TeachUp: Facilitating Early-Stage Teachers to Learn Instructional Strategies from Classroom Videos with Reflective Support

Haoxiang Fan
School of Artificial Intelligence
Sun Yat-sen University
Zhuhai, China
fanhx6@mail2.sysu.edu.cn

Ding Lei
Institute of Artificial Intelligence and Brain Sciences
University of Macau
Macau, China
mc56574@um.edu.mo

Zaihong Zheng*
School of Artificial Intelligence
Sun Yat-sen University
Zhuhai, China
cheng256@mail2.sysu.edu.cn

Jiale Li*
School of Artificial Intelligence
Sun Yat-sen University
Zhuhai, China
lijle27@mail2.sysu.edu.cn

Jionghao Lin
The University of Hong Kong
Hong Kong, China
jionghao@hku.hk

Jian Yin
School of Artificial Intelligence
Sun Yat-sen University
Zhuhai, China
issjyin@mail.sysu.edu.cn

Zhenhui Peng†
School of Artificial Intelligence
Sun Yat-sen University
Zhuhai, China
pengzhh29@mail.sysu.edu.cn

Abstract

Recorded videos of offline open classes provide good examples for early-stage teachers to learn instructional strategies, e.g., how to organize cooperative learning. However, learning by watching these videos is challenging, as these strategies are implicitly performed, and it lacks in-situ reflective support. In this paper, via a formative study (N=9), we design *TeachUp* to support the learning of instructional strategies from classroom teaching videos. *TeachUp* adopts an LLM-powered pipeline to detect nine instructional strategies in videos (precision = 63.4%), provides reflective questions and hints while watching, and generates customized practices with reflective feedback. A within-subjects study (N=16) shows that compared to a traditional video-playing and self-practicing baseline, early-stage teachers with *TeachUp* are more engaged in learning and perform better in applying learned strategies to new tasks. Interviews with four in-service teachers further generalize our findings and *TeachUp*'s use cases. We discuss practical implications for fostering video-based learning of instructional strategies.

CCS Concepts

• **Human-centered computing → Interactive systems and tools**; • **Applied computing → Interactive learning environments**.

Keywords

Video-based learning support, instructional strategy, reflective support

ACM Reference Format:
Haoxiang Fan, Ding Lei, Zaihong Zheng, Jiale Li, Jionghao Lin, Jian Yin, and Zhenhui Peng. 2026. TeachUp: Facilitating Early-Stage Teachers to Learn Instructional Strategies from Classroom Videos with Reflective Support . In *The 39th Annual ACM Symposium on User Interface Software and Technology (UIST '26), November 02–05, 2026, Detroit, MI, USA.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3830398.3830693

*These authors contributed equally to this work.
†Corresponding author.

UIST '26, Detroit, MI, USA

ACM ISBN 979-8-4007-2856-3/2026/11
https://doi.org/10.1145/3830398.3830693

1 Introduction

Instructional strategies, such as cooperative learning, setting objectives, and providing feedback [36], are useful for enhancing student achievement in offline classrooms. This paper targets **early-stage teachers**, including pre-service teachers (*e.g.,* students in normal universities) and in-service teachers with no more than one year of teaching experience, and their learning of instructional strategies with classroom teaching videos. Specifically, the excellent open classes like the award-winning ones are usually recorded and available on video-sharing platforms. These videos serve as valuable learning materials, in which well-trained teachers demonstrate good practices of instructional strategies in a real classroom. Besides, such video-based learning (VBL) has been increasingly adopted in professional development programs for both pre-service and in-service teachers [17, 30, 47]. Teacher-oriented VBL includes observing other teachers' classroom videos and reviewing one's own teaching videos [17], and this work focuses on the former.

Despite the benefits, learning instructional strategies from classroom videos is challenging for early-stage teachers. Unlike directly replicable actions, an instructional strategy is defined by when and why it is enacted: because pedagogical content knowledge integrates pedagogy with disciplinary content [9, 49], the same strategy can be enacted differently across subjects and student states. Yet classroom videos mainly expose teachers' observable actions and leave the underlying pedagogical reasoning implicit [52], which brings two challenges. First, these teachers have to recognize and analyze the instructional strategies performed by the teachers in the videos, while they usually lack teaching experience using these strategies. Recorded videos of open classes are often pedagogically complex, because they feature teachers effectively demonstrating multiple strategies, yet lack the descriptions and tags needed to navigate the video content based on instructional strategies [3]. As a result, this mentally demanding and time-consuming process often leads early-stage teachers to avoid longer, more complex videos, causing potentially high-value pedagogical examples to be ignored [3, 47]. Second, early-stage teachers lack in-situ reflective support while watching classroom videos and opportunities to practice the observed strategies afterward, both of which are essential to learning instructional strategies. In fact, many teachers are prone to copying the teaching activities in the videos rather than interpreting and evaluating them during VBL [7], missing crucial opportunities for professional growth. Moreover, improving teaching skills depends heavily on practices in scenarios close to real classroom settings [19, 53].

Yet existing systems fall short on both sides. For the first, video annotation tools help teachers structure reflection on classroom recordings [46], and computational approaches summarize classroom videos by student emotions [64] or engagement [48]; neither surfaces which strategy a teacher is enacting, nor why it fits that classroom moment. For the second, VBL systems for other skills, *e.g.*, surgical procedures [40], public speaking [59], and running [32], assume predefined skills with observable performance dimensions, and thus do not support rehearsing a strategy whose enactment depends on the subject, the students, and the timing. Scenario-based teacher-training systems [43, 56, 58], meanwhile, are text-based or pre-scripted, and do not connect practice to the strategies just observed in a video. Given these two challenges, we ask the following research question: **How to design a reflective interaction loop to make early-stage teachers' learning of instructional strategies from classroom videos more effective and engaging?**

In this paper, we propose *TeachUp*, an interactive system that supports early-stage teachers to learn instructional strategies based on offline classroom videos. We followed a user-centered design approach to develop and evaluate *TeachUp*. Initially, we interviewed four early-stage teachers to understand their practices and challenges of learning to teach from classroom videos. We also interviewed five experienced teachers to gather feedback and requirements for *TeachUp*'s prototype from experts' views. The findings highlight the needs to deepen reflection on instructional strategies while watching the videos and to facilitate practicing these strategies in authentic scenarios after watching. To address these needs, we designed and implemented *TeachUp* based on the interview findings and theories of teacher education [9, 21, 49]. *TeachUp* structures reflection and practice into a three-stage interaction loop. To help users locate relevant examples, an LLM-powered pipeline suggests clips containing nine instructional-strategy categories defined by Marzano et al. [38] (average precision: 0.634). While watching the videos, users receive generated reflective questions on the teachers' behaviors and alternative ways to enact a strategy. Subsequently, *TeachUp* provides a generated scenario, tasks as hints for users to practice the learned strategy by speaking to the camera, similar to microteaching training [1]. Finally, users can get a personalized report generated by a multi-modal large language model, which helps reflection by contrasting the user's performance with that in the watched video.

To evaluate *TeachUp*'s effectiveness and user experience, we conducted a within-subjects study with 16 early-stage teachers, using a baseline system that supports conventional video watching and self-practicing in front of a camera. Results showed that *TeachUp* significantly improves participants' performance in applying the learned strategy from the video to a new teaching task. Users with *TeachUp* reported being significantly more engaged in the learning session and increased teaching confidence. Participants especially praised the adaptive support of reflections during the watching, practicing, and evaluation stages. We further gathered in-depth insights via interviews with two early-stage teachers and two experienced teachers who freely used *TeachUp* to learn any instruction strategy in videos of their specialized subjects. They indicated that the LLM-generated content in *TeachUp* was overall helpful but sometimes distracting. Together, these findings suggest that *TeachUp* could support teachers to learn from others' teaching videos effectively for professional development at scale.

In summary, this paper has three contributions. First, we contribute the design and implementation of a reflective interaction loop that integrates watching, microteaching, and comparison-based feedback to help early-stage teachers locate, transfer, rehearse, compare, and revise instructional strategies learned from classroom videos. Second, via a within-subjects study, we demonstrate *TeachUp*'s usefulness in improving learning engagement, strategy noticing, and performance in applying the learned strategy to new teaching tasks. Third, as an enabling component, we contribute the first benchmark for instructional strategy detection in classroom videos and a preliminary computational pipeline, intended as starting points for future research.[1]

2 Related Work

2.1 Video-Based Learning for Teachers' Professional Development

Over the past 20 years, video-based learning (VBL) — including watching videos of other teachers as well as self-teaching videos [68] — has been increasingly integrated into teacher education [17, 47, 57], with empirical studies examining its effectiveness in professional development (PD) programs [17] and online self-learning [7]. For example, *STeLLA* provides the Lesson Analysis Protocol (LAP), a tool that helps teachers analyze classroom videos through a *claim–evidence–reasoning–alternative* process focused on science content storyline, to facilitate reflection and critical thinking during PD programs [54], which we integrate into *TeachUp*. When watching these videos, teachers tend to focus on general teaching moves [6] and, more specifically, on learning how to implement instructional strategies [17, 50]. However, the effectiveness of VBL

[1]We provide examples, prompts, and dataset in supplementary materials.

remains debated [51, 67]. Two key reasons have been identified: the sheer volume and pedagogical complexity of available videos makes it hard for early-stage teachers to locate useful clips [3, 47], and the absence of reflective thinking — *e.g.,* evaluating video quality [51, 67] and considering how to transfer strategies to one's own classroom [19, 53] — limits competence development.

Beyond teacher education, VBL has supported training in surgical procedures [40], voice modulation in public speaking through exemplar matching [59], and running technique through novice–expert pose comparison [32]. These systems rely on predefined target skills and observable performance dimensions, and they do not address a central need in teacher VBL: locating implicit strategies in complex classroom videos and interpreting their context-dependent rationale to guide transfer, rehearsal, and revision. Our work is motivated by the benefits and gaps of VBL for developing teaching competence and developing *TeachUp* with strategy detection and reflective support to facilitate VBL of instructional strategies.

2.2 Analyses of Classroom Videos

Prior work has extracted actionable structure from videos for skill acquisition, from crowdsourced segmentation in how-to videos [23] to multimodal analysis for formative feedback [59]. In education, such analysis has supported reflection and orchestration by surfacing classroom affective dynamics, online learning status, teamwork processes, and discussion patterns [14, 34, 42, 64, 66]. For example, Zeng et al. [64] help teachers better understand classroom affective dynamics, provide timely interventions in both online and offline settings via in-class video analysis. *Glancee* [34] provides an adaptable system for instructors to grasp student learning status in synchronous online classes. Additionally, Ngoon et al. [42] support teachers' personalized reflection by visualizing classroom discussion data. Instead of relying on videos, which require considerable effort, it uses transcripts with further processing, suggesting that classroom transcripts alone have the potential to support applications for effective teacher professional development. Nevertheless, there exists a gap in computationally identifying instructional strategies in the classroom videos and analyzing learners' performances of practicing these strategies. We develop an LLM-powered computational pipeline to detect instructional strategies based on video scripts and prompt multi-modal large models to evaluate learners' performances in practicing videos.

2.3 Reflective Support Tools for Learners

Reflective learning—the process of internally examining an experience to gain deeper understanding and new perspectives [12]—has motivated HCI research on interactive systems that promote reflection [8, 16], spanning writing [35, 41], well-being [24, 28], and professional skills [4, 22, 71]. For example, Rich and Hannafin [46] introduced a video annotation tool to assist teachers in reflecting on their classroom teaching and proved its effectiveness. Researchers have begun to leverage large language models to provide personalized feedback and richer reflective interactions [5, 63]. For example, *Friction* [65] provides LLM-driven immediate feedback in writing tasks, inspiring us to explore similar instant reflection support for teachers. Additionally, *TutorUp* [43] leverages LLM-generated content to help tutors reflect on their practice to improve their ability to address engagement challenges within text-based simulated student environments. However, previous reflective learning support tools focus on either pre-scripted or text-based settings, and they seldom look into learners' needs during VBL of instructional strategies. We conduct interviews with both early-stage and experienced teachers to inform the design of reflective support in *TeachUp*'s watching, practicing, and evaluation stages.

3 Formative Study

To inform the design of *TeachUp*, we conducted a formative study with nine participants to understand their practices, challenges, and requirements in video-based learning of instructional strategies.

3.1 Participants and Procedure

We used a snowballing approach to recruit nine participants (Table 3 in Appendix), starting from a local normal university and an elementary school. Four of them (FN1-4) are our targeted early-stage teachers who are under training in teaching chemistry, history, physics, or music. We also include three very experienced teachers (FE1-3) with over 20 years of teaching experience, as well as two relatively experienced teachers (FE4-5) with three years of teaching experience in Chinese or Mathematics in the elementary school. These participants with varying teaching experiences can inform *TeachUp*'s designs from both learners' and trainers' perspectives.

We conducted one-on-one semi-structured interviews with each participant, either online (FN1-4) or offline (FE1-5), lasting between 45 and 60 minutes. We began with participants sharing their experiences with and viewpoints on VBL. Then, we presented a low-fidelity prototype of a VBL support tool in a slide deck to gather feedback and design requirements. The prototype consists of two main pages: one for playing videos, which provides some hints and basic functionalities, and the other that outlines a possible user assessment. The slide deck used in the formative study is shown in Figure 5 and Figure 6 in Appendix. We applied Braun and Clarke's six-phase thematic analysis framework [18] to analyze the interview data. One researcher initially coded all the qualitative data, and another carefully reviewed the codes to ensure accuracy and completeness. Through iterative discussions, the two authors reached a consensus and identified three themes about early-stage teachers' focuses during learning, challenges, and suggestions for the VBL support tool.

3.2 Findings and Design Requirements

Finding 1: Early-stage teachers focus on learning instructional strategies within the context of specific subjects. All early-stage teachers reported focusing on learning from videos about what instructional strategies were used and how to use them within subject-specific contexts. For example, FN2 said, *"I would pay attention to how to give feedback on students' pitch problems and how to explain abstract concepts such as harmony."* This finding reveals that when teachers notice instructional strategies in the videos, their observations on the use of these strategies are highly contextualized to the teaching content.

Finding 2: Developing teaching competence in traditional VBL can be challenging for early-stage teachers. All early-stage teachers expressed a desire to develop a deeper understanding of instructional strategies during the VBL process. However, there are two challenges that hinder the development of teaching competence. First, traditional videos tend to showcase only surface-level activities while concealing the internal cognitive processes, thus lacking the "thinking cues" that would explain the pedagogical reasoning behind those actions. FE2 and FE3 highlighted the risk that early-stage teachers who only imitate the activities of expert teachers, without possessing the necessary skills, may struggle to maintain classroom control. Second, six participants (FE1, FE2, FE4, FE5, FN1, FN2) noted that the instructional strategies and pacing in the videos were tailored to students with specific prior knowledge and learning abilities, which often did not match their own classroom contexts—especially in award-winning open classes featuring highly cooperative students. FE4, who already had three years of teaching experience, noted, *"Sometimes when watching how famous teachers teach in developed regions, I realized that the students are in very good condition. What if I face students with weak knowledge foundations?"*

Finding 3: *TeachUp* should provide reflective support throughout the video-based learning process. After reviewing the prototype slide deck we provided, both early-stage and experienced teachers recognized the potential of reflection and evaluation to enhance VBL outcomes. FE2 and FN1 suggested that the system should prompt users to reflect right after watching a video and after they practice teaching on camera, rather than only providing feedback at the end.

The focus on learning instructional strategies within subject-specific contexts (Finding 1) and need to uncover the pedagogical reasoning underlying teachers' actions point to our first design requirement: **DR1) *TeachUp* should help users accumulate observations of subject-specific contextual teaching strategies from the videos.** This requirement aligns with pedagogical content knowledge that integrates pedagogy and disciplinary content to promote professional development [49]. Besides, as early-stage teachers often lack teaching competence and would face classroom contexts different from the videos' (Finding 2), personalized practices and reflective feedback based on the video content (Finding 3) are needed: **DR2) *TeachUp* should provide personalized training tasks and feedback to bridge the gap between VBL and real classroom teaching.** Lastly, as reflection is a key mechanism for transforming experience into learning [11, 29] and participants expect reflective support throughout the learning process (Finding 3), we have: **DR3) *TeachUp* should encourage continuous and timely reflection to enhance early-stage teachers' understanding of instructional strategies throughout the VBL process.**

3.3 Marzano's Nine Instructional Strategies

As suggested by our highly experienced teachers (FE1-3), *TeachUp* aims at supporting the learning of nine instructional strategies, which are identified by Marzano et al. [38] to be most effective among forty strategies in improving students' outcomes in the classroom. These categories are especially suitable here because they describe observable, cross-subject instructional moves that can appear as video-identifiable classroom events and can later be rehearsed in practice [37]. The nine strategies are: (1) identifying similarities and differences; (2) summarizing and note taking; (3) reinforcing effort and providing recognition; (4) homework and practice; (5) nonlinguistic representations; (6) cooperative learning; (7) setting objectives and providing feedback; (8) generating and testing hypotheses; and (9) activating prior knowledge with cues, questions, and advance organizers. The detailed definitions and some examples of the nine strategies are shown in Table 2 in Appendix. This same framework has also been extended to technology-mediated instruction, including work that maps the nine categories to classroom technologies, online / blended course design, and empirical studies of teacher uptake [2, 10, 44].

4 *TeachUp* Design and Implementation

In this section, we first describe a user scenario to walk through *TeachUp*'s interface design, followed by its implementation.

4.1 Interface and Interaction Design

Consider Harry, an early-stage teacher who just graduated from a normal university and will teach fourth-grade mathematics for the first time next semester. He turns to *TeachUp*, hoping to learn how to effectively *activate students' prior knowledge* from others' classroom videos. After selecting a 10-minute video clip which is tagged with this instructional strategy and about "The Meaning of Fractions" in *TeachUp*'s database, Harry engages in the following three learning stages.

Watching a video clip related to targeted strategy. As shown in Figure 1, Harry can review the classroom background (A), a well-contextualized description generated by analyzing the selected clip in relation to its full lesson video (DR1). At any time of the watching stage, he can check a series of reflective hints (DR3), based on a revised version of the Lesson Analysis Protocol (LAP) [54]. These include a claim identifying the target strategy and a learner-led reflection on its suitability (B), evidence-rich reflection questions with clickable timestamps that jump the video to specific moments (C), and an alternative prompt encouraging him to articulate his own lesson design ideas (D). Harry can respond to the reflection question and get feedback (C'). He can click "Set Practice Background & Prepare" to proceed to the "Practice" stage.

Practicing the learned strategy. *TeachUp* adapts the Microteaching Cycle (*plan* → *teach* → *feedback*) [1, 45] to support personalized training (DR2). Specifically, Harry sets a practice goal of teaching "the meaning of percentages" and receives a generated related teaching scenario (Figure 2 A), along with three possible starting points for implementing the teaching strategy (B). He selects "Introduce percentage after reviewing fractions" (B') and makes plans for three related tasks (C) — two emphasizing imitation of the expert's strategy and one focusing on language, gestures, and tone. Once ready, Harry conducts a four-minute microteaching session. After that, he records a self-reflection and submits the practice (D).

Evaluating performance of practiced strategy. After submission, Harry will receive a personalized report on his task performance (DR2, DR3). The report covers verbal content, tone, gestures, and facial expressions (Figure 2 C'), and is grounded in his practicing video and the watched clip. It highlights Harry's strengths based on multimodal analysis, expands on his self-identified areas for improvement with suggestions grounded in the instructional

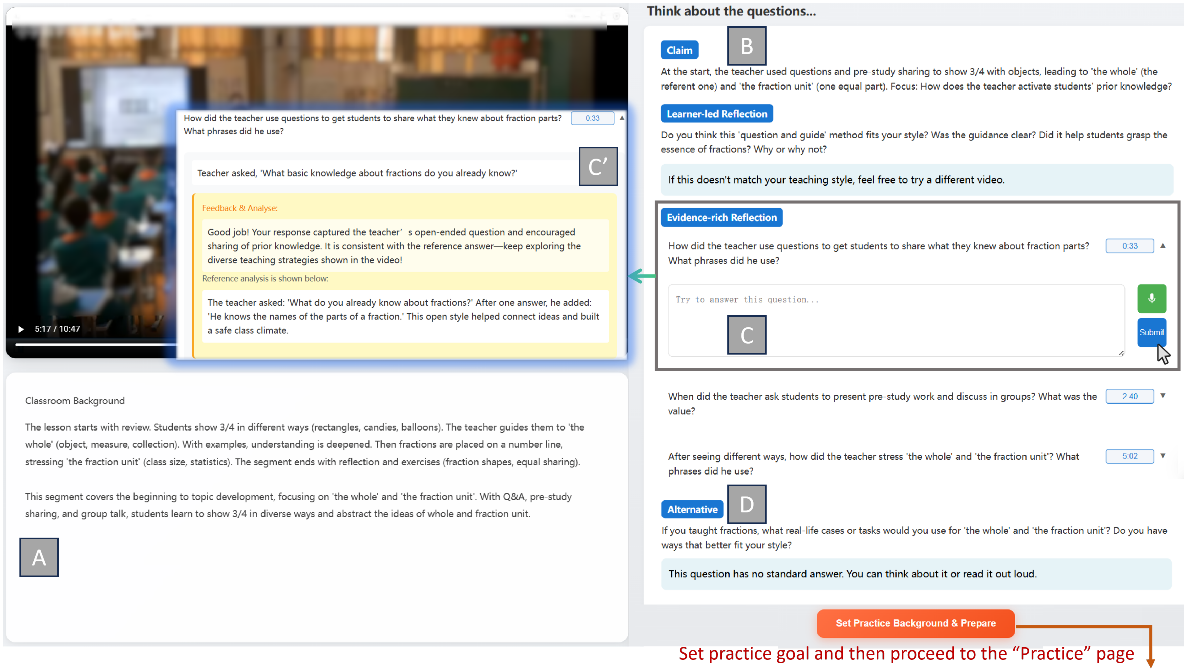

Figure 1: *TeachUp*'s interface in the watching stage. The interface is originally in Chinese and translated by GPT-4.1.

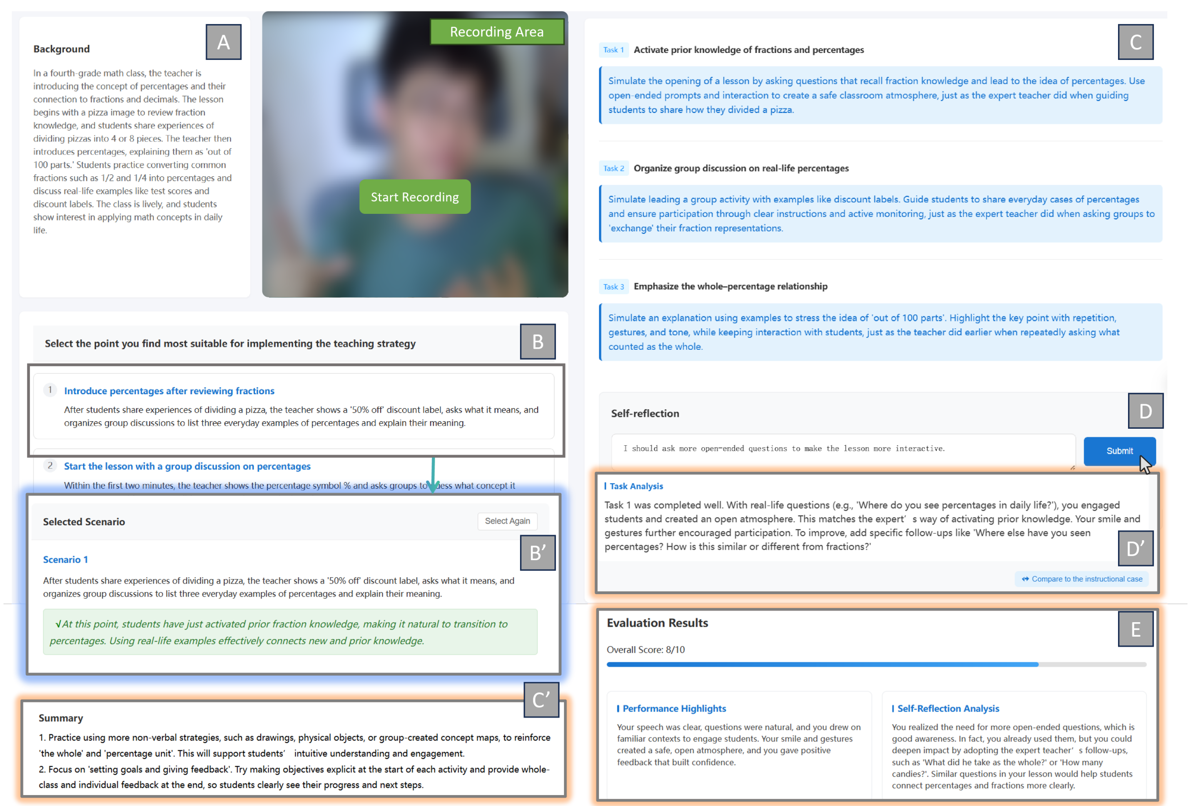

Figure 2: *TeachUp*'s interface in the practicing stage (A, B, C, D) and evaluating stage (A, B', C', D', E). The interface is originally in Chinese and translated by GPT-4.1.

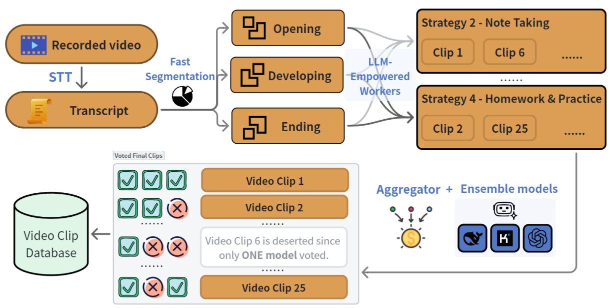

Figure 3: The computational pipeline of detecting instructional strategies in classroom videos.

case, and provides task-by-task feedback (D') with a "Compare to the instructional case" button for side-by-side review. Finally, Harry can check a summary that recommends specific categories of strategies to study next. Harry now has learned how to effectively *activate students' prior knowledge* from the classroom video.

4.2 System Implementation

4.2.1 Detecting Instructional Strategies in Classroom Videos. To enable the system to provide relevant examples, a strategy-detection pipeline that prepares a database of candidate video clips for the instructional strategies in Section 3.3 is needed.

Benchmark. We randomly selected a grade (Grade 1 to Grade 12), a subject, and then one textbook unit as the candidate topic. We searched the topic keywords on *Bilibili* and collected three videos per topic that satisfied the criteria of "40–50 minutes in length" and complete classroom recordings under the default search setting. In total, three units were selected, yielding nine videos. Two authors independently annotated the videos after agreeing on the classification criteria. To assess annotation reliability, segments with identical labels and temporal $IoU \geq 0.5$ were considered the same instructional strategy instance. The two annotators agreed on 85 instructional events; one annotator additionally marked 1 event and the other 4 events missed by their counterpart, corresponding to an agreement rate of 94.4%. The remaining disagreements (five segments) were resolved through discussion.Because the task involves detecting multiple temporal segments in long videos rather than assigning a single label per instance, traditional inter-rater statistics such as Cohen's κ or ICC are not applicable. The final benchmark contains 89 strategy instances in nine instructional strategy categories, with 6–15 segments per category. We released this benchmark as supplementary material.

Pipeline. On the benchmark, we explored multiple modalities for instructional strategy detection, including vision (*e.g.,* large vision-language models, pose detection, emotion recognition), audio (*e.g.,* Qwen-2-Audio), and text, as well as different workflow architectures. However, including more modalities does not improve performance of our pipeline, potentially because of the small size of benchmark. Balancing scalability, efficiency, and accuracy, we finally adopted a pipeline that primarily relies on transcribed spoken text as input. As shown in Figure 3, the pipeline inputs the sentence-level transcript (by Tingwu Big Model [2]) of a classroom teaching video and outputs a set of video clips (s_i), each with a strategy category (cat_i), a confidence score (conf_i), and precise timestamps (t_{s_i}, t_{e_i}). We first conduct a fast segmentation, as the nine teaching strategies are typically associated with different lesson stages, *e.g.,* "cues, questions, and advance organizers" often occur at the beginning, while "homework and practice" typically appears in the middle to late stages. Specifically, we prompt an LLM to coarsely divide the lesson into three stages: the *opening* spans from the start of class to the last review of prior knowledge; the *developing* covers the introduction and teaching of new content; and the *ending* begins with the first in-class exercise. The three stages overlap partially to reduce missed detections. Then, we designed parallelized LLM workers that collectively detect the nine instructional strategies, with each worker responsible for detecting events in its assigned lesson stage. Each worker outputs strategy segments in the form $(t_{s_i}, t_{e_i}, \mathrm{cat}_i)$. Lastly, we designed a rule-based aggregator. For any two segments $S_i = (t_{s_i}, t_{e_i})$ and $S_j = (t_{s_j}, t_{e_j})$ with the same category cat_k, if $t_{s_i} \leq t_{e_j} \wedge t_{s_j} \leq t_{e_i}$, they are merged into $S_{\text{new}} = (\min(t_{s_i}, t_{s_j}), \max(t_{e_i}, t_{e_j}))$. Merging is repeated until convergence, producing the final output $[(s_1, \mathrm{cat}_1, \mathrm{conf}_1), (s_2, \mathrm{cat}_2, \mathrm{conf}_2), \ldots]$.

To improve detection precision, we first applied a zero-shot prompt on eight carefully selected high-quality videos and analyzed the models' typical errors (*e.g.,* for cooperative learning, the model only detected "how to organize group discussions" but missed "how to evaluate the outcomes"). We then incorporated transcript segments as positive few-shot examples and frequently misclassified segments as negative **few-shot** examples. For each event, we designed two to four few-shot examples. To improve reliability, we applied **ensemble** with three LLMs—DeepSeek-v3 [31], Kimi-k2, and GPT-4.1—in parallel to determine the final video clips. Specifically, for each instructional event type, predictions from all models are clustered. The first prediction initializes a cluster, and each subsequent event is merged if its temporal overlap ratio exceeds 0.5; otherwise, it starts a new cluster. A cluster is accepted only if it contains predictions from at least two models. For each accepted cluster, the final boundary is defined by the earliest start and latest end times, and the confidence score is the average of member scores.

Technical evaluation. We evaluate the proposed pipeline using IoU, Precision, Recall, and F1-Score. For a predicted segment S_{pred} and a ground-truth segment S_{gt}, We prioritized precision because the downstream task requires reliable instructional segments for learners, even at the cost of lower recall. We compared the ensemble with individual models on the benchmark across IoU thresholds from 0.1 to 0.8. The ensemble consistently achieves higher precision than individual models, while performance for all models gradually decreases as the IoU threshold becomes stricter. At $IoU = 0.5$—a common threshold in object detection [15]—the ensemble achieves a precision of 0.634 of detecting nine strategies. In contrast, a zero-shot baseline using a single LLM prompted only with strategy definitions performed poorly (precision = 0.095, recall = 0.191, F1 = 0.127, at $IoU = 0.5$). Overall, these results provide an initial feasibility check of the pipeline as an enabling component for *TeachUp*.

[2]https://tingwu.aliyun.com

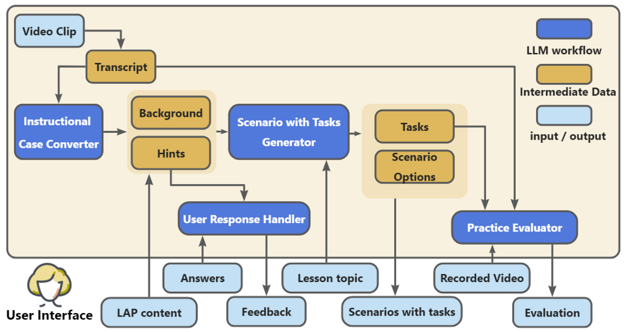

Figure 4: The architecture of *TeachUp*. The Video Clip is the output of the pipeline introduced in Section 4.2.1.

4.2.2 Supporting User Interactions. As shown in Figure 4, we developed four modules that enable user interaction with *TeachUp*. These modules are driven by an LLM (GPT-4.1) and a multimodal LLM (Qwen-2.5-Omni [61]), connected through a modular workflow that balances context richness and response latency.

With the user-selected video clip, the *Instructional Case Converter* transforms it into instructional cases, including background explanations and hints. To minimize user waiting time, we preprocess all video clips offline before releasing *TeachUp* to users. Once users respond to a reflection question in the watching stage, the *User Response Handler* provides immediate feedback using the user's answer, the claim, reflection questions, and reference analysis, excluding video-related information such as transcripts or background. Keeping the prompt short reduces LLM latency and enables real-time feedback. The *Scenario with Tasks Generator* generates practice scenarios and associated tasks by prompting an LLM to remain neutral and objective, avoiding evaluative cues (*e.g.*, "however", "although...but...") that might reveal the correct answer. This prompt includes the background and hints from the watching stage, so that the generated tasks reflect the instructional strategies observed in the video. The *Practice Evaluator* uses a multimodal LLM to evaluate users' microteaching practices. Because our used model accepts at most 40 seconds of video, the recording is resized and divided into 30-second windows with a 5-second overlap. For each segment, the model outputs a JSON containing a transcript, audio–visual observations, and evaluation suggestions. Segment-level results are concatenated and summarized into a final assessment. To reduce inference time, multimodal evaluation runs in parallel workers, enabling a full microteaching video to be processed in under five minutes.

5 User Study

We conducted a within-subjects study with 16 early-stage teachers to evaluate *TeachUp*, guided by two research questions: **RQ1:** How effectively does *TeachUp* support teachers in learning instructional strategies from classroom videos? **RQ2:** What are users' perceptions of *TeachUp* in terms of usability and usefulness?

5.1 Participants

We recruited 16 participants (P1-P16, 8 males and 8 females, age: Mean = 22.19, SD = 2.01) through word-of-mouth. All participants were pre-service teachers recruited from normal universities; P9–P16 additionally reported teaching experience, including internship experience (see Table 4 in the Appendix for detailed backgrounds). Five participants reported conducting video-based learning for their professional development weekly, four monthly, two yearly, three rarely, and two never. Twelve participants are daily users of large language models, and the remaining four are weekly users.

5.2 Experimental Setup

5.2.1 Conditions. In the ***TeachUp* condition**, participants used *TeachUp* following the "watching–practice–evaluation" stages within one instructional case. In the **baseline condition**, participants used a baseline system with a user interface similar to *TeachUp* but without AI-generated hints (Figure 1 B, C, D), practice tasks (Figure 2 C), and evaluation (Figure 2 C', D', E). To ensure a fair comparison and reduce participant workload, we provided the description of the most suitable scenario (Figure 2 B') in the baseline condition.

5.2.2 Task. We controlled the subjects and instructional strategies to compare participants' learning outcome and experience in two conditions. Specifically, we carefully selected four easy-to-understand content knowledge topics about Chinese Language and Mathematics in elementary schools. The Mathematics Task focused on the strategy of *Cues, Questions, and Advance Organizers*, with *Meaning of Fractions* as the VBL content and *Meaning of Percentages* for the test after the learning session. The Chinese Language Task focused on *Cooperative Learning*, with *The Feet of the Boston Ivy* as the VBL content and *The Cricket's Dwelling* for the test. The orders of participants' tasks and used systems were counterbalanced.

5.2.3 Procedure. Each experimental session was conducted either offline or online and lasted 80–90 minutes. One day before the session, participants received brief descriptions of the two instructional strategies and annotated PDFs of the relevant textbooks and lesson plans. They were asked to spend at least 10 minutes reviewing them. On the day of the experiment, participants were prompted to prepare themselves for a classroom lecture. As an incentive, participants with top-3 performances in the test would receive a 50 RMB bonus in addition to 100 RMB compensation. Each condition began with a 10-minute introduction in which a researcher explained the teaching strategies and demonstrated the system. Participants then spent 12 minutes watching instructional examples, followed by orally answering *"What did you learn from the video about the instructional strategies?"* and a 10-minute simulated teaching practice. In the *TeachUp* condition, participants additionally spent three minutes reviewing the system-generated evaluation report. At the end of each condition, participants completed a questionnaire and a teaching test, in which they were given a new scenario, three minutes to prepare, and recorded a 3–5 minute teaching demonstration. There was a 10-minute break between the two conditions. The entire procedure concluded with a 10-minute interview, asking about participants' perceptions of the system, usability of AI-generated content, opinions on the hints and feedback provided at each stage of the system for improving their learning outcomes. Throughout the entire process, we recorded both the screen and audio for subsequent analysis.

5.2.4 Measurements. For **RQ1**, inspired by Wu et al. [60], we examined participants' verbal reflection of what they learned from the video about the instructional strategies after the watching stage. We coded the number of *valid inspiration points* mentioned in their oral reports. In the specific task, a valid point had to describe at least one concrete implementation element of the strategy in the video, such as its timing, task design, or teacher behavior; merely restating the strategy definition was not counted. We treated sub-sentence phrasal segments as the basic coding unit. Furthermore, we invited a mathematics teacher (Male, 9 years experience) and a Chinese teacher (Female, 32 years experience) to evaluate the recorded videos of participants' teaching tests about their respective subjects. The evaluation process was fully blind, with raters unaware of the experimental condition for each video. Based on the literature on microteaching [20, 55] and the scenario of VBL, we established the following four *metrics of teaching performance* on 7-point Likert Scale (1/7 – worst/best performance): a) instructional competence, b) application of teaching strategies learned from videos (*e.g.,* choosing the right moment and appropriate way to introduce a strategy), c) imitation of concrete implementations of the demonstrated strategies (*e.g.,* reproducing the teacher's gestures or phrasing), and d) the authenticity of the scenario.

For **RQ2**, we adapted measures from [27, 33, 43] (Appendix C) and adopted 7-point Likert Scale (1/7 – strongly disagree/agree) to measure the following items after each learning session: Q1) confidence in teaching, Q2) gained insights, Q3) benefits of practicing sessions, Q4) engagement in the learning process, Q5) ease of use, and Q6) intention to use.

Table 1: Quantitative comparison between *TeachUp* and Baseline. Q1–Q6 are 7-point Likert items; Valid Inspiration Points are raw counts. Expert-rated scores (Dim. 1–4) are Z-score standardized across raters. +: $p < .1$, *: $p < .05$, **: $p < .01$.

RQ	Item	TeachUp Mean (SD)	Baseline Mean (SD)	p	Sig.
RQ1	Valid Inspiration Points	2.56 (1.09)	1.81 (1.11)	.049	*
	Dim1: Instructional competence	+0.27 (0.99)	−0.27 (1.00)	.065	+
	Dim2: Strategy application	+0.32 (0.95)	−0.32 (1.01)	.003	**
	Dim3: Imitation of implementations	+0.09 (1.22)	−0.09 (0.79)	.516	
	Dim4: Authenticity of scenario	+0.24 (1.13)	−0.24 (0.86)	.086	+
RQ2	Q1: Confidence in teaching	5.69 (0.79)	5.00 (1.21)	.009	**
	Q2: Gained insights	5.88 (1.02)	5.75 (0.86)	.527	
	Q3: Benefits of practicing sessions	6.25 (0.58)	5.63 (0.72)	.014	*
	Q4: Engagement in the learning process	6.13 (0.50)	5.31 (1.20)	.026	*
	Q5: Ease of use	6.06 (0.77)	5.63 (0.62)	.053	+
	Q6: Intention to use	6.13 (0.50)	5.25 (1.18)	.017	*

6 Analyses and Results

In this section, we present the quantitative and qualitative results for each research question. Table 1 summarizes all quantitative comparisons between *TeachUp* and the baseline. For the numbers of valid inspiration points (RQ1) and items about usability and usefulness (RQ2), we employ the Wilcoxon signed-rank tests to compare the differences between the *TeachUp* and baseline condition. As for the expert evaluation of each participant's performance in the test (RQ1) after using *TeachUp* and the baseline, we first apply Z-score standardization to mitigate inconsistencies in scoring standards among different raters. To compare the conditions, we then assessed the normality of the score differences using the Shapiro-Wilk test. A paired t-test was used for normally distributed data; otherwise, the Wilcoxon signed-rank test was applied. For the qualitative data, two authors applied Braun and Clarke's six-phase thematic analysis framework [18] using a deductive approach, with predefined themes of effectiveness (RQ1), and usability & usefulness (RQ2).

6.1 RQ1: Effectiveness of *TeachUp*

As shown in Table 1, participants identified significantly more valid inspiration points—that is, concrete implementation elements of the focal strategy—with *TeachUp* ($M = 2.56$, $SD = 1.09$) than with the baseline ($M = 1.81$, $SD = 1.11$; $W = 18.00$, $p = .049$). **Expert-rated performance** revealed that *TeachUp* significantly improved participants' performance on applying the learned instructional strategies from the videos to a new teaching task ($p = .003$), with marginal improvements in instructional competence ($p = .065$) and scenario authenticity ($p = .086$). No significant difference was found in imitation of implementations of the strategies on the new teaching task ($p = .516$).

Participants shared how *TeachUp* supported the effectiveness of their learning. In the video-watching stage, teachers regarded the AI-generated hints as a form of reflection for identifying and addressing gaps. Some participants (P3, P11, P5, P6, P8) preferred to watch the entire video before answering the questions, and they expressed that these hints helped them recall details they had missed, as P11 said, *"There are indeed some teaching methods, or some teaching steps, that I overlooked. I think AI in this respect can really serve as a good reminder."* In the scenario-based practicing stage, almost all users agreed that the tasks were helpful and guided them to improve their performance (except P11). For example, participants shared comments such as, *"It gave me a task to think about my own wording"* (P14); *"The coherence phrases I learned in the first stage guided me to apply them here"* (P16); *"It (the task) helped me grasp the pace of teaching"* (P15). Overall, participants expressed that the AI-generated tasks not only improved their performance from different aspects (*e.g.,* language, pacing) but also encouraged deeper reflection on how teaching methods could be approached and refined.

6.2 RQ2: Usability and Usefulness of *TeachUp*

As shown in Table 1, compared to the baseline system, participants with *TeachUp* reported significantly higher confidence in classroom teaching after the learning session ($p = .009$). There is no significant difference between the two conditions regarding perceived gained insights from the video cases that participants can directly apply to improve their teaching practices (p = .527). Nevertheless, participants with *TeachUp* found that its practice sessions were significantly more beneficial for applying the teaching skills learned from the videos in classroom settings (p = .014). Overall, participants perceived that *TeachUp* is generally easier to use (p = .053), reported significantly higher intention to use *TeachUp* for video-based learning and teaching reflection (p = .017), and were more interested and engaged in the training process with *TeachUp* (p = .026), compared to the baseline system.

All participants appreciated *TeachUp*'s reflective support in the structured three-stage learning process. In the watching stage, most participants found the reflective questions helpful for learning from the videos (except P2, P3, and P15). These hints were seen as a form of guidance, encouraging users to pay attention to the micro-level of teacher's actions in the videos, sometimes even just from a glimpse. As P9 stated: *"The analysis is very clear, for example explaining step by step how things continue, which lets me understand how the teacher interacts with the students step by step."* All participants except P11 and P15 also found that the tasks in the practicing stage were helpful. They emphasized the importance of maintaining their own subjectivity rather than strictly following the tasks, as P15 remarked, *"When connecting to these tasks it gave, sometimes I would suddenly get stuck; in contrast, (without these tasks), I know what I need to do and which skills I can use to make the teaching smoother and more complete."* We recognize this as a positive phenomenon, as it reflects the teachers' reflective engagement during the practice process. Lastly, participants praised the precision of the system's evaluation of their practices. P8 remarked, *"Its analyses were very accurate; it truly pointed out what I failed to accomplish."* Similarly, P14 appreciated that *"It gave me a specific positive example to learn from."* Together, these comments illustrate how the evaluation supported reflection by identifying unmet practice goals and providing a concrete reference for improvement.

Apart from the benefits, participants also reported room for improving *TeachUp*. For example, several participants mentioned that having too many tasks in the practicing stage limited *TeachUp*'s usefulness. Three users (P10, P13, P8) pointed out that excessively lengthy generated outputs can detract from the user experience, as P13 noted, *"The language could be more concise. Sometimes just providing a few keywords or key points is enough to understand, without displaying so much."* P3 further noted that in the evaluative report, some spatial or gestural details about his practice video were missed.

7 Interviews with Teachers

Following the within-subjects study, we conducted an interview study with four in-service teachers who used *TeachUp* on their subjects of interest and video clips to enhance the generalizability of our findings. Two of them are early-stage teachers with one year of teaching experience – one teaches science in an elementary school (N1, male), and the other teaches English in a middle school (N2, female). The remaining two are experienced teachers, including one Chinese teacher with 32 years of experience in an elementary school (E1, female) and one mathematics teacher (E2, male) with 29 years of experience in a high school.

The day before the interview, we asked each participant to provide a topic within their subject with which they were highly familiar. We then created a video resource library for each topic by downloading five recorded classroom videos from Bilibili, segmenting them, detecting the instructional strategies using our pipeline, and storing them in the database. Each interview lasted 60 minutes and began with a brief introduction to *TeachUp* and the nine teaching strategies, followed by a 10-minute system tutorial. Participants then freely explored the system for 50 minutes using a think-aloud protocol, completing at least two instructional cases and the full watch–practice–evaluate cycle. Brief contextual interviews were conducted at each stage transition to capture immediate feedback. The session concluded with questions on the potential of *TeachUp* to support teachers' professional development. We analyzed the recorded screens of system interactions and audio of interviews and reported the results below.

The strategy detection pipeline is useful. Because in-service teachers could freely explore tens of video clips segmented by the proposed pipeline, we mainly focus on the evaluation of the pipeline's usefulness in this interview study. The teachers agree that the video segmentation had high precision. E2 further praised the diversity across clips: *"Although the clips covered the same teaching point, they unfolded from many different angles, so I could find the clips that best fit my own teaching situation."*

The reflective support is helpful in the whole learning session. Experts generally praised our "watch-practice-evaluate" framework, affirming that it could enhance teachers' general competencies within a single session. E1 emphasized that benefits extend beyond verbal expression to classroom organization, discipline, and the habit of self-reflection: *"It can also cultivate the habit of reflecting on whether one's classroom has met its objectives."* E2 similarly stressed that reflective thinking needs to be verified through practice to take effect, and valued the platform's ability to support such simulation at any time.

Experienced teachers hold different views on *TeachUp*'s long-term effectiveness. Regarding the long-term effectiveness of early-stage teacher professional development, the two experienced teachers offered differing views. E2 highlighted that reflective hints could support teachers in developing a more comprehensive understanding of their classrooms: *"In everyday teaching, there may be some fixed mindsets... By learning through the system, thinking can be expanded, which is very meaningful."* In contrast, E1 questioned the system's impact on sustained competence development, especially concerning knowledge transfer: *"There may be some shortcomings in transfer, because the whole evaluation is based on comparing with the original video. I think teachers may need to watch the corresponding videos for all their classes."*

The practicing tasks are inflexible and could be risky. In-service teachers indicated that fixed tasks made the training less flexible and posed risks for applying the learned instructional strategies in the real classroom. First, they pointed out that overly detailed task descriptions might lead to the development of inappropriate classroom habits. E2 noted, *"if each step is too detailed, it may affect the efficiency of classroom teaching."* Second, they raised concerns about microteaching without students, relying solely on 'imagined' students and imitation of the teachers in the videos carries certain risks. N2 worried: *"If the students' foundation is weak, some of the steps I just practiced may not be feasible in reality."* To address this issue, E1 emphasized the importance of simulating student feedback: *"I think student feedback is very important. It should include not only these pre-generated things, but also the feedback from students on what you just taught. So I suggest adding simulated student feedback in the future."* This indicates the need for additional features to provide teachers with a more authentic and customizable environment, enriched with synchronous feedback.

The generated content can be further improved. Similar to the findings in the user study, E1, N1, and N2 noted that excessively lengthy generated content can detract from the user experience. N2 suggested, *"Maybe keywords could be bolded or specially highlighted, so users can quickly and clearly identify them at a glance."* E1 commented on the reflective hints, based on her familiarity with the selected Chinese class: *"Because the video is short, one task may already be enough... The questions are quite repetitive, meaning the same question keeps being asked again and again."* Besides, N1 noted that some of his actions might not have been fully analyzed in the evaluative report.

8 Discussion

8.1 Empowering Scalable Video-Based Learning for Teachers' Professional Development

As an enabling component of *TeachUp*, we implemented an LLM-powered pipeline for automatically segmenting classroom video clips and detecting the teachers' instructional strategies. On our benchmark with 89 labeled video clips, our pipeline reaches a precision of 0.634 for detecting nine strategies, and the in-service teachers in our interview study found the resulting clips usable for their own learning. This component addresses the design challenges of video-based learning (VBL) platforms identified by Bates et al. [7], who found that merely providing a large video library can overwhelm teachers; instead, teachers prefer short, practical segments that focus on specific instructional strategies and are "ready-to-use". Our pipeline directly meets this need by automatically extracting such content, thereby saving teachers' time in searching for valuable learning materials online [17]. Given the pipeline's preliminary performance, teachers could first screen detected clips by ranking them according to confidence or adjust the confidence threshold to balance reliability and breadth. Concise clip summaries and evidence linked to precise video moments could then help them quickly inspect why a strategy was suggested and decide whether the clip meets their learning needs. Besides, as highlighted by an experienced teacher in the interview, the diversity of video segments extracted by our pipeline demonstrates its potential for wide-ranging applications as a standalone tool. We designed this pipeline as a reusable, "plug-and-play middleware" that can be deployed between any video resource database and online learning platform. Its generic input–output interface—taking a classroom video and producing timestamped strategy clips with category metadata—enables integration into diverse educational systems without platform-specific modification. This decoupled design means that our technical contribution holds independent value and can enhance numerous existing online teacher education platforms.

8.2 "Watch-Practice-Evaluate": A Kolb-Inspired Framework for Reflective VBL

We developed *TeachUp*, an interactive system that supports a three-stage *"watching-practicing-evaluating"* cycle for VBL, based on our formative study with both early-stage and experienced teachers. This cycle can be viewed as a concrete implementation of Kolb's Reflective Cycle [29], tailored to the VBL context. Specifically, observing a teacher's well-performed instructional case serves as *"Concrete Experience"*, acting as a substitute for early-stage teachers' limited teaching experience. The reflective hints in the watching stage implement *"Reflective Observation"* and *"Abstract Conceptualization"*, encouraging users to link observations with personal experiences and transform them into new ideas. Finally, the "microteaching" practice session realizes *"Active Experimentation"*, where users apply newly acquired concepts and receive immediate, personalized feedback, completing the cycle. In our user study, *TeachUp* improved participants' ability to envision authentic teaching scenarios and enhanced their overall performance, suggesting that this cycle effectively enhances early-stage teachers' pedagogical knowledge. We thus present both an effective implementation framework and empirical evidence to guide the design of reflective VBL systems.

8.3 Implications for HCI Research

Our findings lead to three implications for HCI research.

8.3.1 *AI-powered learning tools should bridge observation and action.*

The results show that *TeachUp* led to significantly more implementation elements of the focal strategy than the baseline. Also, the qualitative results indicate that the watching-stage interactions play a grounding role in the integrated framework, providing strategy anchors for later practice and comparison-based evaluation. While many existing tools focus on helping users better observe and annotate video content [42, 46], our findings suggest that these observations should be connected to enactment through contextualized practice, structured rehearsal, and personalized feedback grounded in the watched content. This perspective may generalize to other domains where people learn by watching experts' demonstrations, such as medical training and public speaking.

8.3.2 *Using LLMs and multimodal large models to transform passive video resources into active learning environments.*

TeachUp demonstrates that large models can be prompted to extract pedagogical structure from videos and generate contextualized practice and evaluation around them, turning static recordings into interactive learning cycles. This extends emerging work on in-context video assessments [39] and interactive notes [70] by enabling multimodal feedback grounded in video content. However, our findings indicate that generated content can sometimes be verbose or insufficiently attuned to nonverbal and temporal nuances in teaching videos. To address this, future systems could employ video generation models to synthesize diverse, fine-grained video variants for richer long-term rehearsal [62], and leverage more specialized video-oriented large models capable of capturing subtle nonverbal cues (*e.g.,* gestures, posture) and cross-time-span pedagogical features [25].

8.3.3 Connecting practicing tasks during VBL with real-world teaching scenarios. In the practice stage, *TeachUp* generates teaching scenarios and tasks based on the watched classroom video clip and user-input practice goal. While participants in the user study generally appreciated the practicing tasks and maintained their own subjectivity during practices, teachers in the interviews were concerned that the tasks were inflexible and risky. These concerns reveal a transfer-distance tension in connecting video-based practice with real-world teaching. If a practice task is too close to the original video context, users may reproduce the case rather than develop the ability to apply the underlying strategy flexibly; conversely, if the context is too different, early-stage teachers may struggle to transfer the newly learned strategy. Future work could address this tension by creating more realistic and adaptive classroom environments for practicing strategies learned from videos. For example, future work could code and model the students' behaviors in classroom videos to simulate student agents with varied profiles and responses [13, 43, 69]. The VBL support tool can therefore offer a virtual classroom for teaching practices [26], in which student agents can interact with teachers by raising hands, speaking, eye contact, and so on. Future systems could further adapt how far each practice scenario departs from the watched video according to users' demonstrated understanding and performance.

8.4 Limitation and Future Work

This study has several limitations. First, the participants in the within-subjects controlled lab study were not actual elementary Chinese or Math teachers but were prompted to act as them. This mismatch with specialized teaching subjects may impair their performance in the tests after the learning session. While our interview with in-service teachers helps generalize our findings, future research should involve more early-stage teachers using *TeachUp* within their specialized subjects to evaluate its effectiveness and usability. Second, our study lacks an evaluation of user behavior (*e.g.,* whether their teaching behaviors changed after the training) and results (*e.g.,* whether their students demonstrated significant improvement) in their following teaching practices, which are important in Kirkpatrick's four levels of evaluation [27]. Future case studies could be conducted to comprehensively assess the impact of *TeachUp* on teachers' professional development in the long term. Third, the current baseline is not a component-level ablation, so our study evaluates the integrated reflective loop as a whole and does not isolate the effectiveness of individual interaction components. Fourth, our strategy detection pipeline was developed and validated on a small benchmark of 89 video scripts. The controlled study used manually annotated ground-truth clips, and the in-service teacher interviews did not identify any incorrectly tagged clips. Consequently, our current studies do not establish how errors in the pipeline would affect early-stage teachers' learning. Alongside examining the effects of such errors on early-stage teachers' learning, future work can use our benchmark and pipeline as starting points to iteratively improve them by using the pipeline to give initial labels to more video scripts and incorporating more multimodal video features.

9 Conclusion

In this paper, we designed and developed an interactive system, *TeachUp*, to support teachers to learn instructional strategies from videos of others' well-executed classroom lectures. We implemented a computational pipeline powered by large language models to detect nine instructional strategies enacted by teachers in the short video clips segmented from the classroom videos. With these video clips, *TeachUp* provides reflective hints and adaptive microteaching practices to engage teachers in a "watch–practice–evaluate" video-based learning cycle. Our within-subjects user study with 16 early-stage teachers and four interviews with in-service teachers demonstrated that *TeachUp* can help teachers effectively learn instructional strategies and improve their instructional competencies. With these findings, our computational pipeline, and developed system, we offer practical implications to future video-based learning systems for teachers' professional development.

Acknowledgments

This work is supported by Guangdong S&T Programme (2025B010 1120004) and General Projects Fund of the Natural Science Foundation of Guangdong Province in China grant (2024A1515012226).

References

[1] D.W. Allen and K. Ryan. 1969. *Microteaching.* Addison-Wesley Publishing Company. https://books.google.com.sg/books?id=iL1RAQAAIAAJ
[2] Abdurrahman Ghaleb Almekhlafi, Sadiq Abdulwahed Ismail, and Abdelmoniem Ahmed Hassan. 2020. Teachers' Reported Use of Marzano's Instructional Strategies in United Arab Emirates K-12 Schools. *International Journal of Instruction* 13, 1 (2020), 325–340.
[3] Julie M Amador, Jode Keehr, Abraham Wallin, and Christopher Chilton. 2020. Video complexity: Describing videos used for teacher learning. *Eurasia Journal of Mathematics, Science and Technology Education* 16, 4 (2020), em1834.
[4] Riku Arakawa and Hiromu Yakura. 2020. INWARD: A computer-supported tool for video-reflection improves efficiency and effectiveness in executive coaching. In *Proceedings of the 2020 CHI conference on human factors in computing systems.* 1–13.
[5] Riku Arakawa, Hiromu Yakura, and Masataka Goto. 2023. CatAlyst: domain-extensible intervention for preventing task procrastination using large generative models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems.* 1–19.
[6] Laura Baecher, Shiao-Chuan Kung, Sarah Laleman Ward, and Kimberly Kern. 2018. Facilitating video analysis for teacher development: A systematic review of the research. *Journal of Technology and Teacher Education* 26, 2 (2018), 185–216.
[7] Meg Schleppenbach Bates, Lena Phalen, and Cheryl G Moran. 2016. If you build it, will they reflect? Examining teachers' use of an online video-based learning website. *Teaching and teacher education* 58 (2016), 17–27.
[8] Marit Bentvelzen, Paweł Woźniak, Pia SF Herbes, Evropi Stefanidi, and Jasmin Niess. 2022. Revisiting reflection in hci: Four design resources for technologies that support reflection. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 1 (2022), 1–27.
[9] Amanda Berry, Fien Depaepe, and Jan Van Driel. 2016. Pedagogical content knowledge in teacher education. In *International handbook of teacher education: volume 1.* Springer, 347–386.
[10] Cynthia E Bolt-Lee. 2021. Developments in Research-Based Instructional Strategies: Learning-Centered Approaches for Accounting Education. *E-Journal of Business Education and Scholarship of Teaching* 15, 2 (2021), 1–14.
[11] David Boud, Rosemary Keogh, and David Walker. 2013. *Reflection: Turning experience into learning.* Routledge.
[12] Evelyn M Boyd and Ann W Fales. 1983. Reflective learning: Key to learning from experience. *Journal of humanistic psychology* 23, 2 (1983), 99–117.
[13] Jiwon Chun, Yuling Zhuang, Armanto Sutedjo, Colin Xu, Rong Ren, and Meng Xia. 2026. ArguMath: AI-Simulated Environment for Pre-service Teacher Training in Orchestrating Classroom Mathematics Argumentation. In *International Conference on Artificial Intelligence in Education.* Springer, 1–10.
[14] Vanessa Echeverria, Lixiang Yan, Linxuan Zhao, Sophie Abel, Riordan Alfredo, Samantha Dix, Hollie Jaggard, Rosie Wotherspoon, Abra Osborne, Simon Buckingham Shum, et al. 2024. TeamSlides: A multimodal teamwork analytics dashboard for teacher-guided reflection in a physical learning space. In *Proceedings of the 14th learning analytics and knowledge conference.* 112–122.

[15] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision* 88, 2 (2010), 303–338.

[16] Rowanne Fleck and Geraldine Fitzpatrick. 2010. Reflecting on reflection: framing a design landscape. In *Proceedings of the 22nd conference of the computer-human interaction special interest group of australia on computer-human interaction*. 216–223.

[17] Cyrille Gaudin and Sébastien Chaliès. 2015. Video viewing in teacher education and professional development: A literature review. *Educational research review* 16 (2015), 41–67.

[18] Greg Guest, Kathleen M MacQueen, and Emily E Namey. 2011. *Applied thematic analysis*. sage publications.

[19] Björn Haßler, Sara Hennessy, and Riikka Hofmann. 2020. OER4Schools: Outcomes of a sustained professional development intervention in sub-Saharan Africa. In *Frontiers in Education*, Vol. 5. Frontiers Media SA, 146.

[20] Chuanjun He and Chunmei Yan. 2011. Exploring authenticity of microteaching in pre-service teacher education programmes. *Teaching education* 22, 3 (2011), 291–302.

[21] Brad Hokanson and Simon Hooper. 2004. Levels of teaching: A taxonomy for instructional design. *Educational technology* 44, 6 (2004), 14–22.

[22] Mohammed Hoque, Matthieu Courgeon, Jean-Claude Martin, Bilge Mutlu, and Rosalind W Picard. 2013. Mach: My automated conversation coach. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*. 697–706.

[23] Juho Kim, Phu Tran Nguyen, Sarah Weir, Philip J Guo, Robert C Miller, and Krzysztof Z Gajos. 2014. Crowdsourcing step-by-step information extraction to enhance existing how-to videos. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 4017–4026.

[24] Taewan Kim, Seolyeong Bae, Hyun Ah Kim, Su-woo Lee, Hwajung Hong, Chanmo Yang, and Young-Ho Kim. 2024. MindfulDiary: Harnessing large language model to support psychiatric patients' journaling. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–20.

[25] Youngmin Kim, Jiwan Chung, Jisoo Kim, Sunghyun Lee, Sangkyu Lee, Junhyeok Kim, Cheoljong Yang, and Youngjae Yu. 2025. Speaking Beyond Language: A Large-Scale Multimodal Dataset for Learning Nonverbal Cues from Video-Grounded Dialogues. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2247–2265.

[26] Seth King, Joseph Boyer, Tyler Bell, and Anne Estapa. 2022. An automated virtual reality training system for teacher-student interaction: A randomized controlled trial. *JMIR serious games* 10, 4 (2022), e41097.

[27] Donald Kirkpatrick and James Kirkpatrick. 2006. *Evaluating training programs: The four levels*. Berrett-Koehler Publishers.

[28] Rafal Kocielnik, Lillian Xiao, Daniel Avrahami, and Gary Hsieh. 2018. Reflection companion: a conversational system for engaging users in reflection on physical activity. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 2 (2018), 1–26.

[29] David A Kolb. 2014. *Experiential learning: Experience as the source of learning and development*. FT press.

[30] Serge Leblanc. 2018. Analysis of video-based training approaches and professional development. *Contemporary Issues in Technology and Teacher Education* 18, 1 (2018), 125–148.

[31] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437* (2024).

[32] Jingyuan Liu, Nazmus Saquib, Chen Zhutian, Rubaiat Habib Kazi, Li-Yi Wei, Hongbo Fu, and Chiew-Lan Tai. 2022. Posecoach: A customizable analysis and visualization system for video-based running coaching. *IEEE Transactions on Visualization and Computer Graphics* 30, 7 (2022), 3180–3195.

[33] Arnold M Lund. 2001. Measuring usability with the use questionnaire12. *Usability interface* 8, 2 (2001), 3–6.

[34] Shuai Ma, Taichang Zhou, Fei Nie, and Xiaojuan Ma. 2022. Glancee: An adaptable system for instructors to grasp student learning status in synchronous online classes. In *Proceedings of the 2022 CHI conference on human factors in computing systems*. 1–25.

[35] Ahmed Magooda, Diane Litman, Ahmed Ashraf, and Muhsin Menekse. 2022. Improving the quality of students' written reflections using natural language processing: Model design and classroom evaluation. In *International conference on artificial intelligence in education*. Springer, 519–525.

[36] Robert J Marzano. 2003. *What works in schools: Translating research into action*. Ascd.

[37] Robert J Marzano, Jennifer S Norford, Diane E Paynter, Debra J Pickering, and Barbara B Gaddy. 2001. *A Handbook for Classroom Instruction That Works*. ERIC.

[38] Robert J Marzano, Debra Pickering, and Jane E Pollock. 2001. *Classroom instruction that works: Research-based strategies for increasing student achievement*. Ascd.

[39] Toni-Jan Keith Palma Monserrat, Yawen Li, Shengdong Zhao, and Xiang Cao. 2014. L. IVE: an integrated interactive video-based learning environment. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 3399–3402.

[40] Tahmina Nazari, Floyd W van de Graaf, Mary EW Dankbaar, Johan F Lange, Jeroen JG van Merriënboer, and Theo Wiggers. 2020. One step at a time: step by step versus continuous video-based learning to prepare medical students for performing surgical procedures. *Journal of surgical education* 77, 4 (2020), 779–787.

[41] Seyed Parsa Neshaei, Thiemo Wambsganss, Hind El Bouchrifi, and Tanja Käser. 2025. MindMate: Exploring the Effect of Conversational Agents on Reflective Writing. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–9.

[42] Tricia J Ngoon, S Sushil, Angela EB Stewart, Ung-Sang Lee, Saranya Venkatraman, Neil Thawani, Prasenjit Mitra, Sherice Clarke, John Zimmerman, and Amy Ogan. 2024. ClassInSight: Designing Conversation Support Tools to Visualize Classroom Discussion for Personalized Teacher Professional Development. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–15.

[43] Sitong Pan, Robin Schmucker, Bernardo Garcia Bulle Bueno, Salome Aguilar Llanes, Fernanda Albo Alarcón, Hangxiao Zhu, Adam Teo, and Meng Xia. 2025. Tutorup: What if your students were simulated? training tutors to address engagement challenges in online learning. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–18.

[44] Howard Pitler, Elizabeth R Hubbell, and Matt Kuhn. 2012. *Using technology with classroom instruction that works*. Ascd.

[45] Ambili Remesh. 2013. Microteaching, an efficient technique for learning effective teaching. *Journal of research in medical sciences: the official journal of Isfahan University of Medical Sciences* 18, 2 (2013), 158.

[46] Peter J Rich and Michael Hannafin. 2009. Video annotation tools: Technologies to scaffold, structure, and transform teacher reflection. *Journal of teacher education* 60, 1 (2009), 52–67.

[47] Marija Sablić, Ana Mirosavljević, and Alma Škugor. 2021. Video-based learning (VBL)—past, present and future: An overview of the research published from 2008 to 2019. *Technology, Knowledge and Learning* 26, 4 (2021), 1061–1077.

[48] Alpay Sabuncuoglu and T Metin Sezgin. 2023. Developing a multimodal classroom engagement analysis dashboard for higher-education. *Proceedings of the ACM on Human-Computer Interaction* 7, EICS (2023), 1–23.

[49] Lee S Shulman. 1986. Those who understand: Knowledge growth in teaching. *Educational researcher* 15, 2 (1986), 4–14.

[50] Christina Siry and Sonya N Martin. 2014. Facilitating reflexivity in preservice science teacher education using video analysis and cogenerative dialogue in field-based methods courses. *EURASIA journal of mathematics, science and technology education* 10, 5 (2014), 481–508.

[51] Winnie Wing-mui So. 2012. Quality of learning outcomes in an online video-based learning community: Potential and challenges for student teachers. *Asia-Pacific Journal of Teacher Education* 40, 2 (2012), 143–158.

[52] Rand J Spiro, Brian P Collins, and Aparna Ramchandran. 2014. Reflections on a Post-Gutenerg Epistemology for Video Use in Ill-Structured Domains: Fostering Complex Learning and Cognitive Flexibility. In *Video research in the learning sciences*. Routledge, 93–100.

[53] James W Stigler and James Hiebert. 2009. *The teaching gap: Best ideas from the world's teachers for improving education in the classroom*. Simon and Schuster.

[54] Joseph A Taylor, Kathleen Roth, Christopher D Wilson, Molly AM Stuhlsatz, and Elizabeth Tipton. 2017. The effect of an analysis-of-practice, videocase-based, teacher professional development program on elementary students' science achievement. *Journal of Research on Educational Effectiveness* 10, 2 (2017), 241–271.

[55] Pugazhenthan Thangaraju and Bikash Medhi. 2023. Microteaching: Overview and examination evaluation. *Indian Journal of Pharmacology* 55, 4 (2023), 257–262.

[56] Meredith Thompson, Kesiena Owho-Ovuakporie, Kevin Robinson, Yoon Jeon Kim, Rachel Slama, and Justin Reich. 2019. Teacher Moments: A digital simulation for preservice teachers to approximate parent–teacher conversations. *Journal of Digital Learning in Teacher Education* 35, 3 (2019), 144–164.

[57] Isabell Tucholka and Bernadette Gold. 2025. Analysing classroom videos in teacher education—How different instructional settings promote student teachers' professional vision of classroom management. *Learning and Instruction* 97 (2025), 102084.

[58] Xu Wang, Meredith Thompson, Kexin Yang, Dan Roy, Kenneth R Koedinger, Carolyn P Rose, and Justin Reich. 2021. Practice-based teacher questioning strategy training with ELK: A role-playing simulation for eliciting learner knowledge. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–27.

[59] Xingbo Wang, Haipeng Zeng, Yong Wang, Aoyu Wu, Zhida Sun, Xiaojuan Ma, and Huamin Qu. 2020. Voicecoach: Interactive evidence-based training for voice modulation skills in public speaking. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.

[60] Shiwei Wu, Mingxiang Wang, Chuhan Shi, and Zhenhui Peng. 2025. ComViewer: An Interactive Visual Tool to Help Viewers Seek Social Support in Online Mental Health Communities. *Proceedings of the ACM on Human-Computer Interaction* 9, 2 (2025), 1–31.

[61] Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, et al. 2025. Qwen2. 5-omni technical report.

arXiv preprint arXiv:2503.20215 (2025).

[62] Tao Xu, Yuan Liu, Yaru Jin, Yueyao Qu, Jie Bai, Wenlan Zhang, and Yun Zhou. 2025. From recorded to AI-generated instructional videos: A comparison of learning performance and experience. *British Journal of Educational Technology* 56, 4 (2025), 1463–1487.

[63] Xiaotong Xu, Jiayu Yin, Catherine Gu, Jenny Mar, Sydney Zhang, Jane L E, and Steven P Dow. 2024. Jamplate: Exploring llm-enhanced templates for idea reflection. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*. 907–921.

[64] 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 (2020), 3168–3181.

[65] Chao Zhang, Kexin Ju, Peter Bidoshi, Yu-Chun Grace Yen, and Jeffrey M Rzeszotarski. 2025. Friction: Deciphering Writing Feedback into Writing Revisions through LLM-Assisted Reflection. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–27.

[66] Gefei Zhang, Shenming Ji, Yicao Li, Jingwei Tang, Jihong Ding, Meng Xia, Guodao Sun, and Ronghua Liang. 2025. CPVis: Evidence-based Multimodal Learning Analytics for Evaluation in Collaborative Programming. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–26.

[67] Meilan Zhang, Mary Lundeberg, Matthew J Koehler, and Jan Eberhardt. 2011. Understanding affordances and challenges of three types of video for teacher professional development. *Teaching and teacher education* 27, 2 (2011), 454–462.

[68] Meilan Zhang, Mary Lundeberg, Tom J McConnell, Matthew J Koehler, and Jan Eberhardt. 2010. Using questioning to facilitate discussion of science teaching problems in teacher professional development. *Interdisciplinary Journal of Problem-Based Learning* 4, 1 (2010), 5.

[69] Zheyuan Zhang, Daniel Zhang-Li, Jifan Yu, Linlu Gong, Jinchang Zhou, Zhanxin Hao, Jianxiao Jiang, Jie Cao, Huiqin Liu, Zhiyuan Liu, et al. 2025. Simulating classroom education with llm-empowered agents. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 10364–10379.

[70] Running Zhao, Zhihan Jiang, Xinchen Zhang, Chirui Chang, Handi Chen, Weipeng Deng, Luyao Jin, Xiaojuan Qi, Xun Qian, and Edith CH Ngai. 2025. NoteIt: A System Converting Instructional Videos to Interactable Notes Through Multimodal Video Understanding. *arXiv preprint arXiv:2508.14395* (2025).

[71] Hangyu Zhou, Yuichiro Fujimoto, Masayuki Kanbara, and Hirokazu Kato. 2021. Virtual reality as a reflection technique for public speaking training. *Applied Sciences* 11, 9 (2021), 3988.

A Marzano's Nine Events

Table 2 shows the definitions and the examples of the nine broad, high-probability categories of teaching strategies used in classroom.

Table 2: Instructional strategies and examples of implementation in the classroom

Strategy	Definition	Examples of Implementation
(1) Identifying Similarities and Differences	Having students compare, contrast, and classify concepts.	- Use Venn diagrams to compare and classify items. - Ask students to identify similarities and differences on their own, followed by group discussion.
(2) Summarizing and Note Taking	Having students identify key information and restate it concisely in their own words.	- Summarize concepts that have already been taught. - Encourage students to use a consistent format for note taking.
(3) Reinforcing Effort and Providing Recognition	Connecting student effort to achievement and providing recognition for their work.	- Assign timed quizzes and have students report on their performance (*e.g.,* accuracy). - Display exemplary student work in the classroom to recognize their effort.
(4) Homework and Practice	Using assignments and practice to reinforce concepts and skills learned in class.	- Set aside class time for students to practice. - Assign and review homework.
(5) Nonlinguistic Representations	Representing and exploring concepts using visuals, physical models, and movement.	- Use drawings, videos, physical models, or role-playing to demonstrate concepts.
(6) Cooperative Learning	Organizing students into groups to work collaboratively on academic tasks.	- Group students based on interests or common experiences and assign specific roles (*e.g.,* recorder, reporter). - Implement activities such as group presentations.
(7) Setting Objectives and Providing Feedback	Setting clear learning goals and providing students with timely, specific feedback.	- Continuously display unit goals on a whiteboard or in slides, and track progress to provide ongoing feedback.
(8) Generating and Testing Hypotheses	Engaging students in proposing hypotheses and then designing tasks to test them.	- Ask students to predict what would happen if an aspect of a system were changed. - Organize groups to complete a hands-on task and discuss the reasons for their success or failure.
(9) Cues, Questions, and Advance Organizers	Using cues, questions, and organizers to activate prior knowledge before introducing new concepts.	- Before a new lesson, present an outline, graphic image, or mind map as an advance organizer. - Pose key questions and pause after asking to encourage deeper thinking before students answer.

B The Slide Deck Used in the Formative Study

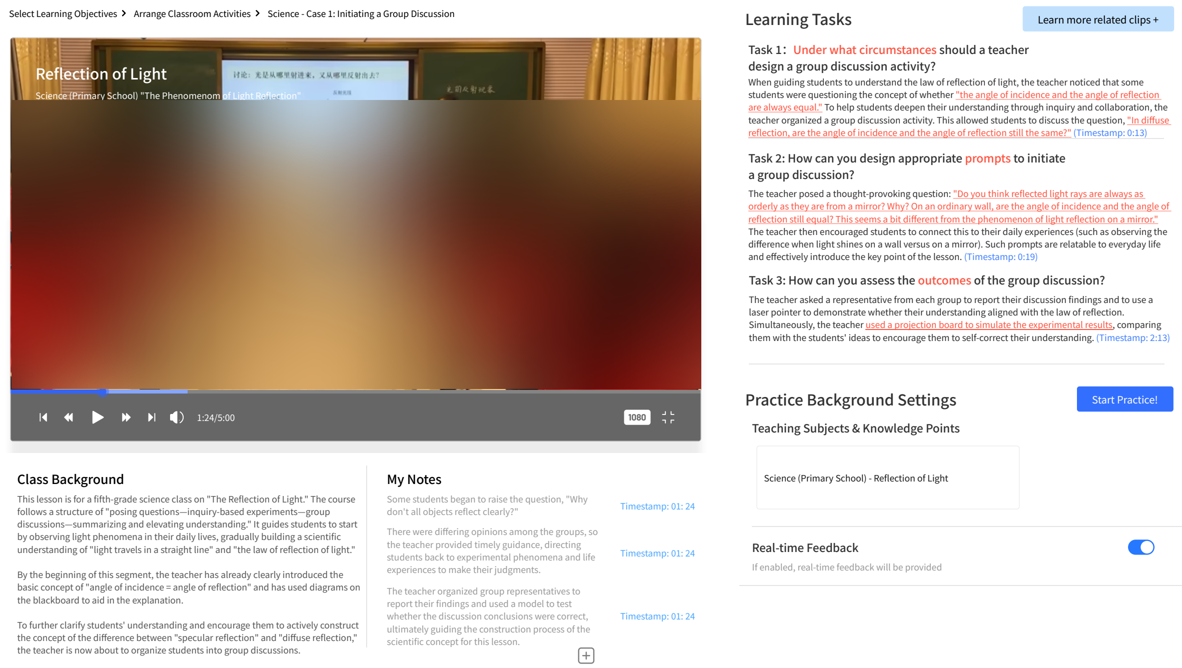

Figure 5: The video-watching page of the prototype. Translated from Chinese.

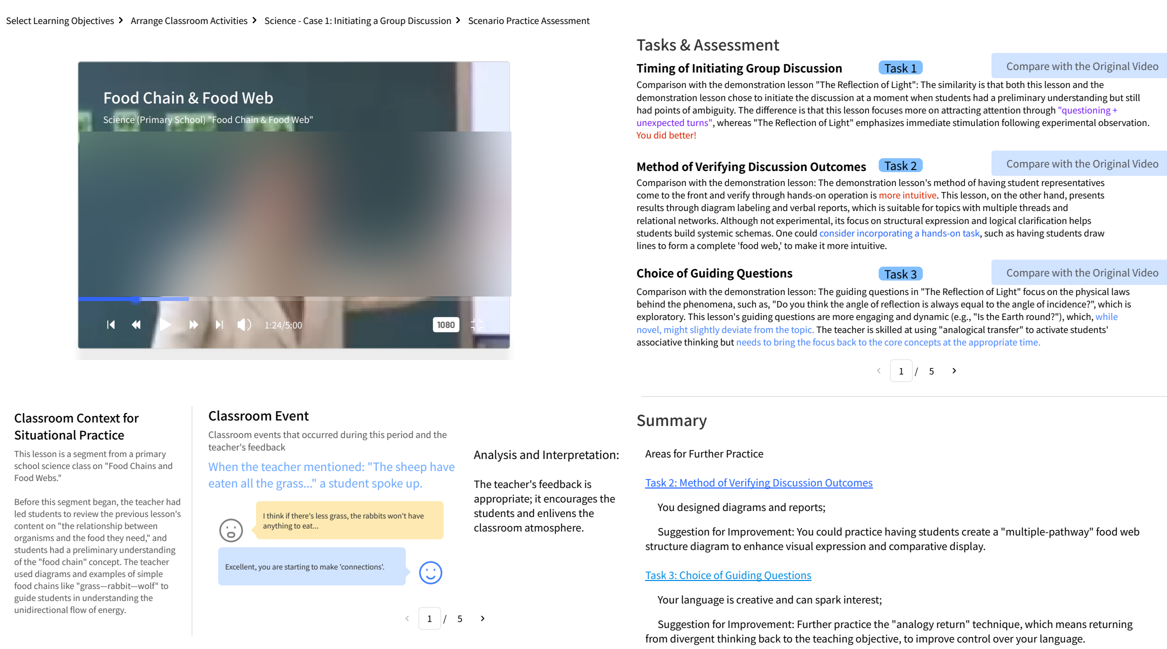

Figure 6: The evaluation page of the prototype. Translated from Chinese.

C Questionnaire Items

The questionnaire items used to evaluate the system are shown below.

- **Q1**: I feel more confident about classroom teaching.
- **Q2**: The video cases provided insights that I can directly apply to improve my teaching practice.
- **Q3**: The practice sessions were beneficial for applying the teaching skills learned from the videos in classroom settings.
- **Q4**: The training process kept me interested and engaged throughout.
- **Q5**: I found the system easy to use.
- **Q6**: I would like to use this system frequently for video-based learning and teaching reflection.

D Detailed Information of the Participants Involved

In this section in appendix, the detailed information of the participants involved is listed below. Note: FE1 works at a Teacher Development Center, an institution under the local Education Bureau responsible for curriculum research, teacher training, and professional development.

Table 3: Experienced and early-stage teachers involved in the formative study

ID	Gender	Background	Subjects
FE1	Female	38 years of teaching experience	–
FE2	Female	21 years of teaching experience	Chinese in elementary school
FE3	Female	25 years of teaching experience	Mathematics in elementary school
FE4	Male	3 years of teaching experience	Chinese in elementary school
FE5	Male	3 years of teaching experience	Mathematics in elementary school
FN1	Male	Normal university student	Chemistry in senior high school
FN2	Female	Normal university student	History in senior high school
FN3	Male	Normal university student	Physics in junior high school
FN4	Male	Normal university student	Music in senior high school

Table 4: Participants involved in the within-subjects study. VBL stands for video-based learning, and PD stands for professional development. All participants were pre-service teachers from normal universities; P9–P16 additionally reported teaching experience, including internships.

ID	Gender	Age	Background	Freq. of VBL for PD	Freq. of LLM Usage
P1	Male	24	Normal university student	Monthly	Daily
P2	Male	23	Normal university student	Monthly	Daily
P3	Male	25	Normal university student	Rarely	Weekly
P4	Female	19	Normal university student	Monthly	Weekly
P5	Male	21	Normal university student	Weekly	Daily
P6	Female	21	Normal university student	Never	Daily
P7	Female	22	Normal university student	Weekly	Daily
P8	Female	22	Normal university student	Weekly	Daily
P9	Male	24	4 months of teaching exp.	Yearly	Daily
P10	Male	20	1 year of teaching exp.	Rarely	Weekly
P11	Male	25	1 year of teaching exp.	Weekly	Weekly
P12	Female	20	1 year of teaching exp.	Monthly	Daily
P13	Male	23	1 year of teaching exp.	Never	Daily
P14	Female	23	1 year of teaching exp.	Yearly	Daily
P15	Female	19	10 months of teaching exp.	Rarely	Daily
P16	Female	24	1 year of teaching exp.	Weekly	Daily