VIDEOTEACH: DO VIDEO AND INSTRUCTIONAL QUALITY TRANSLATE INTO LEARNING TRANSFER?

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative models increasingly solve complex multimodal reasoning problems and present their solutions as instructional videos. Yet evaluating video-based teaching requires tracing the full path from the generated video, through teaching, to changes in learner performance. Existing benchmarks typically examine only one part of this path. We introduce VideoTeach[1], a benchmark of 1,176 teaching cases across five academic domains that evaluates all three layers at the problem level. It measures video artifact quality, instructional process quality, and a simulated student's learning outcomes on an anchor problem, a surface-transfer problem and a structural-transfer problem. Evaluation across four representative systems yields two findings. First, video and instructional quality are almost unrelated to learning gains ($|r| \leq 0.05$). Second, transfer remains weak: anchor-problem gains reach $+8.3$ percentage points, whereas surface- and structural-transfer gains peak at $+4.5$ and $+1.2$ points and can be negative. This transfer gap persists with stronger teacher backbones, including GPT-5.5, Gemini-3-Pro, and Claude-Opus-4.7. Additional teaching turns mainly improve anchor-problem performance without producing consistent surface- or structural-transfer gains. Overall, high-quality instructional videos do not guarantee transferable learning.

1 INTRODUCTION

Large language models (LLMs) and vision-language models (VLMs) now solve complex multimodal reasoning tasks (Guo et al., 2025; OpenAI, 2024), achieving strong performance on challenging tasks such as mathematics (Glazer et al., 2024; Balunovic et al., 2025) and physics (Zheng et al., 2025; Luo et al., 2026; Yu et al., 2025). Beyond producing solutions, these models can generate explanations, worked solutions, and feedback on learners' attempts (Saha et al., 2023; Team et al., 2025). However, generating such content is not enough: effective teaching should help learners follow reasoning, address misconceptions, and apply what they learn to new problems (Rozenblit & Keil, 2002; Roscoe & Chi, 2007). Recent systems take a step toward this, generating instructional videos that explain solutions (Ku et al., 2025; Chen et al., 2025b). Yet whether these videos teach learners effectively remains unclear.

Existing evaluations provide only a partial view of teaching effectiveness. Artifact-oriented evaluations assess the generated video, including rendering quality, visual presentation, and semantic correctness (Oli, 2026; Chen et al., 2025a; Ku et al., 2025; Lee et al., 2026; Mariam et al., 2026; Chen et al., 2026). Instruction-oriented evaluations assess logical coherence and pedagogical adequacy in videos, as well as explanation, feedback, and scaffolding in tutoring (Oli, 2026; Chen et al., 2025a; Ku et al., 2025; Lee et al., 2026; Mariam et al., 2026; Chen et al., 2026; Macina et al., 2023; 2025; Srinivasa et al., 2025). Outcome-oriented evaluations measure whether instruction improves learner performance. For example, Code2Video measures target-knowledge recovery with a model learner (Chen et al., 2025b). EducationQ and Teach2Eval measure whether multi-turn guidance improves student models' accuracy on the same target questions (Shi et al., 2025; Zhou et al., 2026). Together, these evaluations separately assess artifact quality, instructional quality, and learning outcomes, but leave two gaps. First, the three dimensions are generally measured in separate settings. Learning outcomes show whether instruction works, while artifact and instructional scores characterize upstream quality. Evaluating all three on the same teaching cases reveals whether upstream

[1]We will release the data and code after acceptance.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

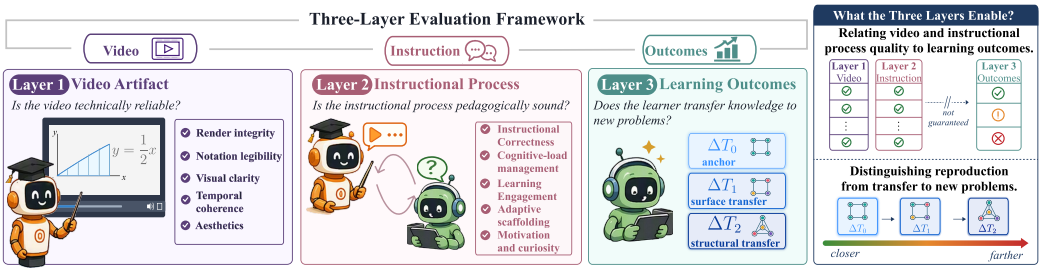

Figure 1: **Three-layer framework for educational video evaluation.** The framework jointly evaluates the video artifact (Layer 1), the instructional process (Layer 2), and the learning outcomes (Layer 3). Learning outcomes are measured on an anchor problem (ΔT_0), a surface transfer problem (ΔT_1), and a structural transfer problem (ΔT_2), representing progressively greater transfer distance.

quality translates into learning and helps locate failures when it does not. Second, existing outcome measures do not distinguish reproduction from transfer. Learners may reproduce taught solutions without transferring the underlying principle to new problems.

We address both gaps with VideoTeach, a benchmark that spans 1,176 teaching cases across five academic domains and aligns three evaluation layers for every problem–video instance (Figure 1). **Layer 1: Video Artifact** assesses the video's technical and visual quality across five dimensions: render integrity, notation legibility, visual clarity, temporal coherence, and aesthetics. **Layer 2: Instructional Process** evaluates the teacher–student interaction along five dimensions: instructional correctness, cognitive load management, learning engagement, adaptive scaffolding, and motivation and curiosity. **Layer 3: Learning Outcomes** measures three problem types at increasing transfer distances: p_0 is the instructional anchor problem; p_1 changes the surface features while preserving the target knowledge and solution procedure; and p_2 recombines the same knowledge within a different problem structure. Together, the three problem types probe reproduction, surface transfer, and structural transfer. The three layers allow us to test whether video and instructional quality translate into learning and to distinguish reproduction from transfer.

Applying VideoTeach to four representative systems spanning text-to-video synthesis and code-based animation yields two main findings. First, video and instructional quality do not reliably indicate learning gains. The system that scores highest on Layers 1 and 2 does not achieve the largest learning gain, and both layer 1 and 2 scores are nearly uncorrelated with learning gains ($|r| \leq 0.01$ and $|r| \leq 0.05$, respectively). Second, current systems largely fail to support transfer beyond the anchor problem. Anchor gains reach $+8.3$ pp, whereas surface- and structural-transfer gains peak at $+4.5$ and $+1.2$ pp and can be negative. This anchor-to-transfer gap persists with stronger teacher backbones and additional teaching turns: stronger backbones improve Layers 1–2, and additional turns improve anchor performance, but neither consistently improves surface or structural transfer. Together, these findings expose a central limitation of current instructional-video systems: they can produce high-quality videos, deliver strong instruction, and improve performance on taught problems without reliably supporting transfer to new problems.

2 Related Work

Instructional-video generation and quality-oriented evaluation. Instructional-video systems are generally divided into two categories: code-based generation and text-to-video synthesis. Code-based approaches either generate a complete Manim program directly or use structured pipelines that decompose the video into scenes and separate planning, coding, and refinement (Oli, 2026; Ku et al., 2025; Chen et al., 2025b). Text-to-video approaches translate LLM-generated scripts or prompts into videos using multimodal models (Oli, 2026; Ji et al., 2025; Lian et al., 2024). Existing evaluations mainly focus on video or instructional quality: video benchmarks assess technical, visual, semantic, and temporal validity (Chen et al., 2025a; Oli, 2026; Lee et al., 2026; Mariam et al., 2026; Chen et al., 2026), while tutoring benchmarks assess explanation, feedback, correctness, and scaffolding, primarily in text-based dialogue (Macina et al., 2023; 2025; Srinivasa et al., 2025). Together, these evaluations characterize the generated video and the instruction it conveys.

They identify technical and pedagogical weaknesses, but do not establish whether learners improve or can apply the taught knowledge to new problems.

Learning-outcome evaluation. Prior work measures learning with model learners at scale or with human learners. EducationQ and Teach2Eval assess whether text-based teaching improves performance on target problems; Code2Video's TeachQuiz tests recovery of target knowledge from generated videos; and TeachBench withholds target questions from the teacher and measures gains on knowledge-aligned questions in text-based teaching (Shi et al., 2025; Zhou et al., 2026; Chen et al., 2025b; Li et al., 2026). Studies with human learners provide direct evidence but report recall or aggregate pre–post gains for limited lessons or topics, making problem-level evaluation costly (Li et al., 2024; Netland et al., 2025; Joshi et al., 2026). Existing outcome measures do not explicitly separate reproduction from transfer at controlled distances. Compared with text-based teaching, video-based teaching adds a multimodal artifact that can fail independently of the instruction. Outcome scores alone cannot distinguish artifact failures from weak instruction or limited transfer. VideoTeach addresses both gaps by pairing each teaching case with an anchor and two controlled transfer problems while jointly measuring video artifact quality, instructional process quality, and learning outcomes.

Learning-science foundations. Cognitive load and multimedia-learning research motivate our instructional-quality rubrics (Sweller et al., 1998; Sweller, 2011; Mayer, 2008; Mayer et al., 2020; Team et al., 2025), while transfer and structure-mapping research motivate assessing learners across explicit transfer distances (Barnett & Ceci, 2002; Gentner, 1983). Evidence that instructor fluency can increase perceived learning without improving measured performance further motivates evaluating artifact quality, instructional quality, and learning outcomes separately (Carpenter et al., 2013).

3 THE VIDEOTEACH BENCHMARK

VideoTeach jointly evaluates video quality, instructional process quality, and learning outcomes across 1,176 problem-centered teaching cases spanning five academic domains. Each case pairs an instructional anchor problem p_0 with a surface-transfer problem p_1 and a structural-transfer problem p_2. Section 3.1 defines the three evaluation layers; Section 3.2 describes benchmark construction and quality control; and Section 3.3 specifies the controlled multi-agent evaluation protocol.

3.1 THREE-LAYER EVALUATION FRAMEWORK

Video-based teaching can fail at three layers: the generated video may be technically or visually deficient; the instructional process may be pedagogically unsound; or the learner may fail to reproduce the taught solution or transfer the underlying knowledge to new problems. We jointly evaluate all three layers on the same teaching cases to capture the full teaching chain and test how video and instructional quality relate to learning and transfer (Figure 1).

Layer 1: Video Artifact. This layer evaluates the generated video as a technical and visual artifact along five dimensions. Render integrity captures error-free and complete rendering; notation legibility assesses the readability of text, equations, and symbols; visual clarity examines the organization and presentation of visual elements; temporal coherence captures the continuity and timing of visual changes; and aesthetics reflects overall visual polish. We score each dimension independently and average the normalized scores. Full criteria appear in Appendix D.

Layer 2: Instructional Process. Unlike Layer 1, which evaluates how the video is rendered and presented, this layer assesses the pedagogical quality of the teacher's instruction given the student's current state and feedback. We adapt the LearnLM Pedagogy Rubric (Team et al., 2025) along five dimensions: *instructional correctness* evaluates the factual and logical validity of the explanation; *cognitive-load management*, the selection, sequencing, and pacing of information; *learning engagement*, support for the student's active participation; *adaptive scaffolding*, responsiveness to the student's demonstrated understanding or confusion; and *motivation and curiosity*, encouragement of continued effort and inquiry. Together, these dimensions characterize how well the teacher explains, adapts, and scaffolds instruction during the interaction. Full rubric details appear in Appendix Table 5.

Table 1: VideoTeach scale and composition. Each case contains an anchor, surface-transfer, and structural-transfer problem.

Benchmark	Cases	Problems	Domains	Sub-fields	Sources
VideoTeach	1,176	3,528	5	24	8

Layer 3: Learning Outcomes. This layer tests whether teaching produces transferable learning rather than only reproduction. Each case contains three fixed assessment problems:

- Anchor p_0: the instructional anchor problem used during teaching;
- Surface transfer p_1 : a new problem that changes the surface features while preserving the target knowledge and solution procedure;
- Structural transfer p_2: a new problem that recombines the same target knowledge within a different problem structure.

The student completes all three problems before and after instruction. For each p_t, where $t \in \{0, 1, 2\}$, let T_t^{pre} and T_t^{post} denote the student's pre- and post-test correctness, respectively. We define the learning gain as

$$\Delta T_t = T_t^{\text{post}} - T_t^{\text{pre}}.$$

Accordingly, ΔT_0 measures reproduction on the anchor problem, while ΔT_1 and ΔT_2 measure surface and structural transfer from the anchor problem. The three problems encode increasing distance from the anchor problem rather than absolute difficulty.

3.2 Benchmark Construction and Quality Control

Data sources and domain coverage. Table 1 summarizes the benchmark's scale, coverage, and composition. We construct VideoTeach from an initial pool of 24,297 problems drawn from eight public benchmarks spanning mathematics, physics, chemistry, humanities, and social sciences. After the filtering stages described below, we sample by source benchmark and sub-field to preserve broad coverage, yielding 1,176 anchor problems p_0 across five domains and 24 sub-fields. Appendix Figure 7 reports the detailed distributions.

Source-problem quality control. We first remove 694 near-duplicate candidates by flagging problem pairs whose sentence-embedding cosine similarity exceeds 0.85. We then retain only self-contained, unambiguous problems with a single fixed answer, removing a further 4,312 candidates. Each remaining problem is solved independently and checked against its stored reference answer, eliminating another 1,529 candidates; disagreements are adjudicated by a stronger model and escalated to human review when unresolved. Finally, three domain experts independently audit a random 10% sample of the final dataset. Under a conservative unanimity criterion, 98% of the audited items pass both problem-validity and reference-answer checks. Appendix A provides the complete curation and quality control procedure.

Transfer-problem construction and quality control. For each anchor problem, the evaluator constructs p_1 and p_2 through a role-separated generate–solve–validate pipeline rather than accepting a single generation. The generator creates candidates under the transfer definitions introduced above: p_1 preserves the target knowledge and solution procedure while changing surface features, whereas p_2 recombines the same knowledge point within a different problem structure. A second model independently solves each candidate, and a third model cross-validates the problem, derived solution, and intended transfer level, checking for ambiguity, unsolvability, answer errors, trivial copying, and violations of the p_1/p_2 definitions. Approximately 3% of initial candidates are flagged for review; these problems are regenerated and re-evaluated until they pass the full pipeline. Separately, three domain experts independently audit a random 10% sample of the generated transfer problems, yielding a 98% pass rate. Appendix B provides the complete generation constraints, validation procedure, prompts, and transfer examples.

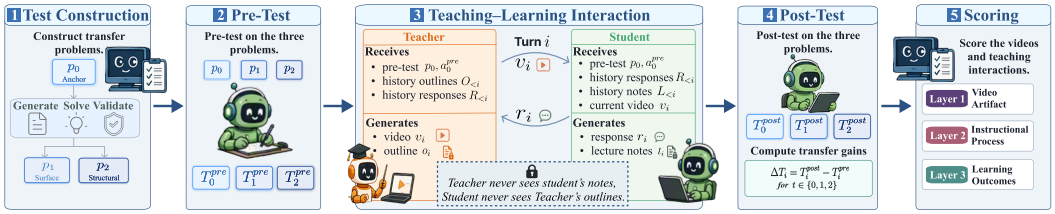

Figure 2: **Multi-agent protocol for the three-layer evaluation.** The protocol proceeds through five stages: test construction, pre-test, teaching-learning interaction, post-test, and scoring. The evaluator scores video artifact quality (Layer 1) and instructional process quality (Layer 2), and computes pre-to-post gains on the anchor, surface-transfer, and structural-transfer problems (Layer 3).

3.3 Multi-Agent Evaluation Protocol

To operationalize the three-layer evaluation, we run each teaching case with separate teacher, student, and evaluator agents. The teacher adapts to the student's responses, while the evaluator withholds the transfer problems from the teacher and conducts the assessment. These roles form the controlled five-stage protocol shown in Figure 2.

Test Construction and Pre-test. To measure problem-level learning gains, we establish a pre-instruction baseline and keep the transfer tests hidden from the teacher. Given an anchor problem p_0, the evaluator generates a surface-transfer problem p_1 and a structural-transfer problem p_2: $(p_1, p_2) = \mathcal{E}_{\text{gen}}(p_0)$, where $\mathcal{E}_{\text{gen}}$ denotes its test-generation procedure. The student then solves all three problems: $a_t^{\text{pre}} = \mathcal{S}_{\text{solve}}(p_t)$ and $T_t^{\text{pre}} = \mathbb{1}[a_t^{\text{pre}} \text{ is correct}]$ for $t \in \{0, 1, 2\}$, where a_t^{pre} contains the student's reasoning. During instruction, the teacher receives only p_0 and a_0^{pre}; p_1, p_2, their pre-test responses, and all correctness labels remain hidden.

Teaching–Learning Interaction. The teacher and student interact for $n \geq 1$ turns. Before turn i, let $O_{<i} = (o_1, \ldots, o_{i-1})$ denote the teacher's private outlines, $R_{<i} = (r_1, \ldots, r_{i-1})$ the student's previous responses, and $L_{<i} = (\ell_1, \ldots, \ell_{i-1})$ the student's private lecture notes. Each turn follows

$$(v_i, o_i) = \mathcal{T}_{\text{teach}}(p_0, a_0^{\text{pre}}, O_{<i}, R_{<i}),$$
$$(r_i, \ell_i) = \mathcal{S}_{\text{learn}}(p_0, a_0^{\text{pre}}, R_{<i}, L_{<i}, v_i).$$

Here, $\mathcal{T}_{\text{teach}}$ produces a video v_i and its outline o_i, while $\mathcal{S}_{\text{learn}}$ returns a response r_i and stores a private lecture note ℓ_i. All histories are empty at $i = 1$. Both agents observe p_0, a_0^{pre}, and $R_{<i}$. The teacher additionally retains $O_{<i}$, whereas the student retains $L_{<i}$ and receives the current video v_i. The student cannot access $O_{<i}$, and the teacher cannot access $L_{<i}$. Replaying all previous videos would cause the context to grow rapidly across turns. Later turns therefore use the teacher's outlines and the student's lecture notes as compact, role-specific representations of the video history. The student's responses remain shared so that the teacher can adapt its instruction.

Post-test. After the final teaching turn, the student retakes all three problems in separate sessions. For each $t \in \{0, 1, 2\}$, it produces $a_t^{\text{post}} = \mathcal{S}_{\text{solve}}(p_0, a_0^{\text{pre}}, L_{\leq n}, R_{\leq n}, p_t)$, with correctness $T_t^{\text{post}} = \mathbb{1}[a_t^{\text{post}} \text{ is correct}]$. Here, a_t^{post} denotes the complete reasoning for p_t. Each session receives the anchor problem p_0, its pre-test response a_0^{pre}, the accumulated lecture notes $L_{\leq n}$, the response history $R_{\leq n}$, and the target problem p_t. Because the teaching history is conditioned on p_0 and a_0^{pre}, including them preserves the instructional context; no correctness label or reference solution is revealed. The teaching videos are not replayed, and outputs from one post-test session are not passed to another, preventing cross-problem information leakage.

Scoring. The evaluator scores each teaching turn independently. Layers 1 and 2 average scores across five dimensions, while Layer 3 records binary pre- and post-test correctness and reports the three learning gains defined in Section 3.1. Appendices D, E, and F report the complete criteria, prompts, and aggregation rules; Table 5 summarizes the Layer 2 rubric.

Reference solutions and scoring rubrics remain private to the evaluator, which does not participate in teaching. These boundaries limit teacher-side test leakage and separate instruction from

measurement. Complete prompts appear in Appendix C, with additional implementation details in Appendix G.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Teacher systems. We evaluate four teacher systems that span text-to-video synthesis and code-based rendering. **T2V** uses an LLM to produce a video-generation prompt that describes the intended visual content, which is then passed to a text-to-video model (Ji et al., 2025; Lian et al., 2024). We use Wan2.2 in the main comparison and include Seedance-v2 and Kling-v3-turbo in additional analyses. The code-based systems render LLM-generated Manim programs. **MONO** generates a complete Manim program in one pass (Oli, 2026). **FREE** and **SCAF** instead use plan-then-code pipelines: both decompose the video into scenes, plan each scene's instructional content, and use a coding agent to generate its Manim program (Ku et al., 2025; Chen et al., 2025b). **FREE** organizes the scenes through a free-form teaching plan and allows the model to determine their visual composition. **SCAF** uses a structured storyboard and generates each scene with a predefined layout scaffold. FREE and SCAF also differ in their prompts, output formats, and rendering constraints; we therefore compare them as end-to-end systems, not as a scaffold ablation.

Student and evaluator agents. Unless otherwise specified, Qwen2.5-VL-7B-Instruct serves as the student and GPT-5.5 as the evaluator. We select a moderate-capability student to avoid severe floor and ceiling effects. All videos are silent: the Student receives the full MP4, whereas the L1/L2 Evaluator receives frames sampled every two seconds (Appendix G); narration and audio–visual coordination are outside our scope.

Full and Lite benchmark configurations. Generating a one-minute video takes approximately 10 minutes in our implementation, making repeated evaluation of all 1,176 cases costly. Following BIG-Bench Lite (BIG-bench authors, 2023), we define two configurations. **VideoTeach** uses all 1,176 cases for the four-system, single-turn comparison. **VideoTeach-Lite** is a fixed stratified subset of 300 cases shared across systems and runs for backbone, domain-level, and multi-turn analyses.

4.2 MAIN RESULTS

Figure 3 reports the main results using DeepSeek-V4-Pro for the three code-based systems and Wan2.2 for T2V. We refer to video artifact quality, instructional process quality, and learning outcomes as L1, L2, and L3, respectively. In the full set, failed generations receive zero, so systems with lower generation success receive lower aggregate scores. We therefore also report results on the 529 common-success cases for which all systems generated a video, enabling a direct comparison of their L1–L3 performance. Appendix Table 7 reports the complete numerical results.

Video artifact and instructional process quality show little linear association with learning gains. System rankings diverge across layers. On the common-success subset, FREE leads L1 (66.4) and L2 (52.1), whereas MONO achieves the largest mean L3 gain, averaged over ΔT_0, ΔT_1, and ΔT_2 ($+3.4$ pp versus $+2.4$ pp for FREE). To examine whether this mismatch also appears across cases, we compute correlations separately for each system. Each case contributes the L1, L2, and L3 scores; Figure 4 reports the Pearson correlations across cases for each pair of layers. The L1–L2 correlations range from -0.02 to 0.16, while correlations of mean L3 gain with L1 and L2 remain near zero ($|r| \leq 0.01$ and $|r| \leq 0.05$, respectively). Thus, neither L1 nor L2 serves as a reliable linear proxy for learning gains, supporting separate reporting of all three layers.

Video-based gains exceed the NO-VIDEO baseline but weaken with transfer distance. NO-VIDEO repeats the pre/post assessment without video-based instruction, capturing changes from repeated testing or self-correction after the pre-test. On the common-success subset, it yields -1.3, -1.7, and -2.3 pp on ΔT_0, ΔT_1, and ΔT_2, respectively. Every video-based system exceeds these values on all three problems, by 2.1–9.6, 0.9–6.2, and 1.4–3.5 pp, respectively. However, outperforming the NO-VIDEO baseline should not be mistaken for successful transfer. Current systems largely fail to support learning beyond the anchor problem: gains generally weaken with transfer

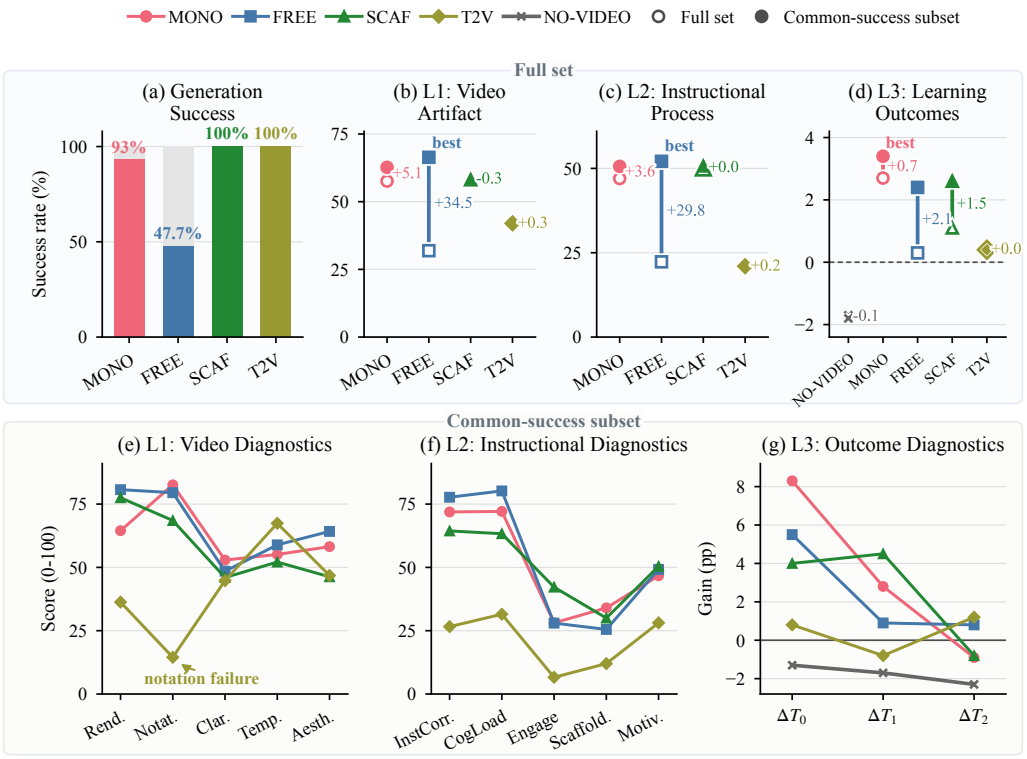

Figure 3: **Main results on three-layer performance and generation reliability.** (a) Video-generation success rates. (b–d) L1, L2, and L3 scores on the full benchmark ($N = 1{,}176$; generation failures receive zero) and the common-success subset ($N = 529$; all systems successfully generate a video). Open and filled markers denote the full and common-success sets, respectively, and annotations show the score differences. (e–g) Dimension-level L1/L2 scores and transfer-specific L3 gains on the common-success subset. L1/L2 range from 0 to 100. L3 reports percentage-point pre-to-post gains. NO-VIDEO is an L3-only baseline that repeats the pre/post test without a video. In (e), Rend., Notat., Clar., Temp., and Aesth. denote render integrity, notation legibility, visual clarity, temporal coherence, and aesthetics. In (f), InstCorr., CogLoad, Engage, Scaffold., and Motiv. denote instructional correctness, cognitive-load management, learning engagement, adaptive scaffolding, and motivation and curiosity. System rankings differ across upstream quality and learning outcomes, while transfer gains remain small.



Figure 4: Within-system item-level correlations between evaluation layers.

T2V backbone	Layer 1	Layer 2	Layer 3
Wan2.2	36.7	17.5	-1.2
Kling-v3-turbo	35	29.8	-1.7
Seedance-v2	45	35.6	-1.5

Table 2: Results for three T2V video backbones on VideoTeach-Lite. Layer 3 is the mean of the three pre/post gains ($\Delta T_0, \Delta T_1, \Delta T_2$).

distance, and structural-transfer gains remain between -0.9 and $+1.2$ pp. Reporting ΔT_0 alone can therefore overstate teaching effectiveness: it measures improvement on the anchor problem, whereas ΔT_1 and ΔT_2 test whether learning transfers to new problems.

T2V trails across layers, while FREE a combines low generation success rate with high successful-case quality. On the common-success subset, T2V ranks last among the video-based

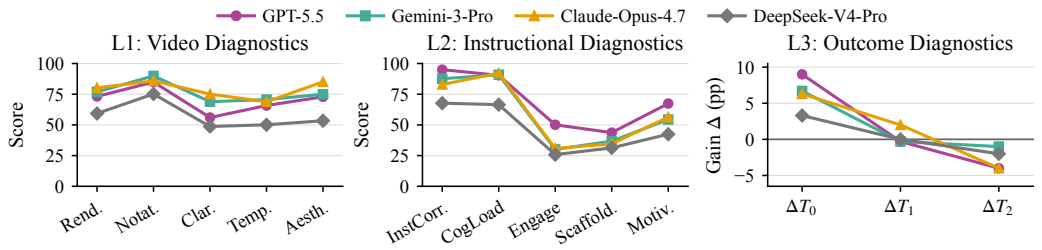

Figure 5: **Effect of the teacher backbone across the three evaluation layers.** We compare the default DeepSeek-V4-Pro teacher with GPT-5.5, Gemini-3-Pro, and Claude-Opus-4.7. The three alternative backbones improve L1, L2, and anchor performance ΔT_0 over the default teacher, but surface-transfer gains ΔT_1 remain small and structural-transfer gains ΔT_2 remain negative. Higher is better; metric abbreviations are defined in Figure 3.

systems in L1 (42.0), L2 (21.0), and mean L3 gain ($+0.4$ pp). Its largest gap from the code-based systems is in notation legibility: T2V scores 14.5 versus 68.5–82.6 for the code-based systems, despite achieving the highest temporal coherence (67.4). Garbled text and mathematical symbols may therefore obscure the instructional content and limit learning. This weakness persists across three T2V backbones on VideoTeach-Lite, where L1 ranges from 35.0 to 45.0, L2 from 17.5 to 35.6, and mean L3 gain from -1.7 to -1.2 pp (Table 2). By contrast, FREE succeeds on only 47.7% of cases, compared with 93–100% for the other systems, but leads L1 (66.4) and L2 (52.1) on successful cases. Both FREE and SCAF plan the video by scene before generating scene-level code; FREE permits free-form scene composition, whereas SCAF uses a predefined layout scaffold. Their results suggest that greater generation flexibility can improve successful outputs at the cost of generation success rate, although differences between the full pipelines prevent attributing this trade-off solely to layout constraints.

4.3 ANALYSIS

The main results show that current systems largely fail to support learning beyond the anchor problem. We therefore test whether this pattern persists under stronger teaching conditions. We focus on MONO, which achieves the highest mean L3 gain in the main comparison, and conduct two controlled interventions: using stronger teacher backbones and increasing the number of teaching turns. We hold the student, evaluator, problems, and remaining protocol fixed.

Stronger teacher backbones improve upstream quality but not transfer. We first strengthen the teacher backbone in MONO. On VideoTeach-Lite, we replace DeepSeek-V4-Pro with GPT-5.5, Gemini-3-Pro, and Claude-Opus-4.7 (Figure 5). All three backbones improve L1 and L2 over the default teacher. Claude-Opus-4.7 achieves the highest L1 score (79.0), while GPT-5.5 achieves the highest L2 score (69.3); no backbone leads both layers. Learning engagement and adaptive scaffolding nevertheless remain among the weakest L2 dimensions. More importantly, learning gains retain the same anchor-to-transfer gap: ΔT_0 ranges from $+6.3$ to $+9.0$ pp, ΔT_1 from -0.3 to $+2.0$ pp, and ΔT_2 remains negative across all three backbones (-4.0 to -1.0 pp). Thus, stronger teachers improve upstream quality and anchor-problem performance, but do not resolve surface or structural transfer. The anchor-to-transfer gap therefore persists even with stronger teacher backbones.

Additional teaching turns improve anchor performance but not transfer. We next extend MONO from one to five teaching turns using Claude-Opus-4.7, Gemini-3-Pro, and GPT-5.5 as teacher backbones, while holding the student, evaluator, and benchmark cases fixed. After each turn, we measure L1 and L2 for the current video and re-evaluate the student using the notes accumulated through that turn (Figure 6). Across all three backbones, L1 and L2 remain broadly stable, while ΔT_0 increases from 7.0–8.3 pp after the first turn to 12.7–15.2 pp after the fifth. By contrast, ΔT_1 and ΔT_2 remain small and non-monotonic. The anchor-to-transfer gap therefore persists as the interaction grows longer: additional turns strengthen anchor performance without producing reliable surface or structural transfer.

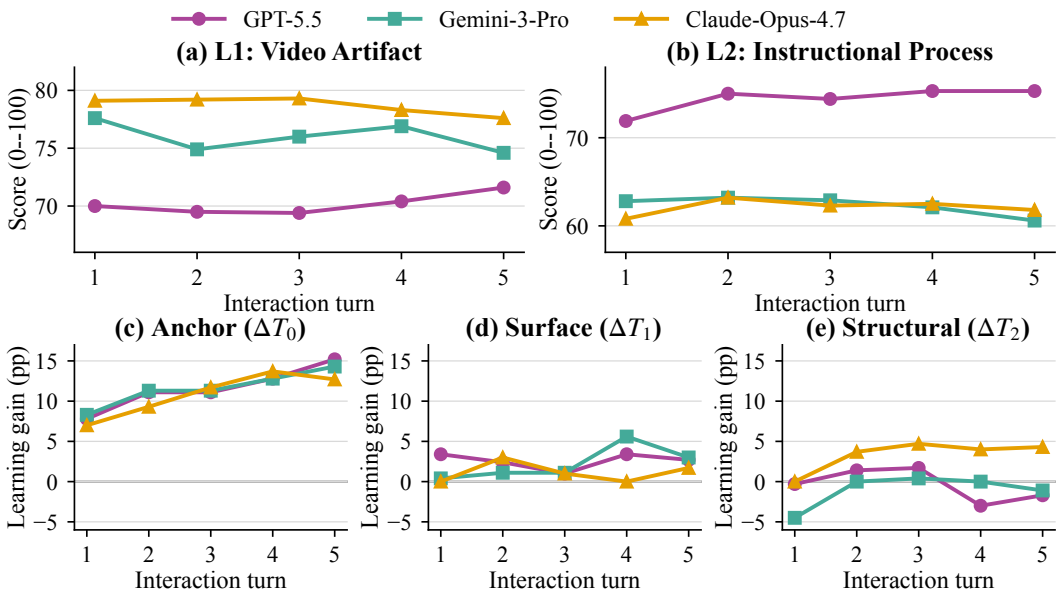

Figure 6: **Effect of additional teaching turns across teacher backbones.** Results for MONO on VideoTeach-Lite over five teaching turns using GPT-5.5, Gemini-3-Pro, and Claude-Opus-4.7. L1/L2 remain broadly stable; additional turns increase anchor performance but do not consistently improve surface or structural transfer. Higher is better.

Additional analyses test robustness and explain the observed gaps. Appendix H examines evaluator sensitivity, human–student-agent agreement, domain-level variation, and qualitative analyses. Rescoring the same cases with three evaluator models yields broadly consistent L1/L2 pipeline rankings ($\tau_b = 0.601$–0.671). A study with 12 human participants also reproduces the aggregate system ranking and the decline from anchor to farther transfer observed with the student agent. Domain-level results reveal distinct system strengths across layers. Qualitative cases clarify why L1/L2 scores do not reliably track L3 outcomes. Videos with high artifact or instructional-process scores can present the correct answer clearly while omitting the underlying rule or leaving a key misconception unresolved, thereby limiting transfer. Conversely, even a low-quality video may expose a salient fact or answer cue that is sufficient to correct the anchor response but does not generalize to related problems. Multi-turn cases reveal a similar pattern: later videos often revisit or refine the same anchor-specific procedure, sometimes with additional examples or formulas, stabilizing performance on p_0 while p_1 and p_2 remain incorrect. Together, these analyses support the system-level comparisons and show that limited transfer remains a central weakness of current methods, motivating future work on teaching strategies that generalize beyond the anchor problem.

5 CONCLUSION

We introduced VideoTeach, a benchmark that evaluates the full instructional-video teaching chain by jointly measuring video artifact quality, instructional process quality, and learning outcomes across 1,176 teaching cases. VideoTeach separates reproduction from surface and structural transfer through problems at increasing transfer distances. Across four representative systems, L1/L2 scores show almost no linear association with learning gains. Improvements also concentrate on the anchor problem, while transfer remains weak. This pattern persists with stronger teacher backbones and additional teaching turns: stronger teachers improve upstream quality, and more interaction turns improve anchor performance, but neither consistently improves surface or structural transfer. Evaluator comparisons and a human study further support these findings. VideoTeach exposes a central disconnect: high video and instructional quality do not reliably lead to transferable learning. Single-layer evaluations would miss this gap. Future instructional-video systems should move beyond optimizing presentation or anchor-problem performance and explicitly target the learner's ability to apply the underlying knowledge to new problems.

GENERATIVE AI USE DISCLOSURE

Generative AI was used in two roles in this work. First, as part of the research methodology, language and vision-language models served as teacher, student, and evaluator agents. They were also used to generate transfer problems and instructional videos and to support automated data curation and answer verification. These uses, including the specific models and quality-control procedures, are documented in the Methods and Appendix.

Second, generative AI tools, including Codex and Claude Code, assisted with language editing and with aggregating, analyzing, and visualizing experimental results. The authors manually checked all reported statistics, tables, and figures against the underlying experimental records. They also reviewed and revised all AI-assisted text to ensure that it did not alter the scientific content, claims, or conclusions. The authors take full responsibility for the final manuscript.

ETHICS STATEMENT

VideoTeach is intended for research on instructional-video generation and evaluation, not for classroom deployment or other high-stakes educational use without human oversight. Generated videos may contain convincing factual or pedagogical errors; benchmark scores therefore do not certify instructional safety. The benchmark itself uses academic problems and requires no sensitive learner data. The primary experiments use model-based student agents. Appendix H.3 reports a study with 12 human participants. The human study was approved by the authors' institutional ethics committee (protocol details withheld for anonymity). All participants provided informed consent and were compensated according to local standards. We collected only the task responses required for evaluation. We will document source licenses and release only materials permitted by their licenses and, where applicable, participant consent.

REPRODUCIBILITY STATEMENT

We support reproducibility by documenting the complete VideoTeach evaluation protocol, information-access boundaries, transfer-level definitions, scoring rules, and aggregation procedures in the main paper. The experimental setup specifies the teacher, student, and evaluator models, video-generation pipelines, evaluation configurations, and treatment of generation failures. The appendix further provides the data sources, curation and deduplication procedures, transfer-problem validation, quality-control results, evaluation rubrics, and prompts. We will publicly release the benchmark data and code after review, subject to the licenses of the source datasets and model providers.

REFERENCES

Mislav Balunovic, Jasper Dekoninck, Ivo Petrov, Nikola Jovanović, and Martin T. Vechev. Matharena: Evaluating LLMs on uncontaminated math competitions. In *Advances in Neural Information Processing Systems*, volume 38, pp. 22851–22888. Neural Information Processing Systems Foundation, Inc., 2025. doi: 10.52202/085713-0679. URL https://doi.org/10.52202/085713-0679.

Susan M. Barnett and Stephen J. Ceci. When and where do we apply what we learn? a taxonomy for far transfer. *Psychological Bulletin*, 128(4):612–637, 2002. doi: 10.1037/0033-2909.128.4.612.

BIG-bench authors. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=uyTL5Bvosj.

Robert A. Bjork. Memory and metamemory considerations in the training of human beings. In *Metacognition*, pp. 185–206. The MIT Press, 1994. doi: 10.7551/mitpress/4561.003.0011. URL https://doi.org/10.7551/mitpress/4561.003.0011.

Shana K. Carpenter, Miko M. Wilford, Nate Kornell, and Kellie M. Mullaney. Appearances can be deceiving: Instructor fluency increases perceptions of learning without increasing actual learning. *Psychonomic Bulletin & Review*, 20(6):1350–1356, 2013. doi: 10.3758/s13423-013-0442-z.

Baoliang Chen, Xinlong Bu, Hanwei Zhu, Lingyu Zhu, and Jieyu Zhan. Eduvqa: Towards concept-aware assessment of educational ai-generated videos, 2026. URL https://arxiv.org/abs/2603.03066.

Hao Chen, Tianyu Shi, Pengran Huang, Zeyuan Li, Jiahui Pan, Qianglong Chen, and Lewei He. VisualEDU: A benchmark for assessing coding and visual comprehension through educational problem-solving video generation. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 16363–16394, Suzhou, China, November 2025a. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.889. URL https://aclanthology.org/2025.findings-emnlp.889/.

Yanzhe Chen, Kevin Qinghong Lin, and Mike Zheng Shou. Code2video: A code-centric paradigm for educational video generation, 2025b. URL https://arxiv.org/abs/2510.01174.

Michelene Chi and Ruth Wylie. The icap framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist*, 49, 10 2014. doi: 10.1080/00461520.2014.965823.

John H. Flavell. Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist*, 34:906–911, 1979. URL https://api.semanticscholar.org/CorpusID:8841485.

Dedre Gentner. Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7(2):155–170, 1983. ISSN 0364-0213. doi: https://doi.org/10.1016/S0364-0213(83)80009-3. URL https://www.sciencedirect.com/science/article/pii/S0364021383800093.

Elliot Glazer, Ege Erdil, Tamay Besiroglu, Diego Chicharro, Evan Chen, Alex Gunning, Caroline Falkman Olsson, Jean-Stanislas Denain, Anson Ho, Emily de Oliveira Santos, Olli Järviniemi, Matthew Barnett, Robert Sandler, Matej Vrzala, Jaime Sevilla, Qiuyu Ren, Elizabeth Pratt, Lionel Levine, Grant Barkley, Natalie Stewart, Bogdan Grechuk, Tetiana Grechuk, Shreepranav Varma Enugandla, and Mark Wildon. Frontiermath: A benchmark for evaluating advanced mathematical reasoning in AI, 2024. URL https://arxiv.org/abs/2411.04872.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang.

DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645 (8081):633–638, September 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09422-z. URL https://doi.org/10.1038/s41586-025-09422-z.

Suzanne E. Hidi and K. Ann Renninger. The four-phase model of interest development. *Educational Psychologist*, 41:111 – 127, 2006. URL https://api.semanticscholar.org/CorpusID:18870838.

Yatai Ji, Jiacheng Zhang, Jie Wu, Shilong Zhang, Shoufa Chen, Chongjian Ge, Peize Sun, Weifeng Chen, Wenqi Shao, Xuefeng Xiao, Weilin Huang, and Ping Luo. Prompt-a-video: Prompt your video diffusion model via preference-aligned llm. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18725–18735, October 2025.

Aastha Joshi, Hongyi Ke, Meet Gajjar, Aaron Christian, Qi Wang, and Jun Chen. LLM2Manim: Pedagogy-aware AI generation of STEM animations, 2026. URL https://arxiv.org/abs/2604.05266.

Max Ku, Cheuk Hei Chong, Jonathan Leung, Krish Shah, Alvin Yu, and Wenhu Chen. TheoremExplainAgent: Towards video-based multimodal explanations for LLM theorem understanding. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6663–6684, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.332. URL https://aclanthology.org/2025.acl-long.332/.

Unggi Lee, Hoyoung Ahn, Yoon Choi, Seonmin Eun, Jahyun Jeong, Seonmin Jin, Harmony Jung, Hye Jin Kim, Chaerin Lee, Hyunji Lee, Jeongjin Lee, Soohwan Lee, Young-Seok Oh, Jaehyeon Park, Sun ok Ryu, Sunyoung Shin, Yoorim Son, Haeun Park, and Yeil Jeong. Are video models zero-shot learners and reasoners in education? eduvideobench, a knowledge-skills-attitude benchmark for educational video generation, 2026. URL https://arxiv.org/abs/2605.26918.

Zheng Li, Siyao Song, Jingyuan Ma, Rui Li, Ying Zeng, Minghao Li, and Zhifang Sui. Teachbench: A syllabus-grounded framework for evaluating teaching ability in large language models. *ArXiv*, abs/2601.21375, 2026. URL https://api.semanticscholar.org/CorpusID:285140783.

Zoe Ruo-Yu Li, Caswell Barry, and Mutlu Cukurova. Adult learners recall and recognition performance and affective feedback when learning from an AI-generated synthetic video, 2024. URL https://arxiv.org/abs/2412.10384.

Long Lian, Baifeng Shi, Adam Yala, Trevor Darrell, and Boyi Li. LLM-grounded video diffusion models. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 50207–50227, 2024. URL https://openreview.net/forum?id=exKHibougU.

George Loewenstein. The psychology of curiosity: A review and reinterpretation. In George Loewenstein (ed.), *Exotic Preferences: Behavioral Economics and Human Motivation*. Oxford University Press, 08 2007. ISBN 9780199257072. doi: 10.1093/oso/9780199257072.003.0006. URL https://doi.org/10.1093/oso/9780199257072.003.0006.

Yun Luo, Futing Wang, Qianjia Cheng, Fangchen Yu, Haodi Lei, Jianhao Yan, Chenxi Li, Jiacheng Chen, Yufeng Zhao, Haiyuan Wan, Yuchen Zhang, Shenghe Zheng, Junchi Yao, Qingyang Zhang, Haonan He, Wenxuan Zeng, Li Sheng, Chengxing Xie, Yuxin Zuo, Yizhuo Li, Yulun Wu, Rui Huang, Dongzhan Zhou, Kai Chen, Yu Qiao, Lei Bai, Yu Cheng, Ning Ding, Bowen Zhou, Peng Ye, and Ganqu Cui. P1-VL: bridging visual perception and scientific reasoning in physics olympiads, 2026. URL https://arxiv.org/abs/2602.09443.

Jakub Macina, Nico Daheim, Sankalan Chowdhury, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. MathDial: A Dialogue Tutoring Dataset with Rich Pedagogical Properties Grounded in Math Reasoning Problems. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023. doi: 10.18653/v1/2023.findings-emnlp.372.

Jakub Macina, Nico Daheim, Ido Hakimi, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. Mathtutorbench: A benchmark for measuring open-ended pedagogical capabilities of LLM tutors. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 204–221. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.11.

K. M. Megha Mariam, Aditya Arun, Zakaria Laskar, and C. V. Jawahar. Phyeduvideo: A benchmark for evaluating text-to-video models for physics education. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2026, Tucson, AZ, USA, March 6-10, 2026*, pp. 8690–8699. IEEE, 2026. doi: 10.1109/WACV61042.2026.00838. URL https://doi.org/10.1109/WACV61042.2026.00838.

Richard Mayer, Logan Fiorella, and Andrew Stull. Five ways to increase the effectiveness of instructional video. *Educational Technology Research and Development*, 68:837–852, 02 2020. doi: 10.1007/s11423-020-09749-6.

Richard E. Mayer. Applying the science of learning: Evidence-based principles for the design of multimedia instruction. *American Psychologist*, 63(8):760–769, 2008. doi: 10.1037/0003-066X.63.8.760. URL https://doi.org/10.1037/0003-066X.63.8.760.

Richard E. Mayer. *Multimedia Learning*. Cambridge University Press, New York, 2nd edition, 2009.

Torbjørn Netland, Oliver von Dzengelevski, Katalin Tesch, and Daniel Kwasnitschka. Comparing human-made and AI-generated teaching videos: An experimental study on learning effects. *Computers & Education*, 224:105164, 2025. doi: 10.1016/j.compedu.2024.105164. URL https://doi.org/10.1016/j.compedu.2024.105164.

Nabin Oli. Manibench: A benchmark for testing visual-logic drift and syntactic hallucinations in manim code generation, 2026. URL https://arxiv.org/abs/2603.13251.

OpenAI. Openai o1 system card, 2024. URL https://arxiv.org/abs/2412.16720.

Jean Piaget. *The equilibration of cognitive structures : the central problem of intellectual development*. Univ. of Chicago Pr., Chicago, 1985. ISBN 0226667812.

Rod D. Roscoe and Michelene T. H. Chi. Understanding tutor learning: Knowledge-building and knowledge-telling in peer tutors' explanations and questions. *Review of Educational Research*, 77(4):534–574, 2007. doi: 10.3102/0034654307309920.

Leonid Rozenblit and Frank Keil. The misunderstood limits of folk science: an illusion of explanatory depth. *Cognitive Science*, 26(5):521–562, 2002. doi: 10.1207/s15516709cog2605_1.

Swarnadeep Saha, Peter Hase, and Mohit Bansal. Can language models teach? teacher explanations improve student performance via personalization. In *Advances in Neural Information Processing Systems 36*, 2023. doi: 10.52202/075280-2746.

Yao Shi, Rongkeng Liang, and Yong Xu. EducationQ: Evaluating LLMs' teaching capabilities through multi-agent dialogue framework. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 32799–32828, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1576. URL https://aclanthology.org/2025.acl-long.1576/.

Rakshith S. Srinivasa, Zora Che, Chen Bo Calvin Zhang, Diego Mares, Ernesto Hernandez, Jayeon Park, Dean Lee, Guillermo Mangialardi, Charmaine Ng, Ed-Yeremai Hernandez Cardona, Anisha Gunjal, Yunzhong He, Bing Liu, and Chen Xing. TutorBench: A benchmark to assess tutoring capabilities of large language models, 2025. URL https://arxiv.org/abs/2510.02663. Accepted as a poster at the DATA-FM Workshop, ICLR 2026.

John Sweller. Cognitive load theory. In *Psychology of Learning and Motivation*, volume 55, pp. 37–76. Elsevier, 2011. doi: 10.1016/B978-0-12-387691-1.00002-8. URL https://doi.org/10.1016/B978-0-12-387691-1.00002-8.

John Sweller, Jeroen J. G. Van Merrienboer, and Fred Paas. Cognitive architecture and instructional design. *Educational Psychology Review*, 10:251–296, 09 1998. doi: 10.1023/a:1022193728205.

LearnLM Team, Abhinit Modi, Aditya Srikanth Veerubhotla, Aliya Rysbek, Andrea Huber, Brett Wiltshire, Brian Veprek, Daniel Gillick, Daniel Kasenberg, Derek Ahmed, Irina Jurenka, James Cohan, Jennifer She, Julia Wilkowski, Kaiz Alarakyia, Kevin R. McKee, Lisa Wang, Markus Kunesch, Mike Schaekermann, Miruna Pîslar, Nikhil Joshi, Parsa Mahmoudieh, Paul Jhun, Sara Wiltberger, Shakir Mohamed, Shashank Agarwal, Shubham Milind Phal, Sun Jae Lee, Theofilos Strinopoulos, Wei-Jen Ko, Amy Wang, Ankit Anand, Avishkar Bhoopchand, Dan Wild, Divya Pandya, Filip Bar, Garth Graham, Holger Winnemoeller, Mahvish Nagda, Prateek Kolhar, Renee Schneider, Shaojian Zhu, Stephanie Chan, Steve Yadlowsky, Viknesh Sounderajah, and Yannis Assael. Learnlm: Improving gemini for learning, 2025. URL https://arxiv.org/abs/2412.16429.

L. S. VYGOTSKY. *Mind in Society: Development of Higher Psychological Processes*. Harvard University Press, 1978. URL http://www.jstor.org/stable/j.ctvjf9vz4.

David Wood, Jerome S. Bruner, and Gail Ross. The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry*, 17(2):89–100, 1976. doi: 10.1111/j.1469-7610.1976.tb00381.x. URL https://doi.org/10.1111/j.1469-7610.1976.tb00381.x.

Fangchen Yu, Haiyuan Wan, Qianjia Cheng, Yuchen Zhang, Jiacheng Chen, Fujun Han, Yulun Wu, Junchi Yao, Ruilizhen Hu, Ning Ding, Yu Cheng, Tao Chen, Lei Bai, Dongzhan Zhou, Yun Luo, Ganqu Cui, and Peng Ye. HiPhO: how far are (m)llms from humans in the latest high school physics olympiad benchmark?, 2025. URL https://arxiv.org/abs/2509.07894.

Shenghe Zheng, Qianjia Cheng, Junchi Yao, Mengsong Wu, Haonan He, Ning Ding, Yu Cheng, Shuyue Hu, Lei Bai, Dongzhan Zhou, Ganqu Cui, and Peng Ye. Scaling physical reasoning with the PHYSICS dataset. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2–7, 2025*, 2025. URL https://papers.neurips.cc/paper_files/paper/2025/hash/eb5b1b97cd9568cff7b1736c01168fb7-Abstract-Datasets_and_Benchmarks_Track.html.

Yuhang Zhou, Xutian Chen, Yixin Cao, Yuchen Ni, Yu He, Siyu Tian, Xiang Liu, Yunwen Chen, Guangnan Ye, Xipeng Qiu, and Hongfeng Chai. Teach2eval: An interaction-driven llms evaluation method via teaching effectiveness. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 124064–124088, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/c98ef086dc70d528e1c1aa1e66893365-Paper-Conference.pdf.

Algorithm 1 VideoTeach multi-agent protocol

Require: Anchor problem p_0; teacher $\mathcal{T}$; student $\mathcal{S}$; evaluator $\mathcal{E}$; number of teaching turns $n \geq 1$

STAGE 1 & 2: TEST CONSTRUCTION AND PRE-TEST

$(p_1, p_2) \leftarrow \mathcal{E}_{\text{gen}}(p_0)$ ▷ constructed before teaching

for $t \in \{0, 1, 2\}$ **do**

  $a_t^{\text{pre}} \leftarrow \mathcal{S}_{\text{solve}}(p_t)$

  $T_t^{\text{pre}} \leftarrow \mathbb{1}[a_t^{\text{pre}} \text{ is correct}]$

end for

$O_{<1} \leftarrow (), \quad R_{<1} \leftarrow (), \quad L_{<1} \leftarrow ()$ ▷ empty ordered histories

STAGE 3: TEACHING–LEARNING INTERACTION

for $i = 1$ to n **do**

  $(v_i, o_i) \leftarrow \mathcal{T}_{\text{teach}}\,(p_0, a_0^{\text{pre}}, O_{<i}, R_{<i})$

  $(r_i, \ell_i) \leftarrow \mathcal{S}_{\text{learn}}\,(p_0, a_0^{\text{pre}}, R_{<i}, L_{<i}, v_i)$

  $S_{1,i} \leftarrow \mathcal{E}_{\text{L1}}(p_0, v_i)$ ▷ L1 scoring

  $S_{2,i} \leftarrow \mathcal{E}_{\text{L2}}(p_0, a_0^{\text{pre}}, r_i, v_i)$ ▷ L2 scoring

  $O_{<i+1} \leftarrow (O_{<i}, o_i)$

  $R_{<i+1} \leftarrow (R_{<i}, r_i)$

  $L_{<i+1} \leftarrow (L_{<i}, \ell_i)$

end for

STAGE 4 & 5: POST-TEST AND SCORING

$R_{\leq n} \leftarrow R_{<n+1}, \quad L_{\leq n} \leftarrow L_{<n+1}$

for $t \in \{0, 1, 2\}$ **do**

  $a_t^{\text{post}} \leftarrow \mathcal{S}_{\text{solve}}\,(p_0, a_0^{\text{pre}}, L_{\leq n}, R_{\leq n}, p_t)$

  $T_t^{\text{post}} \leftarrow \mathbb{1}[a_t^{\text{post}} \text{ is correct}]$

  $\Delta T_t \leftarrow T_t^{\text{post}} - T_t^{\text{pre}}$

end for

return $\{(S_{1,i}, S_{2,i})\}_{i=1}^{n}; (\Delta T_0, \Delta T_1, \Delta T_2)$

A DATA CURATION AND QUALITY CONTROL

Candidate pool and deduplication. We begin with 24,297 candidate problems drawn from the eight source benchmarks listed in Section 3.2. To reduce cross-source overlap, we encode each problem with a sentence transformer and flag pairs whose cosine similarity exceeds 0.85. We remove 694 near-duplicate candidates before subsequent filtering. The threshold is a conservative operational heuristic rather than a claimed optimum: estimating its precision and recall would require an exhaustively adjudicated set of duplicate and non-duplicate pairs, which is unavailable. We therefore leave formal threshold-sensitivity analysis to a future benchmark release.

Validity and reference-answer verification. We retain only problems that are self-contained, unambiguous, and admit a closed answer. Qwen3-Max performs the initial validity check, removing 4,312 candidates that fail these criteria. It then independently solves each remaining problem and compares its answer with the stored reference answer. When the generated and stored answers disagree, GPT-5.5 acts as an arbiter; cases that remain unresolved are escalated to human review. This verification step removes a further 1,529 candidates. This cascade uses a lightweight model for initial screening and a stronger model or human reviewer only for disagreements.

Stratified sampling. After deduplication and filtering, we sample problems by source benchmark and sub-field to preserve proportional coverage across the five domains. This procedure yields the final 1,176 problems spanning 24 sub-fields. Appendix Tables 3 and 4 report the resulting source- and sub-field-level distributions.

Expert audit. We draw a random 10% sample from the final 1,176 items, yielding 118 audited items after rounding. Three domain experts inspect every sampled item independently using two binary criteria: (i) the problem is self-contained, unambiguous, and admits a single fixed answer; and (ii) the stored reference answer is correct. An item passes only if all three experts judge both

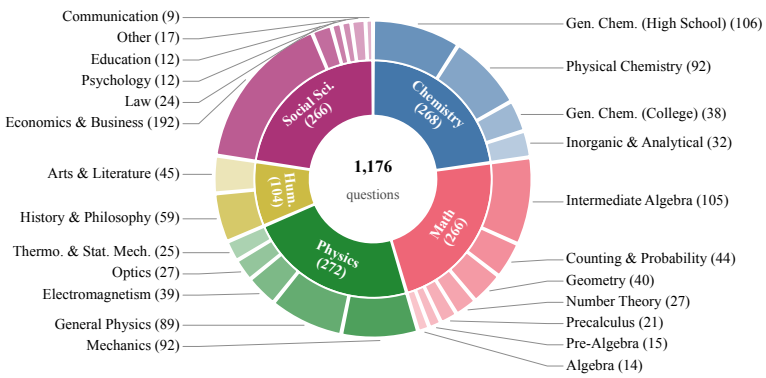

Figure 7: VideoTeach composition: 1,176 anchor problems across five domains (inner ring) and 24 sub-fields (outer ring).

Table 3: Data sources of VideoTeach.

Subject	**Source**	**# Questions**
Chemistry	MMLU	144
	SciBench	124
Math	MATH	266
Physics	UGPhysics	183
	SciEval	89
Humanities	SuperGPQA	91
	MMLU-Pro	13
Social Sciences	EconLogicQA	144
	SuperGPQA	81
	MMLU-Pro	41
Total		**1,176**

criteria satisfied. Any disagreement is conservatively recorded as unqualified rather than resolved by majority vote. Under this unanimity rule, 98% of the audited items pass.

Public-source exposure. The source benchmarks are public and may have appeared in model pretraining. Our item-matched pre-test measures whether the student can answer each item before teaching, so the reported gain is not simply its post-test accuracy. The deliberately moderate student and its low pre-test performance reduce the likelihood that memorized answers dominate the aggregate gain, but they do not constitute a contamination audit or prove that individual items were unseen. We therefore qualify claims accordingly and identify evaluation on newly authored or sequestered items as future work.

B Transfer-Problem Construction and Quality Control

We do not accept problems p_1 and p_2 from a single generation pass. Instead, we use a generate–solve–validate pipeline that separates variant construction from quality control. Given the anchor problem, metadata, knowledge points, and reference solution, GPT-5.5 generates candidate transfer problems under explicit operational constraints specified in the two prompts below: p_1 must preserve the knowledge and procedure while changing only numbers, context, or surface form, whereas p_2 must combine the knowledge with related concepts or alter the problem's structure. Gemini-3-Flash then independently solves each candidate to derive its answer and reasoning, providing a separate check of solvability and answer consistency. Finally, Claude-Opus-4.6 cross-validates the candidate problem, independently derived solution, and intended transfer level, checking for errors, ambiguity,

Table 4: Sub-field distribution of VideoTeach.

Subject	Sub-field	# Questions
Chemistry	Gen. Chem. (High School)	106
	Physical Chemistry	92
	Gen. Chem. (College)	38
	Inorganic & Analytical	32
Math	Intermediate Algebra	105
	Geometry	40
	Counting & Probability	44
	Number Theory	27
	Precalculus	21
	Algebra	14
	Pre-Algebra	15
Physics	Mechanics	92
	General Physics	89
	Electromagnetism	39
	Thermo. & Stat. Mech.	25
	Optics	27
Humanities	History & Philosophy	59
	Arts & Literature	45
Social Sciences	Economics & Business	192
	Law	24
	Psychology	12
	Education	12
	Communication	9
	Other	17
Total		**1,176**

unsolvability, trivial copying, and violations of p_1 and p_2 definitions. Candidates that fail any check are marked as `needs_review` rather than silently retained. This flag occurs for approximately 3% of initial generations; all flagged problems are regenerated and re-evaluated until they pass the full pipeline. This role separation reduces single-model self-validation: one model proposes each variant, a second solves it, and a third verifies its correctness and intended transfer level. A candidate is retained only if it is solvable and satisfies the operational definition of p_1 or p_2.

Finally, three domain experts independently audit a random 10% sample. A problem passes only if all three experts agree that it is well-formed, its reference answer is correct, and its relationship to the anchor matches the intended transfer level: surface transfer for p_1 or structural transfer for p_2; disagreements are counted as failures. Under this unanimity criterion, 98% of the audited items pass.

B.1 Anchor–Surface Transfer–Structural Transfer Triplets

To illustrate the distinction among p_0, p_1, and p_2, we provide ten example triplets—two from each domain.

B.1.1 Mathematics

`M0004`. Anchor (p_0). Mikka wants to order a pizza with two different toppings. He has 8 different toppings to choose from. How many different pizzas could he order?

Surface transfer (p_1). Sarah wants to choose 3 different books to bring on vacation from a shelf of 10 books. How many different combinations of books could she choose?

Structural transfer (p_2). Mikka wants to order a pizza with two or three different toppings. He has 8 different toppings to choose from. How many different pizzas could he order?

M0022. Anchor (p_0). Suppose that there is a right triangle $\triangle ABC$ such that the legs AB and BC have lengths $4\sqrt{2}$ and 2, respectively. What is the length of the median BM?

Surface transfer (p_1). Suppose that there is a right triangle $\triangle PQR$ such that the legs PQ and QR have lengths $3\sqrt{3}$ and 3, respectively. What is the length of the median QM, where M is the midpoint of the hypotenuse PR?

Structural transfer (p_2). In right triangle $\triangle ABC$ with the right angle at B, the median from B to the midpoint M of hypotenuse AC has length 5. If leg $AB = 6$, what is the length of leg BC?

B.1.2 PHYSICS

SCIEVAL_0794. Anchor (p_0). A charge of $25\,\mathrm{C}$ is passing through points A and B on a circuit. If the charge's electric potential changes from $25\,\mathrm{J}$ to $3\,\mathrm{J}$, what is the voltage between points A and B? A. 0.44 Volts B. 1.12 Volts C. 2.20 Volts D. 0.88 Volts Answer:

Surface transfer (p_1). A charge of $40\,\mathrm{C}$ passes through points P and Q in a circuit. If the charge's electric potential energy decreases from $60\,\mathrm{J}$ to $12\,\mathrm{J}$ as it moves from P to Q, what is the voltage between points P and Q? A. 0.80 Volts B. 1.20 Volts C. 1.50 Volts D. 2.40 Volts

Structural transfer (p_2). In a circuit, a charge moves from point X to point Y through a resistor. The voltage between X and Y is $1.5\,\mathrm{V}$, and the charge's electric potential energy at point X is $45\,\mathrm{J}$. If the electric potential energy at point Y is $15\,\mathrm{J}$, what is the amount of charge (in coulombs) that passed through the resistor? A. 10 C B. 20 C C. 30 C D. 45 C

UGPHYS_0001. Anchor (p_0). A small ball with mass m is attached to one end of an inextensible light string, passes through a small hole on the table, and moves on a frictionless horizontal table surface. The string has no friction with the hole. What force is needed at the other end of the string to make the ball perform circular motion with radius r_1 and angular velocity ω_1 around the hole as the center?

Surface transfer (p_1). A puck of mass m is tied to one end of a light string and moves on a smooth horizontal table in a circle around a small hole. The string passes frictionlessly through the hole. If the puck is to move with radius R and angular speed 2ω, what pulling force must be applied to the other end of the string? Express your answer in terms of m, R, and ω.

Structural transfer (p_2). A small ball of mass m moves on a frictionless horizontal table, attached to a light string passing through a hole at the center. A constant force F is applied downward at the other end of the string, so the string tension equals F. If the ball rotates uniformly in a circle of radius r around the hole, what is its angular velocity ω in terms of F, m, and r?

B.1.3 CHEMISTRY

MMLU_CHEM_0128. Anchor (p_0). Each resonance form of the nitrate ion, $\mathrm{NO_3^-}$, has how many sigma and how many pi bonds? Choices: (A) 1 sigma and 2 pi (B) 2 sigma and 1 pi (C) 1 sigma and 1 pi (D) 3 sigma and 1 pi

Surface transfer (p_1). Each resonance form of the carbonate ion, $\mathrm{CO_3^{2-}}$, has how many sigma and how many pi bonds? Choices: (A) 1 sigma and 2 pi (B) 2 sigma and 1 pi (C) 3 sigma and 1 pi (D) 3 sigma and 2 pi

Structural transfer (p_2). Consider all three resonance forms of the nitrate ion, $\mathrm{NO_3^-}$. What is the total number of sigma bonds plus the total number of pi bonds when you add up the bonds across all three resonance structures? Choices: (A) 9 sigma and 3 pi (B) 6 sigma and 3 pi (C) 9 sigma and 1 pi (D) 3 sigma and 3 pi

MMLU_CHEM_0103. Anchor (p_0). Approximately how many half-lives will it take to reduce 10.0 kg of a radioactive substance to 1.0 microgram of that substance? Choices: (A) 8 (B) 13 (C) 29 (D) 34

Surface transfer (p_1). Approximately how many half-lives will it take to reduce 1.0 kg of a radioactive substance to 1.0 nanogram of that substance? Choices: (A) 12 (B) 24 (C) 36 (D) 40

Structural transfer (p_2). A radioactive substance has a half-life of 2.5 hours. Approximately how many hours will it take to reduce 1.0 kg of this substance to 1.0 milligram? Choices: (A) 20 hours (B) 35 hours (C) 50 hours (D) 75 hours

B.1.4 HUMANITIES

SGPQA_3057. Anchor (p_0). What type of comedy, often involving humorous actions by characters and mutual teasing, but lacking moral education? Choices: (A) Light-hearted comedy (B) Slapstick comedy (C) Joyful comedy (D) Parodic comedy (E) Mockery comedy (F) Playful comedy (G) Satirical comedy (H) Whimsical comedy (I) Farce (J) Humorous comedy

Surface transfer (p_1). Which type of comedy is characterized by exaggerated physical humor, absurd situations, and characters engaging in ridiculous antics and mutual pranks, without conveying any serious moral or educational message? Choices: (A) Satirical comedy (B) Farce (C) Comedy of manners (D) Romantic comedy

Structural transfer (p_2). A theater critic describes two plays: Play X uses wit and irony to expose societal vices and human folly with the intent of correcting behavior, while Play Y features characters chasing each other through doors, engaging in pie fights, and pulling pranks with no deeper message intended. Which pair correctly identifies the comedy types of Play X and Play Y respectively? Choices: (A) Farce and Comedy of manners (B) Satirical comedy and Farce (C) Comedy of manners and Satirical comedy (D) Farce and Satirical comedy

MMLUPRO_3595. Anchor (p_0). Which Christian denomination promotes a stripped-down Christianity? Choices: (A) Roman Catholic (B) Lutherans (C) Methodists (D) Greek Orthodox (E) Quakers (F) Seventh-day Adventists (G) Baptists (H) Mormons (I) Puritans (J) Anglicans

Surface transfer (p_1). Which Christian denomination is known for its emphasis on simplicity, rejection of formal clergy, and a minimalist approach to worship without creeds or sacraments? Choices: (A) Presbyterians (B) Quakers (C) Roman Catholics (D) Eastern Orthodox

Structural transfer (p_2). Which of the following pairs correctly matches a Christian denomination with its distinctive characteristic regarding worship practices? Choices: (A) Quakers–stripped-down Christianity with silent worship and no formal clergy (B) Roman Catholics–rejection of all sacraments and liturgical traditions (C) Baptists–belief in the real presence of Christ in the Eucharist (D) Anglicans–complete rejection of hierarchical church structure

B.1.5 SOCIAL SCIENCE

SGPQA_0608. Anchor (p_0). Higher education ( ) refers to the regional composition of higher education institutions. Choices: (A) Institutional arrangement (B) Formal structure (C) Sectoral alignment (D) Family structure (E) Academic organization (F) Geographical distribution (G) Cultural framework (H) Regional structure (I) Hierarchical structure (J) Educational network

Surface transfer (p_1). In higher education studies, the term that refers to the distribution and composition of higher education institutions across different geographic areas is called: (A) Temporal structure (B) Regional structure (C) Disciplinary structure (D) Administrative structure

Structural transfer (p_2). When policymakers analyze the balance between the number of higher education institutions in urban versus rural areas within a country, which structural dimension of higher education are they primarily examining? (A) Disciplinary structure, which concerns the distribution of academic subjects offered (B) Hierarchical structure, which concerns the ranking of institutions by prestige (C) Regional structure, which concerns the geographic distribution of institutions (D) Enrollment structure, which concerns the number of students per institution

MMLUPRO_3217. Anchor (p_0). With respect to the kinked demand curve, which of the following is NOT correct? Choices: (A) There is no price leader. (B) There are relatively equal shares of the market. (C) Price is greater than marginal costs. (D) The demand curve is of one continuous function.

Surface transfer (p_1). Regarding the kinked demand curve model of oligopoly, which of the following statements is INCORRECT? (A) Prices tend to remain stable even when costs change moderately. (B) The marginal revenue curve has a discontinuous gap at the kink. (C) Rivals will match price decreases but not price increases. (D) The model assumes that all firms face a single, uninterrupted demand function.

Structural transfer (p_2). In the kinked demand curve model, the marginal revenue (MR) curve has a vertical gap at the output level corresponding to the kink. Which of the following best explains why price remains stable even when marginal costs change? (A) Firms collude to keep prices fixed regardless of cost changes. (B) As long as the MC curve passes through the gap in the MR curve, profit-maximizing output and price do not change. (C) The government regulates prices in oligopolistic markets. (D) Firms in a kinked demand model always act as price takers.

System prompt

```
You are an independent education evaluator.
Your role is to generate high-quality transfer questions for a given anchor problem.

You will receive the anchor problem together with its solution and knowledge points.
Generate two transfer questions (T1 and T2) that test student mastery and transfer.

IMPORTANT: You MUST output a valid JSON object. Do NOT output only analysis or
    explanation.
Your response must contain a JSON code block with the required format.
```

User prompt

```
Here is an anchor problem with its solution and knowledge points.

## Anchor Problem

{problem}

## Subject / Knowledge Points

Subject: {subject}
Knowledge Points: {knowledge_points}

## Reference Solution

{solution}

## Reference Answer

{answer}

## Your Task

Generate TWO transfer questions based on the problem above. You only need to generate
    the QUESTION TEXT -- do NOT provide answers or solutions (those will be generated
    separately).

- **Transfer 1 (T1, surface transfer)**: Same knowledge point(s) as the anchor problem.
  Only change numbers, context, or surface phrasing.
  Tests basic mastery of the same concept. Must be solvable and verifiable.

- **Transfer 2 (T2, structural transfer)**: Requires combining the knowledge point(s)
  with related concepts, and transforms the problem type/structure.
  Tests whether students can generalize the concept to new situations.

Constraints:

- Difficulty must NOT be significantly harder than the anchor.
- Questions must NOT duplicate the anchor problem.
- CRITICAL: Design questions that have a SHORT, DETERMINISTIC, and AUTOMATICALLY
    VERIFIABLE answer.
  The answer should be: a number, a mathematical expression/formula, a fraction, a
     choice letter (A/B/C/D), or a short symbolic result.
  Do NOT design questions whose answer would be long explanatory text or multi-sentence
      prose.
```

```
  If the anchor problem is qualitative/open-ended, you MUST transform it into a
     quantitative or multiple-choice variant that has a unique correct answer.

Please think through this carefully. After your analysis, provide your final answer in
     the following JSON format:
````json
{
  "T1": {
    "question": "full question text"
  },
  "T2": {
    "question": "full question text"
  }
}
````
```

C Multi-Agent Interaction Protocol and Prompts

The templates below instantiate the controlled pre-test, teaching–learning interaction, and post-test protocol. They also make the information boundaries explicit: the teacher receives the Student's observable pre-test results and responses during interaction but not the Student's private lecture notes, while the Student never receives the teacher's private outlines.

C.1 Pre-test Prompt

All pre-test problems (anchor, surface transfer, and structural transfer problems) use the same template, with only the corresponding problem inserted at `{question}`. For multiple-choice items, the final sentence below is appended.

User prompt

```
You are a student taking an exam. Please solve the problem carefully.

Problem:
{question}

Please reason step by step, then write your final answer inside \boxed{}, in the format
     : \boxed{your answer}

If it is a multiple-choice question, write ONLY the choice letter (e.g. \boxed{A}), do
     NOT repeat the option content.
```

C.2 Teacher Interaction Prompt

All teacher pipelines receive the same problem, subject, the student pre-test results, and prior interaction history. We factor pipeline-dependent generation instructions, output schemas, and rendering constraints into the `{method_specific_prompt}` placeholder. This keeps the teacher's observable student state identical across pipelines while preserving each pipeline's native generation mechanism. For compactness, the templates below use one placeholder; in the implementation, pipeline-specific content may be divided between the system and user messages as `{method_specific_system_prompt}` and `{method_specific_user_prompt}`.

System prompt

```
You are a teacher in an ongoing teaching session. A student has just completed a Pre-
     test on one exam item. Your output will become the teacher's {turn_description}.

{method_specific_prompt}
```

User prompt

```
## Problem

{problem}

Subject: {subject}

## Student Pre-test

Answer: {pretest_answer}
Reasoning: {pretest_reasoning}

{interaction_context}

## Your Task

Generate the next teaching video for the student.

{method_specific_prompt}
```

Turn-dependent context. For the first turn, `{turn_description}` is "initial teaching video after the Pre-test," and `{interaction_context}` is:

User prompt

```
This is the first teaching turn (Turn 1).
```

For later turns, `{turn_description}` is "follow-up teaching video after prior interaction turns," and `{interaction_context}` is:

User prompt

```
## Interaction History ({num_previous_turns} prior turns)

[Turn 1]
What you taught:
{teacher_outline_1}

Student's feedback:
{student_response_1}

...

[Turn {t-1}]
What you taught:
{teacher_outline_t-1}

Student's feedback:
{student_response_t-1}

This is now teaching turn {t}. Generate a supplementary teaching video that addresses
     what the student still does not understand.
```

Pipeline-specific instructions. The placeholder specifies the pipeline-specific prompts: a complete, executable Manim CE program with technical constraints for MONO; a free-form teaching brief with a topic and description for FREE; a structured storyboard JSON with 3–5 fixed sections for SCAF; and a 200–400-word video-generation prompt describing the visuals and instructional content for T2V.

C.3 Student Learning-and-Feedback Prompt

At teaching turn i, the student receives the anchor problem, its pre-test answer and reasoning, and the lecture notes and feedback retained from previous rounds. This history block is empty at the first turn. The current teaching video is supplied as a multimodal input in the same request, followed

by the instruction below. The returned `lecture_note` is stored as the student's private learning memory and used in the post-test, whereas `response_to_teacher` is passed to the teacher in the next round. Both are retained in the Student's post-test context, but the Teacher never receives the full lecture note.

User prompt

```
You are a student attending a tutoring session. Your teacher sends you teaching videos
    to help you understand the problem.

Anchor problem:
{anchor_problem}

Your initial attempt (pre-test):
Answer: {pretest_answer}
Reasoning: {pretest_reasoning}

## Your Lecture Notes from Previous Rounds

[Round 1 Lecture Note]
{previous_lecture_note}
My response to teacher: {previous_response}

...

The teacher has sent you a new teaching video. Watch it carefully, then respond with a
    JSON object containing:

1. "lecture_note": Write detailed lecture notes capturing the key concepts, methods,
    formulas, and step-by-step procedures taught in the video. Be SPECIFIC--include
    exact formulas, derivation steps, key identities, worked examples, and important
    distinctions or common mistakes. Write as if you are creating a reference sheet
    you can use later to solve similar problems.

2. "response_to_teacher": Your honest response to the teacher--questions about things
    you still do not understand, aspects that confused you, or confirmation that the
    explanation was clear. Be specific.

Output ONLY a valid JSON object:
{"lecture_note": "...", "response_to_teacher": "..."}

The video is as follows: {VIDEO}
```

C.4 Post-test Prompts

The post-tests are administered in independent sessions. Each session begins with the anchor problem and the Student's anchor pre-test response, followed by the Student's lecture notes and feedback from all completed rounds; the target problem is appended last. For transfer post-tests, the Student's earlier response to the target transfer problem is not included. Neither template re-inputs the teaching videos or the Teacher's outlines. The round block shown below is repeated for every completed teaching round.

C.4.1 Anchor Problem Post-test

User prompt

```
You are a student who has just finished a tutoring session. Now you need to re-answer
    the anchor problem using what you learned.

Anchor problem:
{anchor_problem}

Your initial attempt on the anchor problem (before tutoring):
Answer: {anchor_pretest_answer}
Reasoning: {anchor_pretest_reasoning}

## Your Lecture Notes from the Tutoring Session

[Round 1 Lecture Note]
```

```
{lecture_note_1}
My feedback: {response_1}

...

Now please answer the anchor question again.

Question:
{anchor_problem}

Apply the concepts and methods you learned from the teaching session. Reason step by
    step and write your final answer inside \boxed{}, in the format: \boxed{your
    answer}
```

C.4.2 Transfer Problem Post-test

User prompt

```
You are a student who has just finished a tutoring session. Now you need to answer a
    related question using what you learned.

Anchor problem (what you were taught):
{anchor_problem}

Your initial attempt on the anchor problem (before tutoring):
Answer: {anchor_pretest_answer}
Reasoning: {anchor_pretest_reasoning}

## Your Lecture Notes

[Round 1 Lecture Note]
{lecture_note_1}
My feedback: {response_1}

...

Now please answer the following transfer problem:

Question:
{T1_or_T2_question}

Apply the concepts and methods you learned from the teaching session. Reason step by
    step and write your final answer inside \boxed{}, in the format: \boxed{your
    answer}
```

D Layer 1: Video Artifact Evaluation

The evaluator scores each teaching turn independently from sampled video frames. The system and user prompts below define the five video artifact dimensions, their ordinal scale, and the required output schema.

D.1 Evaluation Dimensions and Prompt

System prompt

```
You are an expert evaluator for educational video artifact quality. You will assess a
    generated instructional video on five INDEPENDENT dimensions.

CRITICAL RULES:
1. Evaluate each dimension independently. Do NOT let a low score on one dimension
    influence scores on others. A video can score 3 on render integrity but 0 on
    aesthetics.
2. Stay strictly within each dimension's scope. Each dimension states what NOT to
    assess -- those aspects belong to a different evaluation layer.
3. Base your judgment solely on what is visible in the provided frames. Do not
    speculate about content beyond what is shown.
```

```
4. You are NOT evaluating teaching quality, pedagogical effectiveness, or learning
     outcomes. You are evaluating artifact quality only.

Scoring scale (0-3, no neutral midpoint):
  0 = Unusable: severe failure, video is broken for this dimension
  1 = Major issues: noticeable problems that significantly detract
  2 = Acceptable: minor issues, overall usable
  3 = Strong: no significant issues

IMPORTANT: Output ONLY a valid JSON object.
```

User prompt

```
## Problem
{problem}

You will now inspect sampled frames from the teaching video.
Evaluate only the video's artifact quality based on what is visually observable in
     those frames.

## Sampled Video Frames
{sampled_video_frames}

## Evaluation Task: Video Quality (5 dimensions, 0-3 each)

=== Dimension 1: Render Integrity ===
Assess whether code executes and the generated video preserves the intended rendered
     content without corruption or loss.
Check:
- Are expressions, diagrams, and text visibly rendered rather than missing or blank?
- Are there garbled characters, tofu boxes, or encoding/rendering failures?
- Does the video appear complete rather than truncated mid-animation?
DO NOT assess whether the formula, solution step, or diagram is mathematically or
     semantically correct; that belongs to Layer 2 Instructional Correctness.

=== Dimension 2: Notation Legibility ===
Assess whether mathematical expressions, symbols, labels, and diagrams are visually
     legible.
Check:
- Are formulas shown with the correct notation rather than malformed or wrong notation?
- Are mathematical symbols and labels displayed accurately?
- Are diagrams visually expressed with the intended notation or structure?
DO NOT assess: general render success, readability, or pedagogical correctness. Do not
     penalize mathematically wrong but clearly rendered content here; score such errors
      under Layer 2 Instructional Correctness.

=== Dimension 3: Visual Clarity ===
Assess whether the visuals are legible and easy to read.
Check:
- Is font size large enough to read?
- Is there sufficient contrast between text and background?
- Do visual elements overlap or obscure each other?
- Is spacing between elements adequate?
DO NOT assess: aesthetic appeal or information density.

=== Dimension 4: Temporal Coherence ===
Assess visual continuity across the video timeline using the frame sequence.
Check:
- Are scene transitions smooth rather than abrupt?
- Is animation pacing consistent across frames?
- Are there visual discontinuities such as elements disappearing, reappearing, or
     jumping unexpectedly?
DO NOT assess: pedagogical sequencing or teaching quality.

=== Dimension 5: Aesthetics ===
Assess visual appeal of the video.
Check:
- Are colors harmonious?
- Is the composition balanced rather than cluttered or lopsided?
- Is there clear visual hierarchy so important elements stand out?
DO NOT assess: information density or cognitive load.

Output ONLY valid JSON, no markdown fences:

{
  "render_integrity": {
```

```
    "score": 0,
    "rationale": "<one sentence citing specific visual evidence>"
  },
  "notational_legibility": {
    "score": 0,
    "rationale": "<one sentence citing specific visual evidence>"
  },
  "visual_clarity": {
    "score": 0,
    "rationale": "<one sentence citing specific visual evidence>"
  },
  "temporal_coherence": {
    "score": 0,
    "rationale": "<one sentence citing specific visual evidence>"
  },
  "aesthetics": {
    "score": 0,
    "rationale": "<one sentence citing specific visual evidence>"
  }
}

Replace each 0 with your actual score (0-3). Score each dimension INDEPENDENTLY.
```

D.2 Score Normalization and Aggregation

Each of the five dimensions receives one rating in $\{0, 1, 2, 3\}$, mapped to $\{0, 33.3, 66.7, 100\}$. Let $x_{1,d}$ denote the normalized score for dimension $d \in \mathcal{D}_1$, where $|\mathcal{D}_1| = 5$. The Layer 1 score is

$$S_1 = \frac{1}{5} \sum_{d \in \mathcal{D}_1} x_{1,d}.$$

Thus, all production-quality dimensions contribute equally.

E Layer 2: Instructional Process Evaluation

The evaluator scores each teaching turn from the sampled video frames, the student's pre-test response, and the student's feedback. The rubric, prompt, and aggregation rule are reported together below.

E.1 Evaluation Dimensions and Rubric

The rubrics are shown in Table 5.

E.2 Evaluator Prompt

System prompt

```
You are an independent education assessment expert evaluating a SINGLE teaching turn
    using the Layer 2 pedagogy rubric.

You will observe:
- The teaching video frames (key frames extracted from the video)
- The student's pre-test answer (baseline)
- The student's feedback after watching

You do NOT have access to the teacher's outline or script. Evaluate ONLY from what is
    visible in the video frames and the student's observable reactions.

If no video frames are provided, you CANNOT assess the teaching quality -- output all "
    No" scores and set observed_weaknesses to [{"weakness_type": "no_video", "evidence
    ": "No teaching video was available for evaluation."}].

Evaluate this SINGLE teaching turn across 5 dimensions with 15 sub-indicators.
Each sub-indicator: "Yes" / "To some extent" / "No"

Also identify any student weaknesses you observe (concepts the student still doesn't
    understand).
```

Table 5: **Layer 2 rubric for instructional process quality.** The five complementary dimensions provide broad coverage of the instructional process and are grounded in established principles from the learning sciences, including cognitive load, multimedia learning, scaffolding, learner engagement, and motivation.

Dim.	Item	Theory	Evaluation Content
D1 — Instructional Correctness			
	Content Accuracy	Mayer (2009)	Are the knowledge, formulas, and solution steps presented in the video factually correct? This item targets mathematical or conceptual errors that are independent of how the content is rendered or sequenced.
	Logical Coherence	Sweller (2011)	Is the reasoning chain internally consistent and free of contradictions? Earlier claims are not contradicted in later segments; inferential steps are explicitly connected rather than left as unexplained jumps.
	Step Completeness	Mayer (2009)	Are all key solution steps shown? No critical step is silently skipped; each transition from one step to the next is made visible to the learner.
D2 — Cognitive Load Management			
	Coherence	Mayer (2009)	Is extraneous material excluded? The video contains no irrelevant information, distracting visuals, or digressions that do not serve the learning objective.
	Key Information Highlighting	Mayer (2009)	Are essential elements highlighted or visually cued? Key concepts, formulas, and transitions are marked with annotations, colour emphasis, or spatial cues.
	Visual Pacing	Mayer (2009)	Is the video appropriately paced, and do visual and verbal elements appear in temporal contiguity? Complex content is broken into digestible segments; labels appear alongside the relevant element rather than before or after it.
	Schema Construction	Sweller et al. (1998)	Does the video progressively build conceptual understanding (e.g., intuition → formalisation → generalisation) rather than presenting isolated facts or procedures?
D3 — Learning Engagement			
	Prediction Prompts	Chi & Wylie (2014); Bjork (1994)	Does the video pause or prompt before revealing key results, creating an opportunity for the student to predict the outcome or reason through a challenging step—thereby combining active participation with desirable difficulty?
	Active Engagement	Chi & Wylie (2014)	Does the video create sustained opportunities for mental engagement—posing questions, presenting puzzles, or inviting the student to think along—rather than supporting passive watching throughout?
	Encourage Self-Reflection	Flavell (1979)	Does the teacher guide the student to discover their own errors rather than directly correcting them? Feedback is framed as a question or hint that prompts self-diagnosis.
	Provide Constructive Feedback	Hidi & Renninger (2006)	Is feedback constructive and actionable—explaining *why* an answer is incorrect and what to do differently—rather than merely marking it wrong?
D4 — Adaptive Scaffolding			
	Leveling	VYGOTSKY (1978)	Is the difficulty of the video-based explanation calibrated to the student's current level—operating within the zone of proximal development, neither trivially easy nor unreachably hard?
	Visual Adaptive Response	Wood et al. (1976)	When the student is stuck, confused, or makes an error, does the teacher adjust the instructional approach at an appropriate level of detail—neither doing the work for the student nor leaving them without direction? Adjustments may target the logical structure of the explanation (e.g., providing a hint, decomposing into a simpler sub-problem, or offering an alternative line of reasoning) or the visual medium itself (e.g., switching from abstract formulas to concrete diagrams, or from static text to animated step-by-step illustrations).
D5 — Motivation & Curiosity			
	Situational Interest	Hidi & Renninger (2006)	Does the video activate situational interest—through real-world connections, surprising facts, or compelling visuals—that draws the learner into the topic?
	Cognitive Conflict	Piaget (1985); Loewenstein (2007)	Does the video present counterintuitive results, paradoxes, or cases that contradict the student's likely prior assumptions, thereby motivating deeper inquiry to resolve the conflict?

```
IMPORTANT: Output ONLY a valid JSON object.
```

User prompt

```
## Problem
{problem}

## Student Pre-test (baseline)
Answer: {pretest_answer}
Reasoning: {pretest_reasoning}

## Student's Feedback After Watching
{student_feedback}

Watch the teaching video below, then complete the evaluation task that follows.

## Teaching Video (Turn {turn})
{sampled_video_frames}

## Evaluation Task: Per-Turn Rubric (15 items)

### A. Instructional Correctness (3 items)
- content_accuracy: The knowledge, formulas, and solution steps shown in this turn are
    factually correct.
- logical_coherence: The reasoning chain in this turn is internally consistent and free
    of contradictions.
- step_completeness: The key solution steps in this turn are shown rather than silently
    skipped.

### B. Cognitive Load Management (4 items)
- coherence: The turn excludes irrelevant material and distracting content.
- key_information_highlighting: Important concepts, formulas, or transitions are
    clearly highlighted.
- visual_pacing: The pacing is appropriate, and visual/verbal elements appear in
    temporal contiguity.
- schema_construction: The explanation progressively builds understanding rather than
    presenting isolated facts.

### C. Learning Engagement (4 items)
- prediction_prompts: The turn creates opportunities for the student to predict or
    think before a key reveal.
- active_engagement: The turn sustains mental engagement instead of passive watching.
- encourage_self_reflection: The turn guides the student to discover their own errors.
- provide_constructive_feedback: The turn explains why an answer is wrong and what to
    do differently.

### D. Adaptive Scaffolding (2 items)
- leveling: The difficulty is calibrated to the student's current level.
- visual_adaptive_response: When the student is stuck or confused, the teacher adjusts
    the explanation or visual strategy appropriately.

### E. Motivation and Curiosity (2 items)
- situational_interest: The turn activates interest through compelling examples,
    visuals, or connections.
- cognitive_conflict: The turn introduces a counterintuitive result, paradox, or
    conflict that motivates deeper inquiry.

Output JSON:
{
  "rubric_scores": {
    "instructional_correctness": {
      "content_accuracy": "Yes" or "To some extent" or "No",
      "logical_coherence": "Yes" or "To some extent" or "No",
      "step_completeness": "Yes" or "To some extent" or "No"
    },
    "cognitive_load_management": {
      "coherence": "Yes" or "To some extent" or "No",
      "key_information_highlighting": "Yes" or "To some extent" or "No",
      "visual_pacing": "Yes" or "To some extent" or "No",
      "schema_construction": "Yes" or "To some extent" or "No"
    },
    "learning_engagement": {
      "prediction_prompts": "Yes" or "To some extent" or "No",
      "active_engagement": "Yes" or "To some extent" or "No",
      "encourage_self_reflection": "Yes" or "To some extent" or "No",
      "provide_constructive_feedback": "Yes" or "To some extent" or "No"
    },
    "adaptive_scaffolding": {
      "leveling": "Yes" or "To some extent" or "No",
      "visual_adaptive_response": "Yes" or "To some extent" or "No"
```

```
      },
      "motivation_curiosity": {
        "situational_interest": "Yes" or "To some extent" or "No",
        "cognitive_conflict": "Yes" or "To some extent" or "No"
      }
    },
    "observed_weaknesses": [
      {"weakness_type": "...", "evidence": "..."}
    ]
  }
```

E.3 Score Normalization and Aggregation

Each rubric item is rated as *No*, *To some extent*, or *Yes*, mapped to 0, 50, and 100. Let $x_{2,d,i}$ denote the normalized score of item i in dimension d, and let $n_{2,d}$ be the number of items in that dimension. We first compute

$$s_{2,d} = \frac{1}{n_{2,d}} \sum_{i=1}^{n_{2,d}} x_{2,d,i},$$

and then average the five dimension scores:

$$S_2 = \frac{1}{5} \sum_{d \in \mathcal{D}_2} s_{2,d}.$$

This hierarchy gives every instructional dimension equal weight regardless of its number of rubric items.

F Layer 3: Learning Outcomes Evaluation

Layer 3 measures learning gains separately at the three transfer distances. The pre-test and post-test templates are reported in Appendix C; this section defines correctness assignment and aggregation.

F.1 Correctness Assignment

For each problem p_t, $t \in \{0, 1, 2\}$, the evaluator assigns binary correctness to the complete pre-test and post-test responses:

$$T_t^{\text{pre}}, T_t^{\text{post}} \in \{0, 1\}.$$

Post-test problems are solved in independent sessions, with no state carried across problems.

F.2 Learning-Gain Computation

The item-level gain at transfer level t is

$$\Delta T_t = T_t^{\text{post}} - T_t^{\text{pre}}.$$

Accordingly, ΔT_0, ΔT_1, and ΔT_2 measure gains on the anchor problem, surface transfer, and structural transfer, respectively.

F.3 Aggregation and Reporting

For N evaluated teaching cases, we report the mean gain at each transfer level,

$$\overline{\Delta T}_t = \frac{1}{N} \sum_{j=1}^{N} \Delta T_t^{(j)},$$

expressed in percentage points.

G VIDEO-INPUT CONFIGURATION

Video input. For Layers 1 and 2, the evaluator received each video as a chronologically ordered sequence of JPEG frames sampled at two-second intervals. For a video with F frames and frame rate r, we define its duration as

$$D = \frac{F-1}{r}.$$

We sample frames at timestamps

$$t_i = \min(2i, D), \qquad i = 0, \ldots, \left\lceil \frac{D}{2} \right\rceil,$$

with corresponding frame indices

$$k_i = \min\left(F-1, \left\lfloor r t_i + \frac{1}{2} \right\rfloor\right).$$

This procedure includes the first and final frames and produces

$$N = \left\lceil \frac{D}{2} \right\rceil + 1$$

sampled frames for a video of duration D. Frames were encoded as JPEG images at quality 75 and provided to the evaluator in temporal order with their timestamps.

We applied the same sampling procedure independently to every video in the single-turn and five-turn experiments. In the five-turn setting, each teaching turn was represented by a separate timestamped frame sequence; videos were not resampled to a fixed number of frames across turns.

Audio. All generated videos were silent. The evaluator therefore received only sampled frames, while the student received the silent MP4. We performed no audio extraction, transcription, speech recognition, resampling, or normalization.

H EXPERIMENT ANALYSIS

H.1 WHY LAYER 3 ALONE IS INSUFFICIENT

Layer 3 measures the primary outcome of interest—whether the student's performance improves after instruction—but is insufficient as the sole evaluation criterion. In our results, gains are largest on the anchor problem, while surface- and structural-transfer gains remain small and non-monotonic. A low transfer gain cannot distinguish among an unusable video artifact, an ineffective instructional process, and a student that fails to generalize.

Layers 1 and 2 provide this missing diagnosis. On the common-success subset, for example, T2V achieves the highest temporal coherence (67.4) but scores only 14.5 in notation legibility and 21.0 in overall instructional process quality. These scores identify artifact and instructional-process weaknesses that a learning-gain scalar cannot localize. Conversely, high artifact and instructional process quality do not guarantee larger learning gains: FREE leads L1 and L2, whereas MONO achieves the largest mean Layer 3 gain, and within-system correlations of mean learning gain with L1 and L2 remain near zero.

The three layers therefore capture distinct, non-substitutable aspects of instructional-video performance. Layer 1 evaluates the usability and presentation of the video artifact, Layer 2 evaluates the quality of the instructional process, and Layer 3 measures pre-to-post learning gains across the anchor, surface-transfer, and structural-transfer problems. Reporting all three distinguishes artifact failures, instructional weaknesses, and limited transfer, providing both outcome measurement and actionable diagnosis.

H.2 EVALUATOR AGREEMENT ON PIPELINE RANKINGS

We assess evaluator sensitivity by holding the generated videos and teacher–student interaction traces fixed and rescoring the same 529 common-success cases with GPT-5.5, Claude-Opus-4.7, and

Table 6: Pairwise agreement between evaluators on per-example pipeline rankings, measured by Kendall's τ_b. Higher values indicate stronger agreement.

Evaluator Pair	L1: Video Artifact	L2: Instructional Process
GPT–Claude	0.639	0.619
GPT–Gemini	0.626	0.601
Claude–Gemini	0.630	0.671

	L1: Video Artifact						L2: Instructional Process						L3: Learning Outcomes			
System	Rend.	Notat.	Clar.	Temp.	Aesth.	**Avg**	InstCorr.	CogLoad	Engage	Scaffold.	Motiv.	**Avg**	ΔT_0	ΔT_1	ΔT_2	**Avg**
(a) Full Set; *(N=1,176; a system scores 0 on cases where it fails to produce a video)*																
NO-VIDEO	-	-	-	-	-	-	-	-	-	-	-	-	-1.7	-1.5	-2.0	-1.7
MONO	59.4	**76.1**	**48.1**	50.4	**53.9**	57.6	**66.6**	**66.6**	26.3	**32.1**	43.3	47.0	**+6.3**	**+1.1**	+0.6	**+2.7**
FREE	38.7	38.1	23.4	28.3	30.8	31.9	31.7	32.9	11.4	12.1	23.4	22.3	+2.0	−0.4	−0.7	+0.3
SCAF	**78.0**	67.8	46.4	52.3	47.4	**58.4**	64.4	63.3	**42.3**	30.0	**50.6**	**50.1**	+3.7	+0.0	−0.3	+1.1
T2V	35.4	14.1	44.7	**67.6**	46.7	41.7	25.8	31.4	7.2	11.9	27.8	20.8	+1.1	−1.1	**+1.2**	+0.4
(b) Common-Success Subset; *(N=529; cases where* all *systems produced a video)*																
NO-VIDEO	-	-	-	-	-	-	-	-	-	-	-	-	-1.3	-1.7	-2.3	-1.8
MONO	64.5	**82.6**	**52.9**	55.1	58.2	62.7	71.9	72.1	28.1	**34.0**	46.7	50.6	**+8.3**	+2.8	−0.9	**+3.4**
FREE	**80.7**	79.5	48.5	58.9	**64.2**	**66.4**	**77.7**	**80.2**	28.0	25.5	49.2	**52.1**	+5.5	+0.9	+0.8	+2.4
SCAF	77.5	68.5	46.0	52.1	46.3	58.1	64.4	63.3	**42.2**	30.0	**50.4**	50.1	+4.0	**+4.5**	−0.8	+2.6
T2V	36.3	14.5	44.8	**67.4**	46.8	42.0	26.6	31.5	6.6	12.0	28.1	21.0	+0.8	−0.8	**+1.2**	+0.4

Table 7: **Metric-level results across the three evaluation layers.** Panel (a) reports full-set results for all 1,176 cases, with generation failures receiving zero. Panel (b) reports the 529 common-success cases for which all four video-generation systems produced a video. NO-VIDEO repeats the pre-to-post assessment without videos and appears only in L3. L1/L2 scores range from 0 to 100, while L3 reports percentage-point gains on the anchor problem, surface transfer, and structural transfer. Higher is better. Shaded columns show unweighted within-layer averages; the L3 average is computed over ΔT_0, ΔT_1, and ΔT_2. Boldface marks the best result within each panel. L1 abbreviations are Rend. = Render Integrity, Notat. = Notation Legibility, Clar. = Visual Clarity, Temp. = Temporal Coherence, and Aesth. = Aesthetics. L2 abbreviations are InstCorr. = Instructional Correctness, CogLoad = Cognitive Load Management, Engage = Learning Engagement, Scaffold. = Adaptive Scaffolding, and Motiv. = Motivation and Curiosity.

Gemini-3-Pro. For each case and evaluation layer, each evaluator ranks the four video-generation pipelines, and we compute pairwise agreement using Kendall's τ_b. Agreement ranges from 0.626 to 0.639 for video artifact quality and from 0.601 to 0.671 for instructional process quality (Table 6). Claude-Opus-4.7 and Gemini-3-Pro agree most strongly on instructional process quality ($\tau_b = 0.671$), while GPT-5.5 and Claude-Opus-4.7 show the highest agreement on video artifact quality ($\tau_b = 0.639$). Pipeline rankings are therefore broadly stable across evaluators, although evaluator-specific variation remains.

H.3 Agreement between human learner and the student agent

We replace the student agent with human learners to test whether model-based evaluation reflects human learning outcomes. We sample 50 problems from the common-success subset, with 10 from each domain. The subset is used only for problem selection. We do not reuse previously generated videos because each teaching video is conditioned on the learner's anchor pre-test response and designed to address the difficulties revealed by that response. Instead, the assigned teacher generates a new video for each participant–problem condition.

We recruit 12 participants, each of whom completes all 50 problem-level teaching cases, yielding 600 human evaluations. For each problem, the participants are divided equally among the four teacher systems, so that every problem–teacher condition is evaluated by three independent human learners. Assignments are permuted across problems such that each participant encounters all four teacher systems approximately equally often, but never encounters multiple teacher systems for the same problem.

	L1: Video Artifact						L2: Instructional Process						L3: Learning Outcomes Δ(pp)			
Teacher	Rend.	Notat.	Clar.	Temp.	Aesth.	**Avg**	InstCorr.	CogLoad	Engage	Scaffold.	Motiv.	**Avg**	ΔT_0	ΔT_1	ΔT_2	**Avg**
GPT-5.5	73.2	84.9	56.0	65.8	72.9	70.5	**95.0**	90.4	**50.1**	**43.7**	**67.4**	**69.3**	**+9.0**	−0.3	−4.0	+1.6
Gemini-3-Pro	76.9	**89.9**	68.9	**70.6**	75.0	76.3	87.6	90.9	30.0	36.8	54.4	59.9	+6.7	−0.3	**−1.0**	**+1.8**
Claude-Opus-4.7	**80.2**	85.9	**75.1**	68.8	**85.1**	**79.0**	82.9	**92.2**	30.8	34.6	56.3	59.3	+6.3	**+2.0**	−4.0	+1.4

Table 8: **Teacher-backbone comparison across the three evaluation layers.** Results are reported for $N = 300$ single-turn MONO cases on VideoTeach-Lite, with the Student, Evaluator, and problems fixed; generation failures receive zero. L1/L2 scores range from 0 to 100, while L3 reports percentage-point pre-to-post gains on the anchor problem, surface transfer, and structural transfer. Higher is better. Shaded columns show unweighted within-layer averages, and boldface marks the best result. Claude-Opus-4.7 achieves the highest artifact quality (L1: 79.0), whereas GPT-5.5 achieves the highest instructional process quality (L2: 69.3). Anchor-problem gains range from $+6.3$ to $+9.0$ pp, surface-transfer gains from -0.3 to $+2.0$ pp, and structural-transfer gains remain negative (-4.0 to -1.0 pp). L1 abbreviations are Rend. = Render Integrity, Notat. = Symbol and Notation Legibility, Clar. = Visual Clarity, Temp. = Temporal Coherence, and Aesth. = Aesthetics. L2 abbreviations are InstCorr. = Instructional Correctness, CogLoad = Cognitive Load Management, Engage = Learning Engagement, Scaffold. = Adaptive Scaffolding, and Motiv. = Motivation and Curiosity.

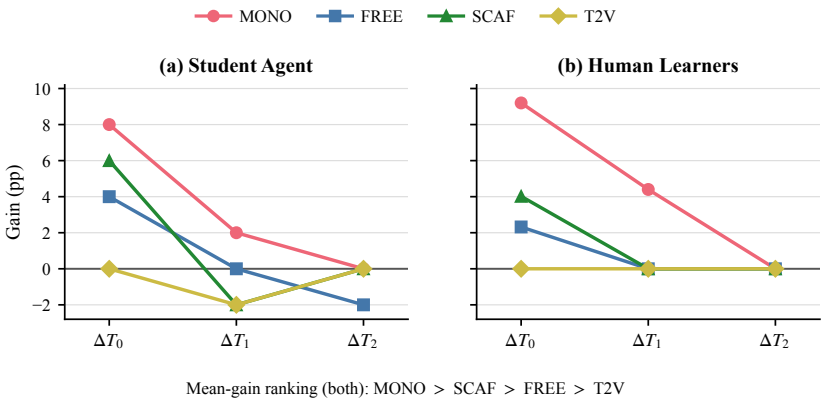

Figure 8: Human–Student-Agent agreement on learning transfer. Panels report mean pre-to-post gains on the anchor problem (ΔT_0), surface transfer (ΔT_1), and structural transfer (ΔT_2). Human gains are averaged over three participants per problem–teacher condition and then macro-averaged over the 50 sampled problems.

For each case, the participant first completes the fixed pre-test on p_0–p_2. Following the VideoTeach protocol, the assigned teacher receives only the anchor problem and the participant's response to p_0 and generates a new single-turn instructional video conditioned on that response. The participant watches the video and then completes the same fixed post-test without receiving answer keys or correctness feedback. For each problem–teacher condition, we first average the gains of its three participants and then macro-average across the 50 problems.

As shown in Figure 8, human learners and the student agent produce the same aggregate system ranking: MONO > SCAF > FREE > T2V. Both also exhibit gains that attenuate from the anchor problem to farther transfer, providing preliminary evidence of agreement in aggregate system ranking and transfer trends within this sample. However, negative gains occur only for the student agent, indicating that this agreement does not imply individual-level fidelity.

H.4 PIPELINES EXHIBIT DISTINCT DOMAIN-SPECIFIC STRENGTHS

Figure 9 shows different domain profiles across methods. On common-success cases, FREE achieves the highest video artifact quality in every domain but leads instructional process quality only in Mathematics. SCAF instead achieves the highest instructional process quality in Chemistry, Hu-

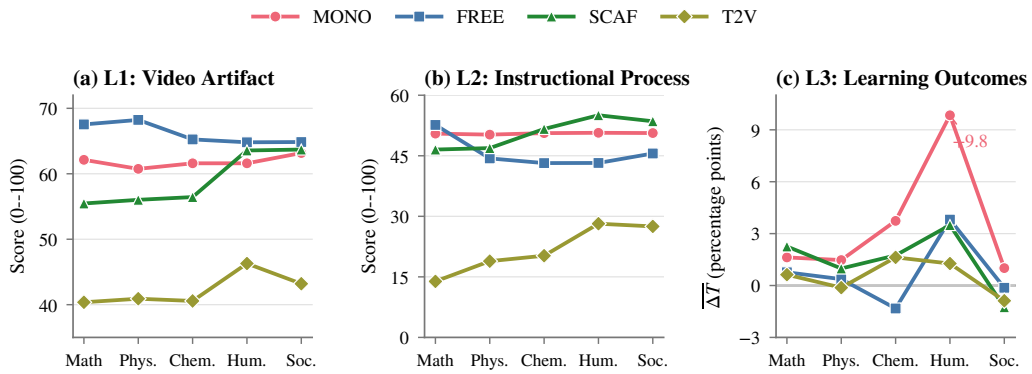

Figure 9: **Domain-wise results on VideoTeach-Lite.** Panels (a)–(c) report video artifact quality (L1), instructional process quality (L2), and mean learning gain (L3), respectively. L1/L2 use common-success cases; L3 averages ΔT_0, ΔT_1, and ΔT_2 in percentage points. Domain abbreviations are Math = Mathematics, Phys. = Physics, Chem. = Chemistry, Hum. = Humanities, and Soc. = Social Sciences.

manities, and Social Sciences. MONO maintains relatively stable artifact and process scores across domains and achieves the largest mean learning gain in four of five domains, reaching $+9.8$ pp in Humanities. Qualitative inspection suggests that many of these gains occur on factual questions about people, works, or dates, consistent with problem-specific recall rather than broadly transferable understanding. T2V ranks last in video artifact and instructional process quality across all domains, suggesting that, relative to the code-based pipelines, end-to-end pixel-space generation provides less reliable control over on-screen text and explanation structure.

H.5 Case Study: Repeated Anchor Refinement without Transfer

We use one case to illustrate how additional teaching turns can improve performance on the anchor problem without producing reliable transfer. The three evaluation problems are as follows.

Anchor problem (p_0). Let

$$P_0(x) = x^3 + 313x^2 - 77x - 8.$$

For integers $n \geq 1$, define

$$P_n(x) = P_{n-1}(x - n).$$

What is the coefficient of x in $P_{20}(x)$?

Gold answer: 763.

Surface-transfer problem (p_1). Let

$$P_0(x) = x^3 + 200x^2 - 50x - 7.$$

For integers $n \geq 1$, define

$$P_n(x) = P_{n-1}(x - n).$$

What is the coefficient of x in $P_{10}(x)$?

Gold answer: -12975.

Structural-transfer problem (p_2). Let

$$P_0(x) = x^3 + 313x^2 - 77x - 8.$$

For integers $n \geq 1$, define

$$P_n(x) = P_{n-1}(x - n).$$

Turn	Teacher focus	p_0	p_1	p_2
1	Combines the recursive shifts into $P_{20}(x) = P_0(x - 210)$, expands the polynomial, and demonstrates the complete calculation of the coefficient of x as 763.	×	×	×
2	Directly corrects the student's previous answer of -881, explains why the incorrect recurrence fails, derives $P_n(x) = P_0\left(x - \frac{n(n+1)}{2}\right)$, and recomputes the anchor answer.	✓	×	×
3	Introduces a different polynomial example, derives the general formula $[x]P_n = 3S^2 - 2aS + b$, and then applies the same formula again to the anchor problem.	×	×	×
4	Provides another worked example with different polynomial coefficients, rehearses the composed-shift procedure, and returns to the anchor calculation of $[x]P_{20}$.	✓	×	×
5	Adds a further example and explicitly warns against omitting the $3S^2$ term, then restates the general formula for the coefficient of x and recomputes the anchor answer as 763.	✓	×	×

Table 9: **Case study of a turn-by-turn evolution.** The teacher-focus column is a summary of literal on-screen content in the generated videos. Correctness is determined by exact normalized agreement with the gold answer. Although later turns introduce new examples and a general formula, all five turns remain centered on extracting the coefficient of x. The student becomes more reliable on the taught problem p_0, but never answers either transfer problem correctly.

What is the coefficient of x^2 in $P_{20}(x)$?

Gold answer: -317.

The surface-transfer problem preserves the requested coefficient but changes the initial polynomial and the number of recursive shifts. The structural-transfer problem preserves the anchor setup but changes the target quantity from the coefficient of x to that of x^2. Table 9 traces the teacher's instructional focus and the student's correctness after each teaching turn. This case illustrates that additional turns can stabilize an anchor-specific solution procedure without yielding reliable transfer: despite progressively correcting, exemplifying, and generalizing the extraction of $[x]$, the student improves only on p_0 and never succeeds when applying the same operation under new parameters in p_1 or adapting it to a different target coefficient in p_2.

H.6 Why Do the Three Evaluation Layers Diverge?

Layer 1 video artifact quality and Layer 2 instructional process quality show only weak linear associations with Layer 3 learning gains. To investigate why, we manually inspect cases that span different combinations of upstream quality and learning outcomes. Figure 10 pairs a representative video frame with the corresponding artifact, instructional-process, and learner-outcome measurements.

Artifact quality can diverge from transferable learning. Two contrasting cases illustrate why artifact quality shows little linear association with learning gains (Figure 10). In `SGPQA_4248`, sharp text, a stable layout, and clean answer highlighting yield high artifact quality (L1: 93.3), but the video provides little explanation of why "Xie Lü Duwei" is correct or how to distinguish it from the alternatives, resulting in low instructional process quality (L2: 28.3). This weakness becomes visible on the surface-transfer problem p_1. The title appears as option D in the anchor problem p_0 but moves to option C in p_1. Although the student answers C correctly before instruction, it changes its answer to D afterward. Because the title remains unchanged while its option position changes, this transition suggests that the student retained the highlighted option label rather than the title itself, resulting in negative transfer. The reverse pattern appears in `SGPQA_1904`: overlapping notation reduces artifact quality (L1: 40.0), but the video communicates the invariant rule that the augmented second occurs between scale degrees 6 and 7, yielding higher instructional process quality (L2: 70.8) and correcting the student's answer on p_1. Together, the cases show that artifact quality captures the usability and presentation of the video, but not whether the instruction conveys the target knowledge in a form that supports transfer.

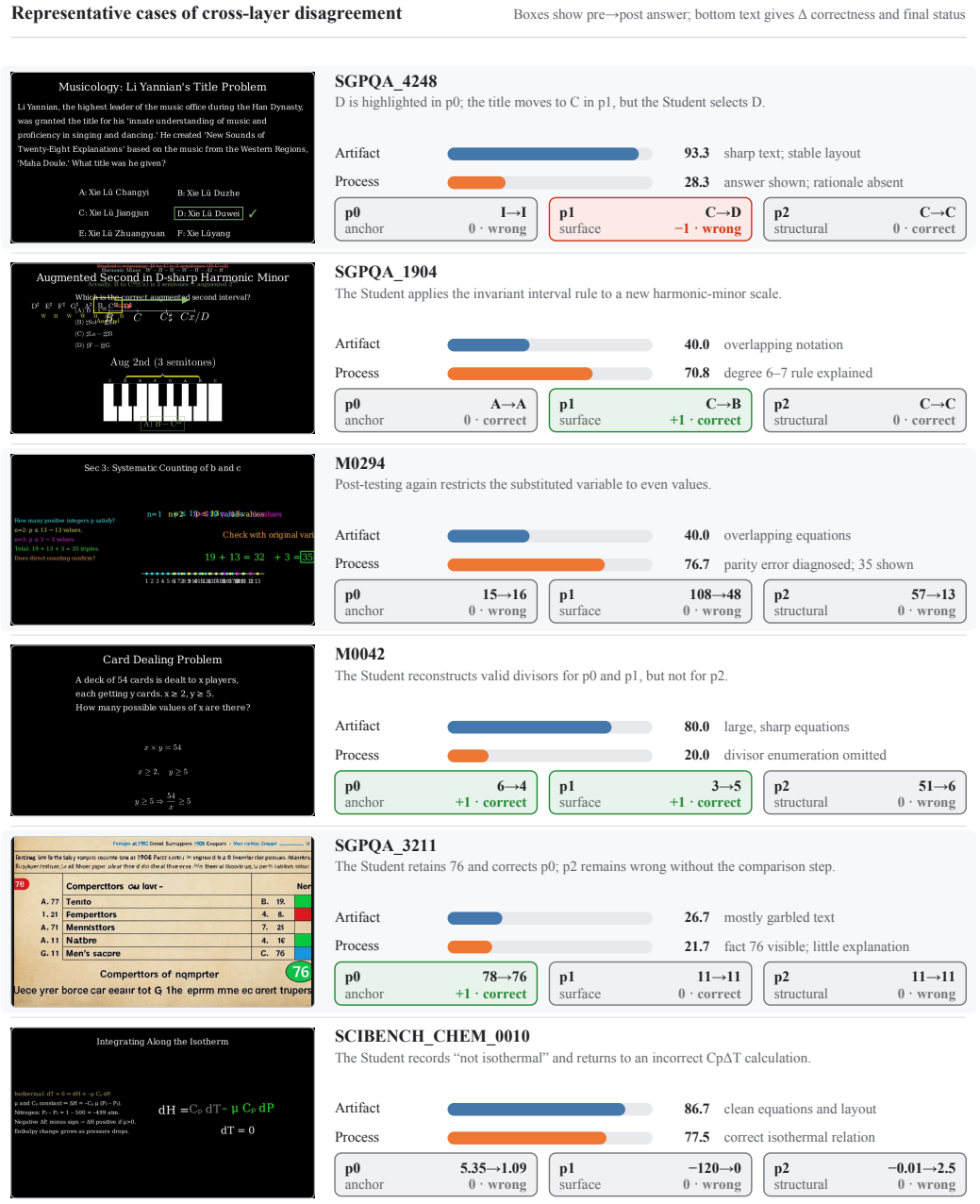

Figure 10: **Representative cases of cross-layer disagreement.** Each row pairs a representative video frame with its Layer 1 video-artifact score, Layer 2 instructional-process score, and Layer 3 outcomes on the anchor problem p_0, surface-transfer problem p_1, and structural-transfer problem p_2. The text beside each upstream score summarizes the dominant qualitative evidence. Each outcome box reports the student's pre-to-post answer, correctness change, and final correctness: $+1$ denotes incorrect-to-correct, 0 no change, and -1 correct-to-incorrect. For example, in `SGPQA_3211`, the visible fact "76" corrects the anchor answer, but the student does not perform the additional comparison required by p_2, which remains incorrect. Representative frames illustrate the principal source of each disagreement; the exact Layer 1 and Layer 2 scores are aggregated over the full multi-dimensional evaluation. Cases are selected for diagnostic contrast rather than prevalence.

High instructional-process scores can mask unresolved misconceptions. Layer 2 rewards instruction that is correct, relevant, and pedagogically structured, but a high aggregate score does not necessarily indicate that the teacher has resolved the specific misconception underlying the student's answer. In `M0294`, the video identifies the parity error and presents the correct count, yielding high instructional process quality (L2: 76.7). However, the instruction largely states the corrected rule and proceeds to the enumeration. The student's subsequent error suggests that the explanation does not sufficiently consolidate why substituting $c = 2p$ already enforces the evenness of c, so that p itself need not be even. Without explicitly contrasting these two constraints or requiring the student to reapply the distinction, the original parity restriction remains operative: the student again enumerates only even values of p, leaving p_0, p_1, and p_2 incorrect.

A related omission appears in `SCIBENCH_CHEM_0010`. The video presents the correct isothermal relation and receives high artifact quality (L1: 86.7) and instructional process quality (L2: 77.5). However, the derivation does not make the applicability boundary of the familiar $C_p\Delta T$ calculation sufficiently explicit. In particular, the instruction needs to emphasize that $dT = 0$ removes the temperature term but does not eliminate the pressure-dependent contribution to enthalpy for a real gas. The student subsequently misclassifies the process and returns to $C_p\Delta T$, leaving all three problems incorrect. Thus, the explanation provides the correct formula but does not sufficiently displace the competing procedure that produced the student's error.

The reverse pattern in `M0042` further shows that not all instructional omissions are equally consequential. Although the video omits the final divisor enumeration and receives low instructional process quality (L2: 20.0), the preceding equation and bounds provide enough information for the student to reconstruct that step, improving on p_0 and p_1, though not on p_2. Together, these cases suggest that learning gains depend not only on the overall quality of the instructional process, but on whether the teacher targets the student's operative misconception and makes the corrected procedure independently applicable.

A salient answer cue improves anchor-problem performance but not transfer. `SGPQA_3211` receives low artifact quality (L1: 26.7) and instructional process quality (L2: 21.7): most frames contain garbled or overlapping text, and the video provides little coherent explanation. Nevertheless, the answer value "76" remains prominently visible throughout the video and is recorded in the student's notes. The student subsequently changes its answer on the anchor problem p_0 from B (78) to the correct option G (76). This gain is therefore consistent with reproduction of a salient answer cue rather than acquisition of the underlying knowledge. No corresponding transfer gain is observed.