MAVE: A Multi-Agent Framework for Grounded Pedagogical Feedback from Online Class Videos

Anonymous ACL submission

Abstract

Timely pedagogical feedback is critical for effective online instruction, yet existing AI-based educational assessment systems still struggle to generate actionable and trustworthy recommendations from authentic classroom environments. Major challenges include long-context noisy videos and weak grounding in real student behaviors. To address these challenges, we introduce **MAVE** (**M**ultimodal **A**gents for **V**ideo-based **E**ducational Feedback), a collaborative multi-agent framework for generating pedagogically grounded feedback from online classroom videos. Built upon the ICAP framework, MAVE formulates feedback generation as a discrepancy reasoning problem between instructors' intended engagement objectives and students' observed cognitive states. MAVE decomposes classroom understanding into specialized agents for instructional structuring, multimodal evidence extraction, engagement assessment, and pedagogical recommendation generation. By grounding feedback in temporally aligned multimodal classroom evidence, MAVE produces interpretable and actionable instructional recommendations. Experiments on real-world online classroom videos demonstrate that MAVE generates more pedagogically grounded and actionable feedback than existing LLM-based feedback generation baselines. Our code is publicly available at [1].

1 Introduction

Online instruction increasingly relies on recorded and screen-sharing-heavy classroom environments, where instructors often struggle to monitor how students respond to instructional activities in real time (Bahmanyar et al., 2015; Rasiah et al., 2020). In such settings, students frequently exhibit passive or minimally observable participation during instructional activities, as shown in Fig. 1. Although expert supervisors can convert classroom observations into actionable pedagogical feedback (Fisher et al., 2025), such observation does not scale across large or asynchronous courses.

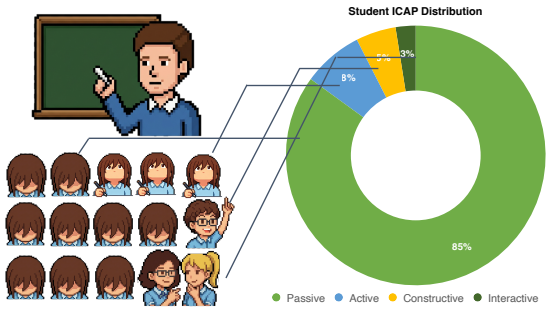

Figure 1: A representative online classroom scenario illustrating the imbalance of student cognitive engagement under the ICAP framework. Most students remain in passive learning states, while only a small proportion exhibit active, constructive, or interactive behaviors, making real-time instructional assessment and feedback difficult and necessary for instructors.

As a result, instructors frequently receive little interpretable evidence about when student engagement breaks down, why specific instructional activities fail, or how interaction patterns could be improved (Gorbunova et al., 2026; Yao et al., 2026). Without such evidence, instructional refinement remains anecdotal: adjustments cannot be tied to specific student moments and cannot be replicated across cohorts. Useful automation must therefore go beyond a single global engagement score; it should localize behaviors within long videos and tie each recommendation to the student evidence behind it.

Meeting this requirement is hard for two reasons. First, classroom videos are inherently *long and noisy*: extended exposition, off-task movement, filler speech, and fragmented participation make engagement cues sparse and segment-local (Nagrani et al., 2025). Second, student engagement

[1] https://anonymous.4open.science/status/MAVE-3D36

is *multimodal and temporally distributed*: note-taking, delayed verbal responses, gaze shifts, and discussion occur asynchronously across modalities. Faced with such input, existing engagement-analysis systems collapse each segment to a single global label or affect category, discarding the cues that explain why a moment succeeded or failed. Recent LLM-based feedback approaches, in turn, rely on unconstrained self-reflection (Shinn et al., 2023; Yuksekgonul et al., 2024) or rubric scoring (Chen et al., 2024; Hong et al., 2026), but do not compare what the lesson aimed to elicit with what students actually showed. Current systems therefore cannot tell whether low engagement reflects expected passive learning, missing instructional scaffolding, or a genuine mismatch between intent and response.

We argue that the missing capability is not better engagement prediction but an explicit per-segment comparison between *intended* and *observed* cognitive engagement. The ICAP framework (Chi and Wylie, 2014) supplies the vocabulary for this comparison: it organizes learning behaviors into four ordinal modes: Passive, Active, Constructive, Interactive, so that an instructional segment can be labeled with an intended cognitive demand and the students' response in the same segment can be assessed on the same scale. The gap between the two becomes directly comparable, and pedagogical feedback can be expressed as a function of that gap rather than as a generic improvement suggestion.

To address the aforementioned challenges, we introduce **MAVE** (**M**ultimodal **A**gents for **V**ideo-based **E**ducational Feedback), a collaborative multi-agent framework for generating actionable pedagogical feedback from authentic online classroom videos. Rather than treating engagement analysis as a standalone classification problem, MAVE formulates instructional feedback generation as a discrepancy reasoning process between the instructor's intended ICAP objective and students' observed cognitive engagement states. Specifically, MAVE decomposes classroom understanding into four specialized agents: **Curator**, **Observer**, **Assessor**, and **Advisor**. The Curator segments long-context lectures into temporally aligned instructional units and infers the target ICAP level for each segment. The Observer extracts multimodal student evidence from classroom video, audio, and interaction signals. Based on these observations, the Assessor estimates students' actual ICAP engagement trajectories through calibrated pedagogical reasoning. Finally, the Advisor compares the intended and observed ICAP states to generate time-aligned, evidence-grounded instructional recommendations. By grounding feedback generation in real classroom evidence and pedagogical theory, MAVE bridges the gap between classroom perception and actionable instructional intervention.

Our Core Contributions (1) We formulate pedagogical feedback generation as a discrepancy reasoning problem between instructors' intended ICAP objectives and students' observed cognitive engagement states, moving beyond conventional engagement classification toward actionable instructional intervention. (2) We propose **MAVE**, a collaborative multi-agent framework that integrates multimodal classroom perception, temporal reasoning, and ICAP-grounded pedagogical analysis. MAVE decomposes classroom understanding into specialized agents for instructional structuring, student evidence extraction, engagement assessment, and feedback generation. (3) We evaluate MAVE on real-world online classroom videos. Experimental results demonstrate that MAVE produces more pedagogically grounded and actionable feedback compared with existing LLM-based feedback generation baselines.

2 Related Work

2.1 AI in Educational Analytics and Feedback

Automated classroom analysis via computer vision, audio processing and multimodal sensing has been widely explored. Early works focused on face detection, gaze tracking and posture recognition to infer student behaviors and attention patterns (Whitehill et al., 2014), while multimodal learning analytics platforms further integrated physiological and environmental signals for holistic classroom monitoring (Blikstein, 2013). Recent advances have adopted deep learning to extract fine-grained behavioral and affective cues from classroom data (Liu et al., 2025), but existing pipelines remain limited to feature extraction and behavioral labeling, failing to construct structured and pedagogically actionable feedback signals for the optimization of instructional strategies. Student engagement modeling is mainly formulated as multi-class classification or regression tasks (Fredricks et al., 2004), with vision-based methods that rely on facial and gaze features to optimize flat categorical objectives (Whitehill et al., 2014; Monkaresi et al., 2016). While ordinal prediction has been applied in sentiment estimation and quality assessment domains

(Gutiérrez et al., 2015), few engagement prediction works encode ordinal structure into uncertainty modeling. Meanwhile, feedback-driven optimization paradigms such as reinforcement learning from human feedback (RLHF), preference-based model optimization (Rafailov et al., 2023), and adaptive interactive learning systems (Okubo et al., 2022; Yan et al., 2024) predominantly rely on explicit preference annotations or scalar reward signals, rarely exploring the extraction of fine-grained implicit feedback from multimodal classroom environments. In practical educational systems, prior works including EmotionCues (Zeng et al., 2020), Zheng et al. (Zheng et al., 2020), and Gao et al. (Gao et al., 2025) mainly focus on behavior recognition, feature extraction, or domain-specific action evaluation, without linking authentic fine-grained feedback to actionable instructional optimization or establishing a real-scenario feedback utilization paradigm.

2.2 LLMs as Evaluators

Large language models have been widely used as evaluators in open-ended tasks such as dialogue assessment and reasoning verification (Zheng et al., 2023; Wang et al., 2024). Multi-agent frameworks further improve evaluation reliability via iterative reasoning, role decomposition, and cross-verification mechanisms (Du et al., 2024; Liang et al., 2024). Recent studies extend this evaluation paradigm to educational scenarios (Emirtekin, 2025; Yi et al., 2025), demonstrating the potential of LLM-based holistic judgment. However, existing LLM-as-a-judge systems predominantly operate in textual domains and often rely on unconstrained generative judgments without structural constraints, uncertainty calibration, or safety gating. When extended to multimodal educational settings, such paradigms (Yuksekgonul et al., 2024; Shinn et al., 2023) may introduce hallucination, bias, and pedagogical inconsistency risks.

3 Method

Overview. Figure 2 summarizes the pipeline. We formulate automated teaching feedback as a discrepancy reasoning problem between the instructor's intended ICAP objective and the students' observed ICAP state. Given a raw online classroom video $\mathcal{V}$, MAVE decomposes the problem into four coordinated agents: **Curator**, **Observer**, **Assessor**, and **Advisor**. The Curator organizes the lecture into temporally aligned instructional units and infers the pedagogical target for each unit. The Observer extracts multimodal evidence of student behavior from the corresponding classroom segment. The Assessor maps this evidence to an observed ICAP state through calibrated ordinal reasoning. The Advisor compares the intended and observed states and generates actionable feedback grounded in both lesson content and student response.

Division of labor. The **Assessor** is responsible only for *measuring* student cognitive engagement: it estimates $\hat{y}_k$ from multimodal evidence and applies ordinal-consistent calibration (Eqs. (2)–(4)) to reduce noisy LLM judgments. It does not generate instructional recommendations. The **Advisor** is responsible only for *acting on* the pedagogical gap: given c_k from the Curator and $\hat{y}_k$ from the Assessor, it computes the discrepancy Δ_k (Eq. (5)) and synthesizes feedback $\mathcal{F}_k$ (Eq. (6)). Calibration stabilizes engagement estimation; discrepancy reasoning turns that estimate into targeted instructional advice.

Formally, let the classroom video be partitioned into K instructional segments $\{\mathcal{V}_k\}_{k=1}^{K}$. For each segment k, MAVE estimates (1) an intended ICAP target $c_k \in \mathcal{Y}$, where $\mathcal{Y} = \{P, A, C, I\}$, (2) an observed student ICAP state $\hat{y}_k \in \mathcal{Y}$, and (3) a feedback text $\mathcal{F}_k$ that explains the discrepancy between c_k and $\hat{y}_k$.

3.1 Curator: Pedagogical Structuring

The Curator reconstructs the pedagogical structure of the lecture from the raw classroom stream, which aims to infer the cognitively appropriate ICAP objective. In other words, the agent estimates what level of student cognitive engagement a particular knowledge unit should ideally elicit according to pedagogical principles.

Instructional parsing. As an auxiliary tool, the Curator first invokes a Transformer-based automatic speech recognition (ASR) module to obtain a timestamp-aligned transcript $\mathcal{T} = \{(w_n, t_n)\}_{n=1}^{N}$, where w_n denotes the recognized token and t_n denotes its temporal boundary. The transcript is aligned with the classroom timeline and grouped into contiguous instructional windows.

Instructional unit construction. Based on the transcript and temporal structure, the Curator partitions the lecture into pedagogically coherent instructional units $\mathcal{U}_k = (\mathcal{W}_k, \tau_k)$, where $\mathcal{W}_k$ de-

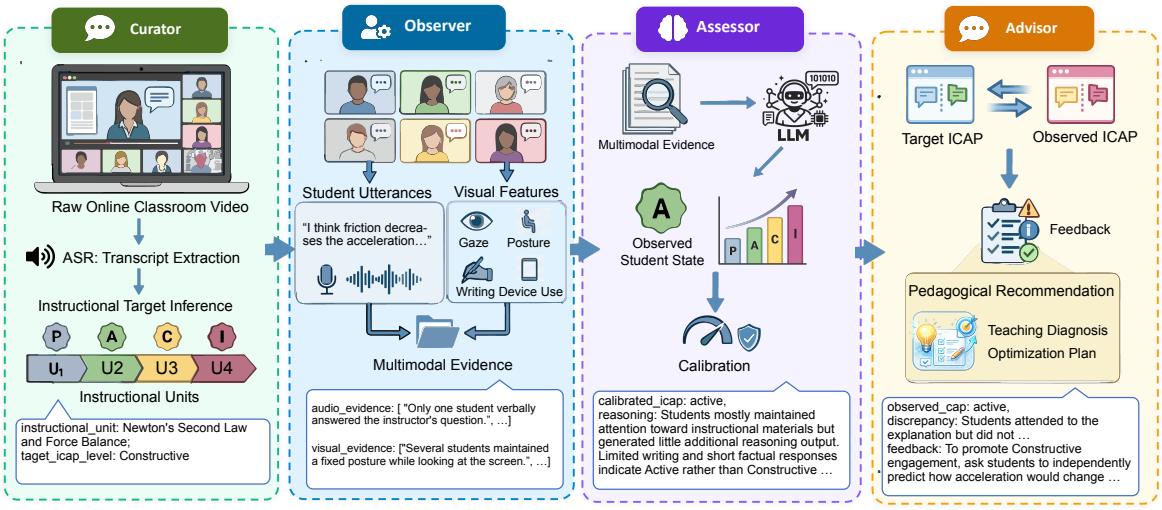

Figure 2: Overview of MAVE. Given a raw online classroom video, the Curator segments the lecture and infers the intended ICAP objective for each instructional unit. The Observer extracts multimodal student evidence from the corresponding temporal window. The Assessor estimates the observed student ICAP state through calibrated ordinal reasoning. Finally, the Advisor compares the intended and observed ICAP states and generates time-aligned pedagogical feedback grounded in classroom evidence.

notes the instructional content within the segment and $\tau_k = [t_k^s, t_k^e]$ denotes its temporal interval. Each instructional unit corresponds to a relatively self-contained knowledge delivery episode. This segmentation establishes a stable temporal abstraction for downstream multimodal reasoning and enables direct alignment between instructional context and student responses.

ICAP target inference. For each instructional unit $\mathcal{U}_k$, the Curator analyzes the instructional content, discourse structure, and pedagogical interaction patterns to infer the cognitively appropriate ICAP target $c_k \in \{P, A, C, I\}$. Instead of predicting what the teacher subjectively intended, the Curator estimates what form of cognitive engagement the knowledge component should ideally promote. For example, factual exposition may correspond to Passive or Active engagement, while conceptual synthesis, reasoning prompts, or collaborative problem solving are expected to induce Constructive or Interactive engagement. The inferred ICAP target serves as the pedagogical reference state for the subsequent discrepancy analysis.

3.2 Observer: Learner Feedback Modeling

The Observer extracts structured student-side evidence from multimodal classroom signals. Instead of directly predicting engagement labels from raw inputs, the Observer first converts heterogeneous classroom observations into temporally aligned natural-language evidence that can support downstream pedagogical reasoning.

Audio interaction parsing. For each instructional unit $\mathcal{U}_k$, the Observer invokes an automatic speech recognition (ASR) module to obtain timestamp-aligned classroom transcripts. Based on speaker turns and interaction context, the Observer further separates student utterances from the instructor's speech, yielding a student-side interaction transcript $\mathcal{A}_k = \{(s_i, t_i)\}$, where s_i denotes a student utterance and t_i denotes its temporal location. These parsed responses provide linguistic evidence about participation, reasoning behaviors, and collaborative interaction patterns.

Visual evidence extraction. For the visual stream, the Observer invokes a video parsing tool to decompose the online classroom layout into individual student sub-windows. Frames are then sampled from each student region at fixed temporal intervals within τ_k. The sampled frames are converted into factual natural-language descriptions that summarize observable behaviors such as gaze direction, posture, writing activity, device usage, and interaction with learning materials.

Formally, the extracted visual evidence for the k-th instructional unit is represented as $\mathcal{V}_k = \{v_k^1, v_k^2, \ldots, v_k^M\}$, where each v_k^m denotes the textual behavioral description of the m-th student window.

The Observer aggregates the linguistic and visual evidence into a temporally aligned multimodal evidence summary $\mathcal{E}_k = \{\mathcal{A}_k, \mathcal{V}_k\}$. This representation preserves local classroom behaviors while reducing frame-level noise, thereby providing a stable foundation for downstream ICAP assessment and discrepancy reasoning.

3.3 Assessor: ICAP State Estimation

The Assessor maps observed learner evidence to a calibrated ICAP state $\hat{y}_k$. Its outputs are consumed by the Advisor and are not used to generate recommendations directly.

ICAP state reasoning. Given the evidence summary $\mathcal{E}_k$, the Assessor first produces an initial ICAP anchor via an LLM-based Judge Agent:

$$\hat{y}_k^{(0)} = \arg\max_{y \in \mathcal{Y}} P_{\text{LLM}}(y \mid \mathcal{E}_k), \tag{1}$$

where $\mathcal{Y} = \{P, A, C, I\}$. The Judge Agent also provides textual justification for the predicted anchor, explaining which behavioral cues support the decision.

Ordinal-consistent calibration. Because LLM judgments may be locally noisy, we apply an ordinal-consistent confidence reconstruction step to convert the anchor prediction into a smoother and more robust distribution over ICAP levels. Let $d(y, \hat{y}_k^{(0)})$ denote the semantic distance between a candidate level y and the anchor level $\hat{y}_k^{(0)}$. We define

$$w_y = \exp(-\alpha d(y, \hat{y}_k^{(0)})), \quad P_k(y) = \frac{w_y}{\sum_{j \in \mathcal{Y}} w_j}, \tag{2}$$

where $\alpha > 0$ controls the decay steepness. This construction enforces ordinal consistency: deviations farther from the anchor receive lower plausibility than nearby states.

Final ICAP decision. The calibrated observed state is obtained by minimizing the expected absolute error under the reconstructed distribution:

$$\hat{y}_k = \arg\min_{y \in \mathcal{Y}} \sum_{j \in \mathcal{Y}} |y - j| P_k(j). \tag{3}$$

This output preserves both interpretability and uncertainty, providing a soft but structured estimate of student cognitive engagement.

3.4 Advisor: Feedback Generation

The Advisor synthesizes teacher-side context, Assessor outputs, and student evidence into time-aligned teaching recommendations. It does not re-estimate ICAP levels from raw video; it reasons over c_k, $\hat{y}_k$, and $\mathcal{E}_k$ to explain why a segment succeeded or failed relative to the intended pedagogical objective.

Discrepancy reasoning. For each instructional unit $\mathcal{U}_k$, the Advisor compares the intended ICAP target c_k with the observed ICAP state $\hat{y}_k$. The resulting discrepancy indicates whether the instructional design and the actual student response are aligned:

$$\Delta_k = \phi(c_k, \hat{y}_k), \tag{4}$$

where $\phi(\cdot)$ measures the pedagogical gap between intended and observed cognitive states.

Recommendation synthesis. Conditioned on the instructional content $\mathcal{U}_k$, the student evidence $\mathcal{E}_k$, and the discrepancy Δ_k, the Advisor generates a time-local recommendation:

$$\mathcal{F}_k = \Psi(\mathcal{U}_k, \mathcal{E}_k, \Delta_k). \tag{5}$$

The generated feedback is grounded in concrete classroom evidence and is explicitly aligned with the intended ICAP transition. For example, when the instructor intends to elicit Constructive or Interactive engagement but the observed state remains Passive or Active, the Advisor can recommend adding prompts, peer discussion, wait time, or more scaffolded questioning.

4 Experiment

4.1 Experiment Setup

To systematically evaluate the generalization performance and robustness of the proposed model, we construct a dedicated dataset for real-world online teaching scenarios. The dataset consists of publicly available screen recordings of live Zoom-based online lectures uploaded to YouTube by educational content creators, with the target learners predominantly being undergraduate students. The lecture content spans two broad domains, STEM and humanities, covering a diverse range of subjects with substantial variation in terminology, speaking rhythm, and instructional style. The final curated dataset contains approximately 10 hours of effective recordings involving around 300 students. To protect privacy and ensure ethical compliance,

Table 1: Quantitative evaluation results across frameworks based on expert human evaluation scores (1–5 scale).

Framework	Domain	ICAP Accuracy	Actionability	Alignment	Average
Vanilla VLM	STEM	3.50 ± 0.29	3.88 ± 0.70	3.50 ± 0.29	3.62 ± 0.36
Reflexion	STEM	3.36 ± 0.25	2.36 ± 0.25	3.55 ± 0.27	3.09 ± 0.22
TextGrad	STEM	3.91 ± 0.09	4.27 ± 0.82	4.00 ± 0.00	4.06 ± 0.15
MAVE (Ours)	STEM	**4.82 ± 0.40**	**4.73 ± 0.65**	**4.45 ± 0.52**	**4.64 ± 0.43**
Vanilla VLM	Humanities	3.50 ± 0.33	3.75 ± 0.25	3.50 ± 0.33	3.58 ± 0.25
Reflexion	Humanities	3.57 ± 0.53	2.57 ± 0.53	3.57 ± 0.53	3.24 ± 0.53
TextGrad	Humanities	4.00 ± 0.00	4.50 ± 0.28	4.10 ± 0.10	4.20 ± 0.03
MAVE (Ours)	Humanities	**4.70 ± 0.48**	**4.80 ± 0.42**	**4.50 ± 0.53**	**4.63 ± 0.40**

all facial and vocal characteristics in the dataset are anonymized prior to analysis. We use GPT-4o-mini as the backbone model for all agents and baseline methods, and provide the detailed prompt templates in Appendix D.

4.2 Metrics and Evaluation

To rigorously assess the quality of the generated pedagogical feedback, we employ a multi-dimensional evaluation framework that measures both the technical reliability and the educational effectiveness of the system. Detailed metric formulations are provided in Appendix A. Specifically, we define several core metrics to evaluate different aspects of the generated feedback: **(1) ICAP Accuracy:** Evaluates the model's ability to correctly infer students' cognitive engagement states under the ICAP framework (Interactive, Constructive, Active, and Passive). This metric measures the consistency between the predicted engagement states and expert annotations, reflecting the model's capability to capture meaningful learning behaviors from multimodal classroom signals. **(2) Alignment:** Measures the extent to which the generated feedback aligns with the predefined teaching objectives, curriculum standards, and the instructional context reflected in the lecture transcripts. This metric also evaluates logical consistency and penalizes hallucinated or self-contradictory suggestions. **(3) Actionability:** Assesses the specificity, clarity, and practical usefulness of the generated recommendations. Effective feedback should provide concrete strategies or interventions that instructors can realistically implement in future teaching sessions. We further adopt a hybrid evaluation framework that combines expert human assessment with LLM-as-a-Judge scoring, and cross-validate both results to improve evaluation reliability and robustness.

4.3 Comparisons and Baselines

To evaluate the effectiveness of MAVE in generating pedagogically grounded instructional feedback, we compare our framework against several representative LLM-based reasoning and optimization paradigms. For fair comparison, all baseline methods are reformulated to produce the same type of output as MAVE, namely time-aligned instructional recommendations conditioned on classroom context and observed student behaviors. **(1) Vanilla VLM:** This baseline directly prompts a vision-language model (VLM) to generate pedagogical feedback from classroom video, lecture transcripts, and observed student behaviors. **(2) Reflexion:** We adapt Reflexion (Shinn et al., 2023) as a self-reflective pedagogical feedback generation framework. The model first produces an initial instructional recommendation based on classroom observations and inferred student engagement states, and subsequently performs self-critique to identify potential weaknesses in its own outputs before generating refined feedback. **(3) TextGrad:** We adapt TextGrad (Yuksekgonul et al., 2024) as an iterative optimization framework based on textual gradients, where the generated gradients are treated as pedagogical feedback signals aligned with the output format of MAVE, enabling iterative refinement of instructional recommendations according to classroom evidence and observed student behaviors.

4.4 Main Results

Table 1 summarizes the quantitative evaluation results across different instructional domains. Overall, MAVE consistently outperforms all baseline methods in terms of ICAP reasoning accuracy, pedagogical alignment, and actionability of the generated feedback. The results demonstrate that generic

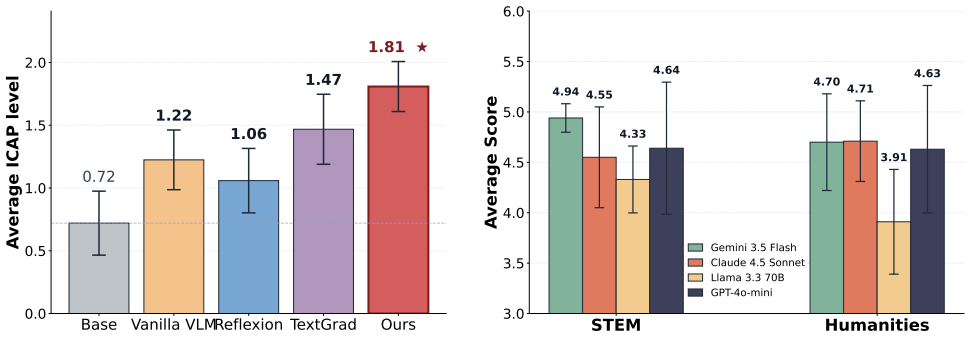

Figure 3: (a) ICAP transition analysis before and after applying pedagogical feedback. The *Base* bar denotes the initial ICAP level before feedback, while the remaining bars represent the post-feedback ICAP levels achieved after applying recommendations generated by different methods. (b) Performance comparison under different backbone LLMs across STEM and Humanities domains.

prompting, self-reflection, or purely text-based optimization strategies are insufficient for producing reliable instructional recommendations in complex multimodal classroom environments. In contrast, by explicitly modeling the discrepancy between intended instructional objectives and observed student engagement states, MAVE generates feedback that is more pedagogically grounded, interpretable, and practically actionable across both STEM and Humanities settings.

4.5 ICAP Transition Analysis

Beyond evaluating feedback quality, we further examine whether the generated recommendations can potentially improve instructional cognitive engagement under the ICAP framework. For each instructional segment, we use an LLM to revise the original teaching script according to the generated recommendation. The revised scripts are then evaluated by expert human annotators, whose ratings provide an estimated post-intervention ICAP level.

As shown in Fig. 3(a), all methods improve the ICAP level of the instructional scripts. However, MAVE achieves the largest ICAP improvement among all compared methods. This suggests that discrepancy-based reasoning and multimodal grounding enable MAVE to generate more effective and pedagogically meaningful instructional recommendations.

4.6 Sensitivity Analysis

We further evaluate the robustness of MAVE under different LLM backbones by replacing the core reasoning model with Gemini 3.5 Flash, Claude 4.5 Sonnet, Llama 3.3 70B, and GPT-4o-mini while keeping all other components unchanged. Fig. 3(b) summarizes the average performance across STEM and Humanities domains.

Table 2: Consistency analysis between expert human evaluation and LLM-as-a-Judge evaluation.

Metric	Domain	Human Score	LLM Score	Percentage Agreement
ICAP Accuracy	STEM	4.95 ± 0.04	4.82 ± 0.40	90.91%
	Humanities	4.90 ± 0.09	4.70 ± 0.48	80.0%
Actionability	STEM	5.00 ± 0.73	4.73 ± 0.65	81.82%
	Humanities	4.80 ± 0.16	4.80 ± 0.42	85.0%
Alignment	STEM	4.22 ± 0.28	4.45 ± 0.52	81.82%
	Humanities	4.60 ± 0.24	4.50 ± 0.53	85.3%

Overall, MAVE maintains relatively stable performance across different backbone models, suggesting that the effectiveness of the framework does not rely heavily on a specific proprietary LLM. Although stronger models generally achieve higher scores, the performance trends remain consistent across domains. These results indicate that the gains primarily come from the proposed discrepancy-based reasoning and multimodal grounding framework rather than solely from raw model capacity. Detailed metric breakdowns are provided in the Appendix. B.

4.7 Evaluation Analysis

To further verify the reliability of the proposed hybrid evaluation framework, we analyze the consistency between expert human evaluation and the automated *LLM-as-a-Judge* assessment. Specifically, both evaluation paradigms independently score the generated pedagogical feedback using the same predefined evaluation rubric. We then compare the resulting scores and agreement statistics across different evaluation dimensions.

As shown in Table 2, the automated evaluator demonstrates strong consistency with expert human judgment across all metrics. This result suggests that the proposed LLM-based evaluation pipeline

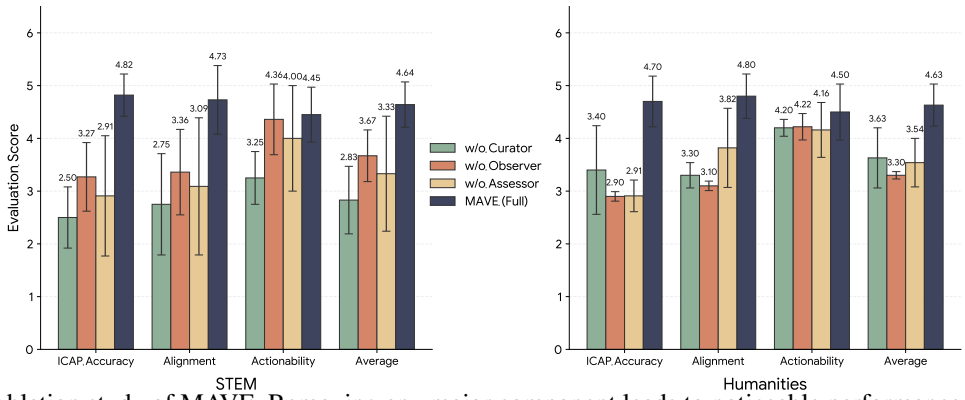

Figure 4: Ablation study of MAVE. Removing any major component leads to noticeable performance degradation across all evaluation dimensions.

can provide stable and reliable assessment signals for pedagogical feedback quality while substantially reducing the cost of large-scale manual annotation. The high agreement between human experts and the automated evaluator further validates the credibility and robustness of the proposed evaluation protocol.

4.8 Ablation Study

To better understand the contribution of each component in MAVE, we conduct an ablation study by progressively removing key modules from the full framework. Specifically, we evaluate the impact of (1) multimodal evidence grounding, (2) ordinal-consistent ICAP calibration, and (3) discrepancy-based pedagogical reasoning. All variants are evaluated using the same evaluation protocol described in Section 4.2.

The results in Figure. 4 demonstrate that each module contributes substantially to the overall performance of the framework. Removing the multimodal Observer causes the largest drop in ICAP Accuracy, indicating that grounded visual and interaction evidence is critical for reliable engagement estimation. Eliminating the ordinal-consistent calibration mechanism further degrades performance, suggesting that modeling the ordinal structure of ICAP states improves robustness and reduces noisy engagement predictions. Finally, removing discrepancy reasoning significantly weakens Alignment and Actionability, as the generated feedback becomes less connected to the gap between intended instructional objectives and observed student cognitive states. Overall, the ablation results validate that the effectiveness of MAVE arises from the synergistic interaction between multimodal grounding, calibrated engagement reasoning, and pedagogically structured discrepancy analysis.

4.9 Efficiency

We further analyze the computational efficiency of MAVE in terms of token consumption, inference latency, processing time, and monetary cost in Appendix C. Since MAVE adopts a multi-agent collaborative architecture with multimodal reasoning and iterative pedagogical analysis, its computational overhead is naturally higher than that of simple prompting-based baselines. However, the additional cost is justified by the substantial improvement in pedagogical grounding, ICAP reasoning accuracy, and actionability of the generated feedback. In practical deployment, most of the computational overhead originates from multimodal evidence extraction and multi-stage LLM reasoning. Nevertheless, the framework remains feasible for offline teaching analysis and post-class instructional feedback scenarios, where response quality and interpretability are more critical than strict real-time latency constraints.

5 Discussion and Conclusion

We presented **MAVE**, a multi-agent framework that grounds pedagogical feedback in ICAP theory and multimodal classroom evidence. By separating engagement estimation (Assessor, with ordinal calibration) from gap-driven recommendation (Advisor, with discrepancy reasoning), MAVE turns long-form online lecture videos into time-aligned, evidence-backed instructional suggestions. On real-world STEM and humanities lectures, expert evaluation shows that MAVE produces feedback that is more accurate with respect to segment-level ICAP labels, better aligned with instructional context, and more actionable than Prompting, Reflexion, and TextGrad under a matched multimodal input protocol.

Limitations. Our study is offline: we analyze recorded lectures and do not measure whether instructors later change their teaching or whether student ICAP levels shift in follow-up sessions. All systems, including baselines, share the same preprocessing and visual evidence summaries, but MAVE incurs higher inference cost due to multi-agent orchestration. Future work may link discrepancy signals to controlled classroom studies and lighter-weight deployment.

References

Reza Bahmanyar, Ambar Murillo Montes de Oca, and Mihai Datcu. 2015. The semantic gap: an exploration of user and computer perspectives in earth observation images. *IEEE Geoscience and Remote Sensing Letters*, 12(10):2046–2050.

Paulo Blikstein. 2013. Multimodal learning analytics. In *Proceedings of the third international conference on learning analytics and knowledge*, pages 102–106.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. 2024. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. In *Forty-first International Conference on Machine Learning*.

Michelene T. H. Chi and Ruth Wylie. 2014. The icap framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist*, 49(4):219–243.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first international conference on machine learning*.

Emrah Emirtekin. 2025. Large language model-powered automated assessment: a systematic review. *Applied Sciences*, 15(10):5683.

Darren Paul Fisher, Gaelle Brotto, Iris Lim, and Colette Southam. 2025. The impact of timely formative feedback on university student motivation. *Assessment & Evaluation in Higher Education*, 50(4):622–631.

Jennifer A Fredricks, Phyllis C Blumenfeld, and Alison H Paris. 2004. School engagement: Potential of the concept, state of the evidence. *Review of educational research*, 74(1):59–109.

Xian Gao, Jiacheng Ruan, Jingsheng Gao, Mingye Xie, Zongyun Zhang, Ting Liu, and Yuzhuo Fu. 2025. From motion signals to insights: A unified framework for student behavior analysis and feedback in physical education classes. *arXiv preprint arXiv:2503.06525*.

Nadezhda Gorbunova, Eduard Vorsin, Yelena Spirina, Saule Zhumagulova, and Nadezhda Popova. 2026. Barriers to artificial intelligence adoption among instructors of it-related disciplines in higher education: a survey-based study. In *Frontiers in Education*, volume 11, page 1804254. Frontiers Media SA.

Pedro Antonio Gutiérrez, Maria Perez-Ortiz, Javier Sanchez-Monedero, Francisco Fernandez-Navarro, and Cesar Hervas-Martinez. 2015. Ordinal regression methods: survey and experimental study. *IEEE Transactions on Knowledge and Data Engineering*, 28(1):127–146.

Yihan Hong, Huaiyuan Yao, Bolin Shen, Wanpeng Xu, Hua Wei, and Yushun Dong. 2026. Rulers: Locked rubrics and evidence-anchored scoring for robust llm evaluation. *arXiv preprint arXiv:2601.08654*.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 17889–17904.

Qingtang Liu, Xinyu Jiang, and Ruyi Jiang. 2025. Classroom behavior recognition using computer vision: A systematic review. *Sensors*, 25(2):373.

Hamed Monkaresi, Nigel Bosch, Rafael A Calvo, and Sidney K D'Mello. 2016. Automated detection of engagement using video-based estimation of facial expressions and heart rate. *IEEE Transactions on Affective Computing*, 8(1):15–28.

Arsha Nagrani, Sachit Menon, Ahmet Iscen, Shyamal Buch, Ramin Mehran, Nilpa Jha, Anja Hauth, Yukun Zhu, Carl Vondrick, Mikhail Sirotenko, and 1 others. 2025. Minerva: Evaluating complex video reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23968–23978.

Fumiya Okubo, Tetsuya Shiino, Tsubasa Minematsu, Yuta Taniguchi, and Atsushi Shimada. 2022. Adaptive learning support system based on automatic recommendation of personalized review materials. *IEEE Transactions on Learning Technologies*, 16(1):92–105.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Ratneswary Rasiah, Harpaljit Kaur, and Vinitha Guptan. 2020. Business continuity plan in the higher education industry: University students' perceptions of the effectiveness of academic continuity plans during covid-19 pandemic. *Applied System Innovation*, 3(4):51.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems*, 36:8634–8652.

Skyler Wang, Ned Cooper, and Margaret Eby. 2024. From human-centered to social-centered artificial intelligence: Assessing chatgpt's impact through disruptive events. *Big data & society*, 11(4):20539517241290220.

Jacob Whitehill, Zewelanji Serpell, Yi-Ching Lin, Aysha Foster, and Javier R Movellan. 2014. The faces of engagement: Automatic recognition of student engagementfrom facial expressions. *IEEE Transactions on Affective Computing*, 5(1):86–98.

Lixiang Yan, Vanessa Echeverria, Yueqiao Jin, Gloria Fernandez-Nieto, Linxuan Zhao, Xinyu Li, Riordan Alfredo, Zachari Swiecki, Dragan Gašević, and Roberto Martinez-Maldonado. 2024. Evidence-based multimodal learning analytics for feedback and reflection in collaborative learning. *British Journal of Educational Technology*, 55(5):1900–1925.

Huaiyuan Yao, Wanpeng Xu, Justin Turnau, Nadia Kellam, and Hua Wei. 2026. Instructional agents: Reducing teaching faculty workload through multi-agent instructional design. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4087–4109, Rabat, Morocco. Association for Computational Linguistics.

Xin Yi, Yue Li, Dongsheng Shi, Linlin Wang, Xiaoling Wang, and Liang He. 2025. Unified defense for large language models against jailbreak and fine-tuning attacks in education. *arXiv preprint arXiv:2511.14423*.

Mert Yuksekgonul, Federico Bianchi, Joseph Boen, Sheng Liu, Zhi Huang, Carlos Guestrin, and James Zou. 2024. Textgrad: Automatic" differentiation" via text. *arXiv preprint arXiv:2406.07496*.

Haipeng Zeng, Xinhuan Shu, Yanbang Wang, Yong Wang, Liguo Zhang, Ting-Chuen Pong, and Huamin Qu. 2020. Emotioncues: Emotion-oriented visual summarization of classroom videos. *IEEE transactions on visualization and computer graphics*, 27(7):3168–3181.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Rui Zheng, Fei Jiang, and Ruimin Shen. 2020. Intelligent student behavior analysis system for real classrooms. pages 9244–9248. IEEE.

A Metric Formulations

This appendix provides the mathematical formulations of the evaluation metrics used in Section 4.2. All metrics are computed at the segment level and averaged across the evaluation set.

ICAP Accuracy. Let $y_k \in \mathcal{Y}$ denote the expert-annotated ICAP label for instructional segment k, and let $\hat{y}_k \in \mathcal{Y}$ denote the predicted ICAP state generated by the Assessor, where

$$\mathcal{Y} = \{P, A, C, I\}.$$

The ICAP classification accuracy is defined as

$$\mathrm{Acc}_{\mathrm{ICAP}} = \frac{1}{N} \sum_{k=1}^{N} \mathbb{I}(\hat{y}_k = y_k), \tag{6}$$

where $\mathbb{I}(\cdot)$ is the indicator function and N is the total number of evaluated segments.

Since ICAP levels are ordinal, we additionally compute the mean absolute error (MAE) between predicted and annotated levels. We map

$$P = 0, \quad A = 1, \quad C = 2, \quad I = 3,$$

and define

$$\mathrm{MAE}_{\mathrm{ICAP}} = \frac{1}{N} \sum_{k=1}^{N} |\hat{y}_k - y_k|. \tag{7}$$

Alignment. Let f_k denote the generated pedagogical feedback for segment k, and let r_k denote the instructional reference constructed from teaching objectives, transcript context, and expert annotations. Alignment measures the semantic consistency between generated feedback and instructional intent:

$$\mathrm{Alignment} = \frac{1}{N} \sum_{k=1}^{N} \mathrm{Sim}(f_k, r_k), \tag{8}$$

where $\mathrm{Sim}(\cdot, \cdot)$ represents a semantic similarity function implemented using rubric-based human scoring and LLM-as-a-Judge evaluation.

Actionability. Actionability evaluates whether the generated feedback provides concrete and practically applicable instructional guidance.

B Sensitivity

Sensitivity analysis results are shown in Table. 4. MAVE maintains consistently strong performance across different LLM backbones, including Gemini 3.5 Flash, Claude 4.5 Sonnet, Llama 3.3 70B Instruct, and GPT-4o-mini, across both STEM and Humanities domains. Although stronger proprietary models generally achieve higher scores, the overall performance remains stable, with GPT-4o-mini also delivering competitive results. These findings indicate that MAVE's effectiveness is not tied to a single backbone model; instead, the gains mainly come from the proposed discrepancy-based reasoning and multimodal grounding framework.

C Efficiency

Efficiency results are reported in Table. 3, showing that MAVE achieves improved feedback quality with moderate additional computational overhead, while remaining feasible for offline teaching analysis.

D Prompt Templates

This appendix provides concise system prompt templates for the four agents used in our framework.

D.1 Curator Agent

```
You are an expert in educational discourse
    structuring.

Task:
Split a fragmented lecture transcript into
    coherent knowledge-point segments and
    infer the intended ICAP level for each
    segment.

Requirements:
1) Preserve temporal order and full
    transcript coverage.
2) Return contiguous and non-overlapping
    segments.
3) For each segment output:
   - topic
   - summary
   - expected_icap in {P, A, C, I}
   - short reason

Return STRICT JSON only.
```

Table 3: Efficiency Analysis

Method	Tokens (M)	Latency (s)	Time (min)	Cost ($)
Prompting	0.01 ± 0.001	1.4 ± 0.2	0.03 ± 0.01	0.01 ± 0.01
Reflexion	0.35 ± 0.02	1.8 ± 0.3	0.15 ± 0.04	0.76 ± 0.02
Textgrad	0.22 ± 0.01	1.8 ± 0.3	0.12 ± 0.03	0.54 ± 0.01
MAVE	0.13 ± 0.06	1.9 ± 0.4	0.20 ± 0.05	0.34 ± 0.03
MAVE detailed	13.34 ± 3.23	2.5 ± 0.5	0.65 ± 0.18	6.00 ± 0.97

Table 4: Sensitivity analysis of MAVE across different LLM backbones. Performance remains relatively stable, demonstrating framework robustness independent of a specific proprietary model.

LLM Backbone	Domain	ICAP Accuracy	Alignment	Actionability	Average
Gemini 3.5 Flash	STEM	5.00 ± 0.00	5.00 ± 0.00	4.91 ± 0.09	4.94 ± 0.02
	Humanities	4.80 ± 0.18	4.70 ± 0.46	4.60 ± 0.27	4.70 ± 0.23
Claude 4.5 Sonnet	STEM	4.55 ± 0.27	4.46 ± 0.47	4.73 ± 0.22	4.55 ± 0.25
	Humanities	4.70 ± 0.23	4.90 ± 0.10	4.60 ± 0.49	4.71 ± 0.16
Llama 3.3 70B Instruct	STEM	4.36 ± 0.26	4.73 ± 0.42	3.91 ± 0.09	4.33 ± 0.11
	Humanities	4.10 ± 0.32	3.80 ± 0.62	3.80 ± 0.18	3.91 ± 0.27
GPT-4o-mini	STEM	4.82 ± 0.40	4.73 ± 0.65	4.45 ± 0.52	4.64 ± 0.43
	Humanities	4.70 ± 0.48	4.80 ± 0.42	4.50 ± 0.53	4.63 ± 0.40

D.2 Observer Agent

```
You are a multimodal classroom behavior
    observer.

Task:
Summarize observable student evidence from
    classroom video, audio, and interaction
     signals.

Requirements:
1) Focus on posture, gaze, gestures, note-
    taking, speech, and interaction.
2) Describe only observable behaviors.
3) Do not infer emotions or intentions.
4) Produce concise evidence summaries for
    ICAP assessment.

Return STRICT JSON only.
```

D.3 Assessor Agent

```
You are an ICAP state assessor.

Task:
Infer the observed ICAP state from
    multimodal student evidence.

Requirements:
1) Predict an anchor ICAP level in {P, A, C,
     I}.
2) Apply ordinal-consistent calibration.
3) Output the final ICAP state with concise
     evidence-based reasoning.

Return STRICT JSON only.
```

D.4 Advisor Agent

```
You are an expert teaching coach.

Task:
Compare the intended ICAP level with the
    observed ICAP state and generate
    pedagogical feedback.

Requirements:
1) Identify the discrepancy between
    intended and observed engagement.
2) Explain the likely instructional issue.
3) Generate one concrete instructional
    revision.
4) Provide one actionable recommendation
    grounded in observed evidence.

Return STRICT JSON only.
```

E Case Study

To qualitatively illustrate the effectiveness of MAVE, we present a representative case study from a real-world online classroom session. In this example, the instructor was explaining a problem-solving process in a STEM lecture and intended to elicit a *Constructive* level of cognitive engagement by encouraging students to actively reason about intermediate solution steps.

However, the Observer agent detected that most students exhibited prolonged passive viewing behaviors like limited verbal interaction and infrequent note-taking activity. Based on the extracted

multimodal evidence, the Assessor estimated the observed ICAP state as predominantly *Passive*. The Advisor subsequently identified a substantial discrepancy between the intended and observed engagement states and generated the following pedagogically grounded recommendation:

In contrast, the Prompting baseline generated generic recommendations such as "increase classroom interaction" without referencing concrete classroom evidence or specifying actionable instructional interventions. Reflexion improved linguistic fluency but still lacked explicit grounding in student behaviors. TextGrad produced more structured suggestions, yet occasionally introduced recommendations weakly related to the actual classroom context.

This case study highlights that MAVE not only recognizes student engagement states more accurately, but also generates feedback that is temporally grounded, pedagogically coherent, and directly connected to observable classroom dynamics.

Observation:

The highest cognitive level among currently speaking students is 'c' (Constructive). Although three students have actively spoken (about 10% coverage), this participation is limited to brief responses and recognition of common cutting issues without individual cognitive construction of knowledge. The transcript shows the teacher diagnosing typical problems and giving tips, but students have not yet independently generated knowledge or produced individual outputs, e.g., personal reflections or individual problem-solving: \"okay so have any of you had trouble cutting the foam board all right i see one i see two three\". This suggests that while some students are vocal, most remain silent and the class-wide cognitive engagement is bottlenecked by low active contributions and lack of individual knowledge construction.

Suggestions:

"proposed_rewrite": "Transform the activity by inviting each student to independently produce a personal diagnostic log of their foam board cutting attempts, identifying their unique challenges and strategies used. This could be facilitated through a written or drawn self-reflection task where students individually analyze their technique, blade condition, and pressure applied, without any peer interaction or discussion. This shift prioritizes solitary cognitive construction and authentic knowledge generation aligned with target Mode C.",

"scaffolding_strategy": "1) Individual Reflection Journaling: Prompt students to silently write a detailed reflection about their personal experience with foam board cutting, explicitly requiring them to identify the problem causes they encountered and propose tailored solutions based on the teacher's tips. 2) Solo Visual Mapping: Require each student to create an independent process map or diagram illustrating their step-by-step approach to cutting foam board and highlighting potential pitfalls, ensuring that this is done alone to foster individual cognitive construction.",

"new_question": "Can you independently describe your specific challenges when cutting foam board and outline a personalized step-by-step strategy you would use to improve your technique based on what you have learned so far?",

"reference_answer": "When I cut the foam board, I noticed that my blade was dull, which made it hard to get a clean cut. I also realized I was pressing down too hard, which caused jagged edges. To improve, I will first check that my blade is sharp before starting. Then, I will hold the blade at a slight angle and use two to three light passes instead of one hard cut. This method should allow me to cut smoothly and prevent tearing the foam.",

Figure 5: Illustrative case study of MAVE. The figure visualizes the instructional context, multimodal student observations, inferred ICAP states, and the grounded pedagogical recommendation generated by the framework.