additionally, they seem to be looking forward, possibly at the teacher or the front of the classroom, which suggests they are paying attention to what is happening in class. The presence of a pencil in hand also indicates that they might be ready to take notes or participate in an activity, further supporting the idea that they are attentive.
Qwen-VL reverse (inat→att): The student in the foreground appears to be attentive. This can be inferred from several observations:
1. **Posture and Focus**: The student is seated upright, which is generally associated with attentiveness.
2. **Hand Position**: The student's hand is resting on the desk, suggesting they might be taking notes or preparing to write, which indicates engagement.
3. **Eye Contact**: Although it's difficult to determine eye contact precisely, the student's gaze seems directed forward, possibly towards the teacher or the front of the classroom, which is typical behavior for students who are paying attention.
These cues suggest that the student is likely focused on the lesson or activity at hand.

Inattentive samples



Figure 53 – checking a phone

Votes for image under caption: checking a phone (Figure 53)
Hard label: inattentive, **Votes**: attentive=1, inattentive=-3, **Model votes**: llava-forward: -1, llava-reverse: 1, qwenvl-forward: -1, qwenvl-reverse: -1
LLaVA forward (att→inat): The student in the foreground appears to be inattentive, as they are looking at their cell phone instead of focusing on the class. This could be due to various reasons, such as feeling bored, disengaged, or distracted by the content on their phone. It is important for students to maintain focus and engage in the classroom environment to fully benefit from their education and develop essential skills.
LLaVA reverse (inat→att): The student in the foreground is attentive, as they are looking at their cell phone. This suggests that they might be engaged in an activity on the phone, such as texting, browsing the internet, or using an app. The presence of multiple students in the classroom indicates that they are all focused on their own tasks, and the student using the cell phone is not necessarily inattentive to the class or the teacher's instructions.
Qwen-VL forward (att→inat): The student in the foreground appears to be inattentive. This is suggested by their focused gaze on the device they are holding, which seems to be a smartphone. In a classroom setting, it is generally expected that students would be paying attention to the teacher or the lesson at hand rather than engaging with personal devices. The posture and focus of the student suggest they are not actively participating in the class activity, which could indicate a lack of attentiveness.
Qwen-VL reverse (inat→att): The student in the foreground appears to be inattentive. This is suggested by their focused gaze on the device they are holding, which could indicate that they are engaged with something other than the lesson being taught. In an educational setting, it is generally expected for students to be attentive and focused on the teacher or the material presented during class time. The fact that this student is looking at a device rather than participating in the lesson suggests a lack of attentiveness.



Figure 54 – dozing

Votes for image under caption: dozing (Figure 54)
Hard label: inattentive, **Votes**: attentive=1, inattentive=-3, **Model votes**: llava-forward: -1, llava-reverse: -1, qwenvl-forward: -1, qwenvl-reverse: 1

LLaVA forward (att→inat): The student in the foreground appears to be inattentive, as they are resting their head on their hand while sitting at the desk. This suggests that the student might be tired, bored, or not fully engaged in the class or activity. It is possible that the student is not interested in the subject matter or is struggling to stay focused, leading to a lack of attention and engagement in the class.
LLaVA reverse (inat→att): The student in the foreground appears to be inattentive, as they are resting their head on their hand while sitting at the desk. This suggests that the student might be tired, bored, or not fully engaged in the class or activity. It is possible that the student is not paying full attention to the lesson or the teacher, or they might be daydreaming or simply taking a break from the task at hand.
Qwen-VL forward (att→inat): The student in the foreground appears to be inattentive. This can be inferred from several observations:
1. **Posture**: The student has their head resting on one hand, which is a common posture associated with boredom, tiredness, or disengagement.
2. **Eye Contact**: The student's eyes do not appear to be focused on the book or any other activity, suggesting they may not be actively engaged.
3. **Body Language**: The overall body language does not convey active participation or concentration.
These cues suggest that the student might be daydreaming, feeling tired, or simply not interested in the current task at hand. However, it's important to note that this interpretation is based on a single image and could be influenced by various factors not visible in the photo.
Qwen-VL reverse (inat→att): The student in the foreground appears to be attentive. This can be inferred from several observations:
1. **Posture**: The student's posture is upright, which is generally associated with attentiveness.
2. **Hand Position**: The student has one hand resting on their cheek and the other on the desk, which could indicate deep thought or concentration rather than inattention.
3. **Focus**: The student's gaze is directed downward towards the book, suggesting they are engaged with the material.
While it's not possible to determine the exact level of engagement without more context, these visual cues suggest that the student is likely attentive to the task at hand.

Votes for image under caption: hunched over (Figure 55)
Hard label: inattentive, **Votes**: attentive=0, inattentive=-4, **Model votes**: llava-forward: -1, llava-reverse: -1, qwenvl-forward: -1, qwenvl-reverse: -1
LLaVA forward (att→inat): The student in the foreground is inattentive, as they are sleeping at their desk. This suggests that the student might be tired or not engaged in the ongoing class or activity. It is possible that the student has been sitting for a long time or has not had enough rest, leading to their inattentiveness. The presence of other students in the classroom also indicates that the student is not the only one who might be feeling tired or disengaged.
LLaVA reverse (inat→att): The student in the foreground appears to be inattentive, as they are sleeping at their desk. This suggests that the student might be tired or not engaged in the ongoing class or activity. It is possible that the student has been sitting for a long time or has not had enough rest, leading to their inattentiveness. The presence of other students in the classroom also indicates that the student is not the only one who might be feeling tired or disengaged.
Qwen-VL forward (att→inat): The student in the foreground appears to be inattentive. This can be inferred from their posture, with their head resting on their arms and their eyes closed, which often indicates a state of rest, boredom, or disengagement. In contrast, attentive students typically sit upright, maintain eye contact, and show signs of engagement such as listening or working on tasks.
Qwen-VL reverse (inat→att): The student in the foreground appears to be inattentive. This is suggested by their posture, with their head resting on their arms and their eyes closed, which typically indicates a lack of engagement or focus. In contrast, attentive students usually maintain eye contact with the teacher or the material being presented, sit upright, and show signs of concentration such as nodding or taking notes.

Votes for image under caption: with silent attention (Figure 56)
Hard label: inattentive, **Votes**: attentive=1, inattentive=-3, **Model votes**: llava-forward: -1, llava-reverse: 1, qwenvl-forward: -1, qwenvl-reverse: -1



Figure 55 – hunched over

LLaVA forward (att→inat): The student in the foreground appears to be inattentive, as she is looking down and not paying attention to the teacher or the class. This could be due to various reasons, such as feeling unwell, being distracted by something outside the classroom, or simply not engaged in the lesson being taught. It is important for the teacher to identify the cause of the student's inattentiveness and address it to ensure that the student remains engaged and focused in the classroom.
LLaVA reverse (inat→att): The student in the foreground appears to be attentive, as she is sitting in a classroom with her eyes focused on the camera. This suggests that she is actively engaged in the situation and paying attention to the person or object capturing her attention. The fact that she is sitting in a classroom with other students also implies that she is likely participating in a learning activity or listening to a teacher, which further supports the idea that she is attentive and engaged in the classroom environment.
Qwen-VL forward (att→inat): The student in the foreground appears to be inattentive. This can be inferred from their posture and facial expression. They have their arms crossed, which is often a sign of disengagement or frustration. Additionally, their gaze seems directed downward, away from the teacher or the activity at hand, which could indicate that they are not focused on what is happening in the classroom. However, it's important to note that body language can sometimes be misleading, and without more context, it's difficult to determine the exact reason for their behavior.
Qwen-VL reverse (inat→att): The student in the foreground appears to be inattentive. This can be inferred from several cues:
1. **Posture**: The student's arms are crossed, which is often a sign of disengagement or frustration.
2. **Expression**: The student's facial expression seems to convey a sense of boredom or lack of interest.
3. **Body Language**: The student is leaning forward with their head down, which might indicate a lack of engagement with the activity at hand.
These non-verbal cues suggest that the student may not be fully focused on the lesson or task being presented. However, it's important to note that this interpretation is based on visual cues and may not accurately reflect the student's internal state or level of understanding.



Figure 56 – with silent attention

Votes for image under caption: immersed in the task (Figure 57)
Hard label: inattentive, **Votes**: attentive=1, inattentive=-3, **Model votes**: llava-forward: -1, llava-reverse: 1, qwenvl-forward: -1, qwenvl-reverse: -1
LLaVA forward (att→inat): The student in the foreground appears to be inattentive, as she is covering her ears with her hands. This behavior suggests that she might be experiencing discomfort or distress, possibly due to noise or other environmental factors. It is important for the student to address the cause of her discomfort to ensure that she remains focused and engaged in her learning activities.
LLaVA reverse (inat→att): The student in the foreground appears to be attentive, as she is focused on her work and has her hands on her ears. This suggests that she is concentrating on her studies or listening to something, possibly a teacher or classmate, and is trying to block out any distractions or noise. The fact that she is wearing headphones further supports the idea that she is engaged in an activity that requires focus and concentration.
Qwen-VL forward (att→inat): The student in the foreground appears to be inattentive. This can be inferred from several visual cues:
1. **Posture and Hand Position**: The student has their hands on their head, which is often a sign of frustration, boredom, or disengagement.
2. **Facial Expression**: Although the face is not fully visible, the overall posture suggests a lack of focus or engagement with the task at hand.
3. **Body Language**: The student's body is leaning forward, but there is no indication of active participation or concentration.
These signs suggest that the student might be feeling overwhelmed, tired, or distracted, rather than being attentive and focused on the work.
Qwen-VL reverse (inat→att): The student in the foreground appears to be inattentive. This can be inferred from several cues:
1. **Posture and Hand Position**: The student has their hands on their head, which is often a sign of frustration, boredom, or disengagement.
2. **Facial Expression**: While it's not possible to see the full facial expression clearly, the overall posture suggests a lack of engagement with the task at hand.



Figure 57 – immersed in the task

3. **Focus on the Task**: The student seems to be looking downward rather than at the teacher or the board, which might indicate they are not paying attention to the lesson.
These observations suggest that the student may be feeling overwhelmed, bored, or distracted, leading to a lack of attentiveness.