Zero-Shot Vision-Language Models for Classroom Engagement Recognition

A Benchmark Study of Prompt Sensitivity & Cross-Dataset Generalization

Kshama Nitin Shah[1*] **Aman Goyal**[2*] **Kemmannu Vineet Venkatesh Rao**[3]

[1] University of Michigan, Ann Arbor [2] Carnegie Mellon University [3] Magna International [*] *equal contribution*

CVPR JUNE 3-7, 2026 DENVER COLORADO

CV4Edu Workshop · non-archival

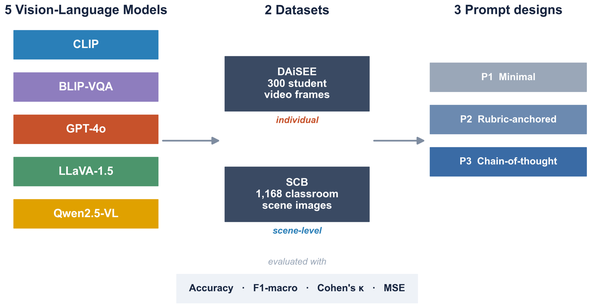
Motivation

Engagement predicts learning outcomes, but measuring it at scale needs trained human observers or large labelled datasets — both hard to deploy.

Vision-Language Models promise *plug-and-play, zero-shot* classroom observation from natural-language prompts — no fine-tuning, no task-specific data.

Our question: do off-the-shelf VLMs actually recognise engagement? We run the first systematic zero-shot benchmark and find a sharp split between individual- and scene-level perception.

The Benchmark



5 VLMs × 2 datasets × 3 prompts, scored on 4 metrics (Acc, F1, Cohen's κ, MSE). ***SCB is reframed as holistic scene-level classification — not per-student detection.***

Contributions

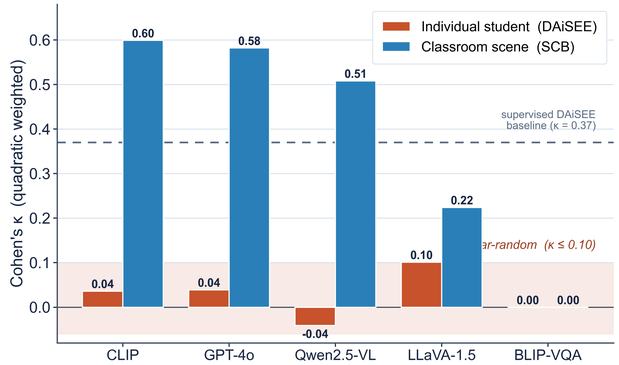
- **First zero-shot VLM benchmark** for classroom engagement across two complementary datasets.
- **Three failure modes** on individuals: near-random agreement, class collapse, extreme prompt sensitivity.
- **Scene-level is tractable** — CLIP & GPT-4o reach κ ≈ 0.60, near supervised methods.
- **A deployment barrier**: GPT-4o safety refusals on faces + low run-to-run self-consistency.
- **A mechanistic explanation**: feature-space analysis showing CLIP clusters frames by student identity, not engagement.

VLMs read classrooms, not students

Zero-shot VLMs are near-random on **individual students** (κ ≤ 0.10), but reach moderate, supervised-comparable agreement on **whole-classroom scenes** (κ ≈ 0.60).

→ **Don't observe the face — aggregate spatially and query the scene.**

Key Result



Best Cohen's κ per model. Scene-level (blue) towers over individual-level (orange).

DAiSEE — individual

Model	Acc	κ	MSE
Supervised*	62.3	0.37	—
CLIP	37.0	0.04	1.00
GPT-4o	36.0	0.04	0.69
LLaVA-1.5	39.0	0.10	0.72
Qwen2.5-VL	35.3	-0.04	0.74
BLIP-VQA	35.3	0.00	0.69

SCB — scene

Model	Acc	κ	MSE
CLIP	67.3	0.60	0.59
GPT-4o	67.9	0.58	0.37
Qwen2.5-VL	64.9	0.51	0.41
LLaVA-1.5	49.2	0.22	2.67
BLIP-VQA	33.6	0.00	0.72

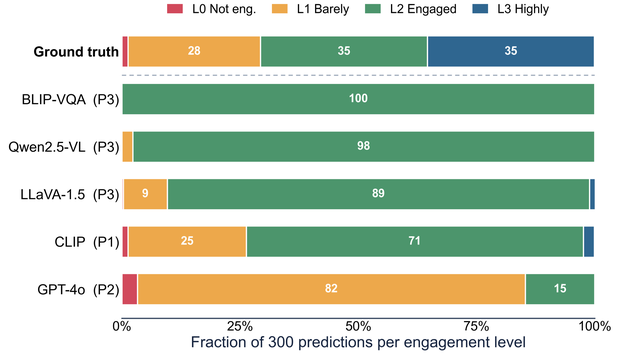
** published supervised DAiSEE baseline (Huang et al.); both tables show best prompt per model.*

Why the gap?

- **Scene cues are language-groundable.** Spatial layout, group dynamics and overt actions (raised hands, phones) match what internet-scale image–text training has seen.
- **A single face is not.** Engagement is a subtle cognitive state with few visual anchors — and it lives in a part of feature space VLMs never learned to separate (see 'Why It Fails').

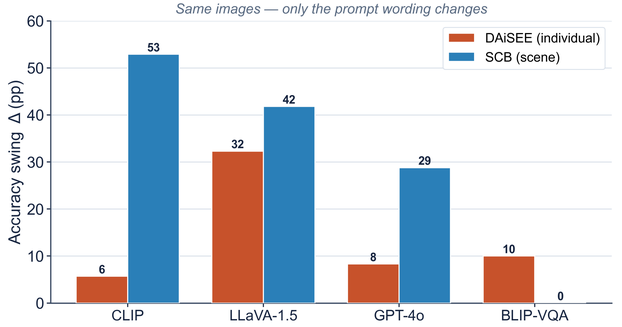
6× higher κ at the scene level than the best individual-level model.

Failure Mode 1 — Class Collapse



Models dump 85–100% of predictions into one level — apparent accuracy just reflects the majority class, not real 4-way discrimination.

Failure Mode 2 — Prompt Sensitivity



Up to 53 pp accuracy swing from prompt wording alone, on identical images.

Failure Mode 3 — Unreliable Deployment

98%

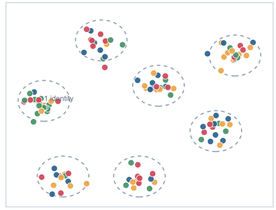
of GPT-4o chain-of-thought requests on student faces are blocked by safety filters

27–32%

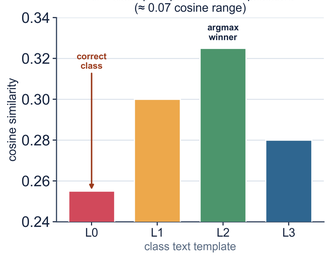
of predictions flip across identical repeated runs — a single reported κ is unreliable

Why It Fails

CLIP groups frames by student identity, not by engagement



All 4 class prompts are near-equidistant (≈ 0.07 cosine range)



CLIP organises its feature space around ***who the student is*** — identity, lighting, background — not how engaged they are. All four engagement classes intermix within every identity cluster, and the four class prompts sit ≈0.07 apart in cosine space, so the predicted label is essentially arbitrary.

Takeaways

- **Don't deploy VLMs as individual observers** — zero-shot face-level engagement is not yet reliable.
- **Aggregate spatially**: query the scene, not the face. More accurate *and* more privacy-respecting for minors.
- **Report prompt-sensitivity (Δκ)** as a first-class evaluation axis.
- **Open 7B models match GPT-4o** at zero API cost, keeping student data on-premise.

Conclusion

Current VLMs can't read *individual* engagement, but scene-level "read-the-room" observation is already viable. Progress needs scene-first pre-aggregation, prompt-robust tuning, and temporal modelling beyond single frames.

Selected References

[1] Gupta et al. DAiSEE: Towards user engagement recognition in the wild. arXiv 2016.
[2] Wang et al. Student classroom behaviour detection (SCB). Systems & Soft Computing 2023.
[3] Radford et al. Learning transferable visual models (CLIP). ICML 2021.
[4] Li et al. BLIP: Bootstrapping language-image pre-training. ICML 2022.
[5] OpenAI. GPT-4o system card. Technical report 2024.
[6] Liu et al. Visual instruction tuning (LLaVA). NeurIPS 2023.
[7] Bai et al. Qwen2.5-VL technical report. arXiv 2025.
[8] Wei et al. Chain-of-thought prompting elicits reasoning in LLMs. NeurIPS 2022.
[9] Whitehill et al. The faces of engagement. IEEE Trans. Affective Computing 2014.
[10] Zhao et al. Calibrate before use: few-shot performance of language models. ICML 2021.
[11] Lu et al. Fantastically ordered prompts and where to find them. ACL 2022.
[12] Menon & Vondrick. Visual classification via description from LLMs. ICLR 2023.
[13] Abedi & Khan. Improving engagement detection with a ResNet+TCN hybrid. CRV 2021.

Contact: **Kshama Nitin Shah** · kshama2705@gmail.com