EduStory: A Unified Framework for Pedagogically-Consistent Multi-Shot STEM Instructional Video Generation

Xinyi Wu [1 2] **Jayant Teotia** [1] **Shuai Zhao** [1] **Erik Cambria** [1]

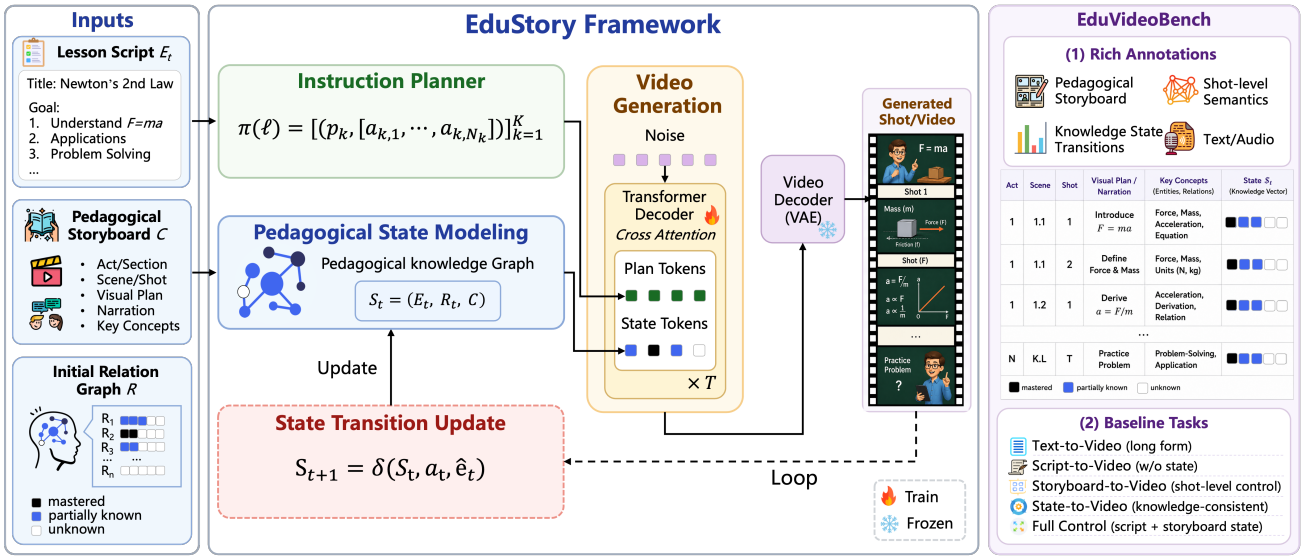

Figure 1. EduStory: A Structured Framework for Knowledge-Consistent Long-Form Educational Video Generation. This figure illustrates the EduStory framework, which integrates pedagogical state modeling, script-guided structured control, and learning-oriented evaluation to enable controllable multi-shot video generation. The pipeline emphasizes persistent knowledge state tracking and structured constraints to ensure narrative coherence and alignment with instructional objectives.

Abstract

Long-horizon video generation has advanced in visual quality, yet existing methods still struggle to maintain knowledge consistency and coherent pedagogical narratives across multi-shot instructional videos, especially in STEM domains. To address these challenges, we propose **EduStory**, a unified framework for reliable instructional video generation. EduStory integrates pedagogical state modeling to track persistent knowledge states, script-guided structured control to organize multi-shot narratives, and learning-oriented evaluation metrics to assess knowledge fidelity and constraint satisfaction. To support rigorous evaluation, we further introduce EduVideoBench, a diagnostic benchmark with multi-granularity annotations, including pedagogical storyboards, shot-level semantics, and knowledge state transitions, together with baseline tasks for controllable instructional video generation. Extensive experiments demonstrate that domain-aware state modeling and structured control substantially reduce narrative breakdown and improve alignment with instructional intent. These results highlight the significance of domain-specific structural constraints and tailored benchmarks for advancing reliable, controllable, and also trustworthy long-horizon video generation.

1. Introduction

Long-horizon video generation has seen remarkable progress, with recent models capable of producing visually coherent clips lasting tens of seconds (Peebles & Xie, 2023; Zheng et al., 2024; Yang et al., 2024). Yet a demanding real-world use case remains largely unsolved: *multi-shot*

[1]Nanyang Technological University [2]Shanghai Jiao Tong University.

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026.

STEM instructional video generation, where a model must maintain strict knowledge consistency across several minutes of generated content. Unlike cinematic storytelling, instructional videos impose hard correctness constraints that are binary and domain-specific. For example, a formula introduced in shot 2 must reappear symbol-for-symbol identically in shot 5, and a force diagram established in the introduction must not silently contradict itself during a later derivation (Jia et al., 2025b). Current models, which are primarily optimized for visual fluency rather than knowledge fidelity, often suffer from *knowledge drift*: entities mutate, formulae lose coefficients, and logical sequences collapse, potentially rendering the generated content educationally misleading or harmful (Wan et al., 2025). To address this gap, we make three contributions as below:

- **EduStory**, a framework that treats instructional video generation as a stateful, constraint-aware process rather than open-ended sequence modeling.
- **EduVideoBench**, the first benchmark specifically designed to diagnose knowledge consistency and pedagogical alignment in multi-shot video generation.
- Preliminary experiments demonstrate that incorporating domain-aware state modeling and constraint verification significantly mitigates narrative breakdown and improves alignment with instructional intent, even when using a lightweight base generator.

2. Related Work

Long-horizon video generation. Recent models such as Open-Sora (Zheng et al., 2024), CogVideoX (Yang et al., 2024), and StreamingT2V (Henschel et al., 2025) extend video generation to tens of seconds through auto-regressive or hierarchical architectures. While these models improve temporal coherence, they lack mechanisms for maintaining domain-specific semantic consistency, which is a requirement orthogonal to visual quality.

Structured and controllable generation. Script-guided generation (Kondratyuk et al., 2024; Sun et al., 2024b;a) and storyboard-conditioned approaches (Liu et al., 2025) decompose long videos into manageable chunks but do not model the *knowledge state* that must persist across segments. Our Instruction Planner builds on this spirit while adding formal state semantics.

Video generation evaluation. Benchmarks such as EvalCrafter (Liu et al., 2024), T2VQA (Wu et al., 2024), and VideoPhy (Bansal et al., 2025) assess visual quality, temporal coherence, and physical plausibility (Kou et al., 2024). None targets the knowledge fidelity and pedagogical structure alignment that define high-quality instructional content. EduVideoBench fills this gap with domain-aware metrics.

Robust AI. Robust AI studies how to ensure model reliability under perturbations, distribution shifts, and safety-critical deployment conditions (Zhao et al., 2024). Early work revealed the vulnerability to adversarial examples (Goodfellow et al., 2015), while RobustBench established standardized evaluation for robustness across models and defenses (Croce et al., 2021). UniFLE (Zhao et al., 2026) has demonstrated promising progress in enhancing the safety of LLMs, especially in mitigating weight-poisoning backdoor attacks. Recent studies further extend robustness to knowledge distillation, few-shot learning, and vision-language models (Dong et al., 2026a; 2024; 2026b; 2025a;c;b). Dong *et al.* made inspiring progress in robust few-shot learning by co-distilling similarity and concept learners, improving robustness under limited supervision (Dong et al., 2024). In contrast, we study a complementary form of robustness: preserving knowledge consistency, symbolic correctness, and pedagogical state transitions across video generation.

3. The EduStory Framework

As shown in Fig. 1, inspired by (Jia et al., 2025a; Wu et al., 2024; 2026), EduStory frames instructional video generation as a *pedagogical state machine* with three tightly coupled components: an Instruction Planner, a State-Conditioned Generator, and a Constraint Verifier.

3.1. Pedagogical State Modeling

At shot t, we define the *pedagogical state*:

$$S_t = \big(E_t,\ R_t,\ \mathcal{C}\big), \tag{1}$$

where E_t is the set of *knowledge entities* introduced through shot t (*e.g.*, $\{\text{force},\ F{=}ma,\ \text{acceleration}\}$); $R_t \subseteq E_t \times E_t \times \mathcal{L}$ is a typed relation graph with label set $\mathcal{L} = \{\textsc{causes}, \textsc{quantifies}, \textsc{derives}, \textsc{instantiates}\}$ encoding logical and physical dependencies; and $\mathcal{C}$ is a domain-specific *constraint set* (*e.g.*, equation balance, unit consistency, directional conventions) that is invariant throughout the video.

State evolves through a deterministic transition:

$$\delta\big(S_t,\ a_t\big) = S_{t+1}, a_t \in \mathcal{A}, \tag{2}$$

where $\mathcal{A}$ is a finite set of *pedagogical actions* (Table 1). Each action specifies precisely which entities and relations are added to the state, making the knowledge accumulation process fully traceable.

3.2. Instruction Planner

The Instruction Planner π maps a plain-text lesson description ℓ to a two-level hierarchical shot plan:

$$\pi(\ell) = \big[\big(p_k,\ [a_{k,1}, \ldots, a_{k,N_k}]\big)\big]_{k=1}^{K}, \tag{3}$$

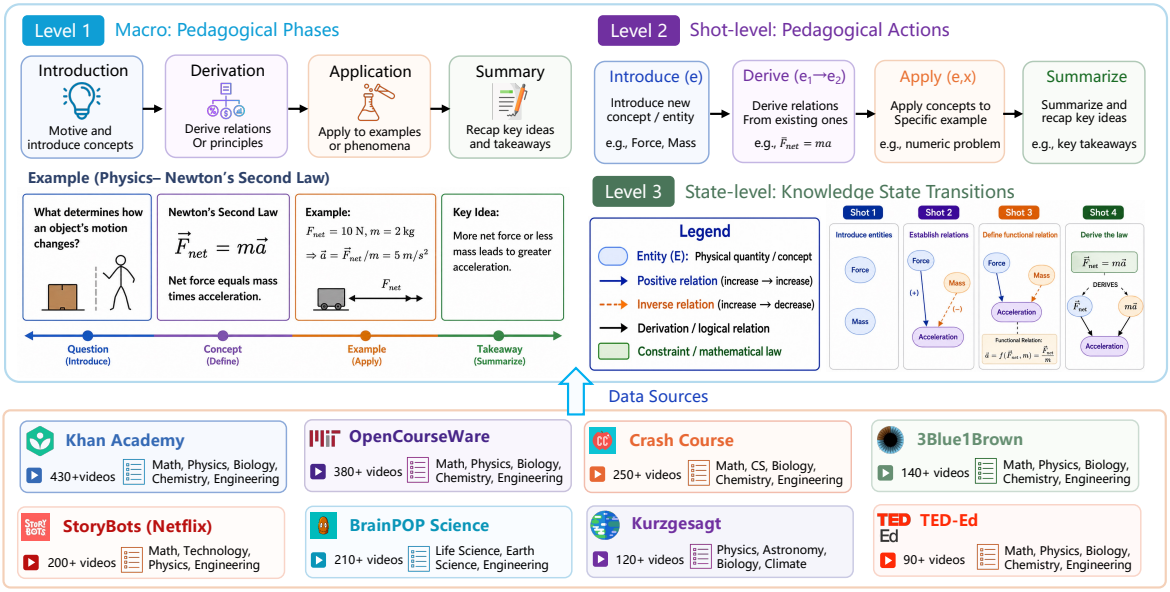

Figure 2. Overview of EduVideoBench, illustrating its multi-source composition and hierarchical annotation for modeling pedagogical structure and knowledge consistency in STEM instructional videos.

Table 1. Pedagogical action set $\mathcal{A}$ and their state effects.

Action a_t	**Effect on** S_{t+1}
INTRODUCE(e)	E += $\{e\}$; add incident edges to R
DERIVE(e_1, e_2, r)	E += $\{e_2\}$; R += $\{(e_1, e_2, r)\}$
APPLY(e, x)	E unchanged; add instantiation edge
SUMMARIZE(E')	S unchanged; trigger recap shot

with K=4 canonical pedagogical phases $\{p_k\}$ = {Introduction, Explanation, Application, Summary}, each subdivided into N_k shot-level actions. We implement π with a prompted LLM that outputs structured JSON, including per-shot action tags, expected entities, and constraint identifiers from $\mathcal{C}$.

3.3. State-Conditioned Generation and Verification

Given a shot plan and current state S_t, video shot v_t is sampled as:

$$v_t \sim P_\theta\big(v \mid \text{prompt}(a_t, S_t)\big), \tag{4}$$

where prompt(a_t, S_t) enriches the shot description with the entities in E_t and their active relations, grounding the generator in accumulated instructional context.

A *Constraint Verifier* $\mathcal{V}$ then evaluates each candidate:

$$\mathcal{V}(v_t, \mathcal{C}, S_t) = \mathbf{1}\big[\forall\, c \in \mathcal{C} :\ \text{CHECK}(v_t, c, S_t) = 1\big], \tag{5}$$

which is implemented as a VLM-based agent (GPT-4o (Hurst et al., 2024)). If $\mathcal{V} = 0$, EduStory regenerates with an augmented prompt encoding the detected violation, up to $K_{\max}$=3 retries. This closed-loop correction is the key distinction from prompt-engineering baselines.

4. EduVideoBench

4.1. Dataset Construction

EduVideoBench comprises **1,800+** multi-shot STEM instructional video clips (30–90 seconds, ≥3 shots each) from eight educational sources. The collected videos cover physics (∼**300**), mathematics (∼**420**), chemistry (∼**220**), engineering and computer science (∼**360**), biology and life science (∼**250**), earth science and astronomy (∼**110**), and other STEM-related topics (∼**40**). To ensure evaluative rigor, the dataset is constructed with the following key characteristics:

(i) Naturalistic educational content: clips are collected from Khan Academy, MIT OpenCourseWare, 3Blue1Brown, CrashCourse, Ask the StoryBots, BrainPOP Science, Kurzgesagt, and TED-Ed, providing diverse instructional styles ranging from lecture-based derivations to animated scientific explanations.

(ii) Programmatic ground truth: we additionally include ∼**100** clips rendered with Manim (Manim Community Developers, 2021), providing controlled reference videos for formula-correctness and symbolic-consistency evaluation. Together, these sources form a dataset with both naturalistic pedagogical variation and verifiable ground-truth structure.

4.2. Three-Level Annotation Taxonomy

Level 1 — Macro: Pedagogical phase sequence. Each clip is labeled with temporal boundaries of up to five phases following the instructional design literature (Gagne et al., 2005): {*phenomenon introduction, hypothesis formulation, formal derivation, example application, summary*}.

Level 2 — Shot: Semantic action tags. Each shot is labeled with a pedagogical action from $\mathcal{A}$, entities present on screen, and any formulae verified against domain rules by GPT-4o (Hurst et al., 2024).

Level 3 — Transition: Knowledge state consistency. For each consecutive pair (v_t, v_{t+1}), we annotate the state delta $\Delta S_t = S_{t+1} \setminus S_t$ and record whether entity continuity, formula symbol matching, and logical ordering are preserved, forming the ground truth for KDR.

4.3. Benchmark Tasks

EduVideoBench defines two diagnostic tasks:

Task I (Script-to-Video): given $\pi(\ell)$, generate the full shot sequence; evaluate with KDR and PAS.

Task II (Continuation): given the first k shots and state S_k, generate the remaining shots under consistency constraints; tests whether models can sustain state without full context.

5. Experiments

5.1. Setup

We use CogVideoX-2B (Yang et al., 2024) as the base generator. Experiments are run on an eight-GPU NVIDIA H100 server. We evaluate on 5,000 held-out EduVideoBench shots. GPT-4o serves as the VLM evaluator for **Knowledge Drift Rate (KDR)** and **Pedagogical Alignment Score (PAS)**; CLIP-S (Hessel et al., 2021) is included as a standard visual quality reference.

Metrics. KDR measures the fraction of consecutive shot pairs exhibiting entity-level or formula-level drift:

$$\mathrm{KDR}(V) = \frac{1}{T-1} \sum_{t=1}^{T-1} \mathbf{1}\big[\mathrm{drift}(v_t, v_{t+1}, S_t) > 0\big]. \quad (6)$$

PAS measures mean shot-level alignment with the intended instructional plan:

$$\mathrm{PAS}(V) = \frac{1}{T} \sum_{t=1}^{T} \mathrm{match}\big(\mathrm{phase}_t,\ \mathrm{plan}_t\big), \quad (7)$$

where $\mathrm{match}(\cdot)$ is scored by the VLM evaluator. KDR is lower-better; PAS and CLIP-S are higher-better.

5.2. Ablation Study

Tab. 2 reports a four-condition ablation isolating the contribution of each EduStory component. **B0** uses a single long prompt without any structural decomposition, corresponding to standard long-video generation. **B1** adds the Instruction Planner, decomposing the lesson into structured per-shot prompts. **B2** additionally conditions generation on the accumulated pedagogical state S_t. **EduStory (Full)** further incorporates the Constraint Verifier with violation-aware regeneration.

Table 2. Ablation results on EduVideoBench Task I. KDR: lower is better (↓). PAS, CLIP-S: higher is better (↑). Boldface denotes the best performance among automated systems, and underlining denotes the second-best performance, respectively.

Method	KDR ↓	PAS ↑	CLIP-S ↑
B0: Baseline (long prompt)	0.41	0.52	0.28
B1: + Instruction Planner	0.33	0.64	**0.29**
B2: + Pedagogical State Model	0.21	0.71	0.28
EduStory (Full)	<u>0.14</u>	<u>0.79</u>	0.27
Human upper bound	**0.00**	**1.00**	—

Analysis. Adding the Instruction Planner (B1) improves PAS by $+12$ points, confirming that structured shot decomposition benefits pedagogical alignment even without explicit state tracking. The largest KDR reduction occurs at the B1→B2 transition (-12 points), demonstrating that explicit state modeling, rather than prompt structure alone, is the primary driver of knowledge consistency. The Constraint Verifier in EduStory Full yields a further -7 KDR reduction via targeted violation correction, at a modest cost of -0.01 CLIP-S, reflecting the expected trade-off between faithfulness and diversity inherent in constrained generation.

6. Conclusion

We introduced EduStory, a pedagogical state machine for multishot STEM instructional video generation, and EduVideoBench, a benchmark with three-level domain-aware annotations. Our ablation study shows that explicit state modeling and constraint verification are necessary and sufficient to reduce knowledge drift and improve pedagogical alignment over strong prompt engineering baselines.

This matters because long-form generative systems fail primarily due to loss of structure over time, not lack of fluency. As sequences grow, models drift, violating prerequisite relationships and breaking conceptual coherence, which directly harms learning and trust.

EduStory addresses this by enforcing a constrained progression through instructional states, ensuring coherent concept ordering and dependency satisfaction, while constraint verification prevents error propagation. EduVideoBench complements this with evaluation of factual accuracy, conceptual structure, and instructional coherence, which are largely missing from existing benchmarks.

Overall, this work shows that reliable long-form generation requires explicit structure and domain-aware validation, not just better prompts or larger models.

References

Bansal, H., Lin, Z., Xie, T., Zong, Z., Yarom, M., Bitton, Y., Jiang, C., Sun, Y., Chang, K.-W., and Grover, A. Videophy: Evaluating physical commonsense for video generation. In *The Thirteenth International Conference on Learning Representations*, 2025. 2

Croce, F., Andriushchenko, M., Sehwag, V., Debenedetti, E., Flammarion, N., Chiang, M., Mittal, P., and Hein, M. Robustbench: a standardized adversarial robustness benchmark. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021. 2

Dong, J., Koniusz, P., Chen, J., Xie, X., and Ong, Y.-S. Adversarially robust few-shot learning via parameter co-distillation of similarity and class concept learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28535–28544, 2024. 2

Dong, J., Koniusz, P., Qu, X., and Ong, Y.-S. Stabilizing modality gap & lowering gradient norms improve zero-shot adversarial robustness of vlms. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pp. 236–247, 2025a. 2

Dong, J., Liu, J., Qu, X., and Ong, Y.-S. Confound from all sides, distill with resilience: Multi-objective adversarial paths to zero-shot robustness. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 624–634, 2025b. 2

Dong, J., Zhang, C., Qu, X., Ma, Z., Koniusz, P., and Ong, Y.-S. Robust superalignment: Weak-to-strong robustness generalization for vision-language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025c. 2

Dong, J., Moayedi, R. Z., Ong, Y.-S., and Moosavi-Dezfooli, S.-M. Allies teach better than enemies: Inverse adversaries for robust knowledge distillation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026a. 2

Dong, J., Zhang, C., Qu, X., Rong, S. Q., Thai, N. D., Pan, W., Li, X., Liu, T., Koniusz, P., and Ong, Y.-S. Tug-of-war no more: Harmonizing accuracy and robustness in vision-language models via stability-aware task vector merging. In *The Fourteenth International Conference on Learning Representations*, 2026b. 2

Gagne, R. M., Wager, W. W., Golas, K. C., Keller, J. M., and Russell, J. D. Principles of instructional design, 2005. 3

Goodfellow, I., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. In *ICLR*, 2015. 2

Henschel, R., Khachatryan, L., Poghosyan, H., Hayrapetyan, D., Tadevosyan, V., Wang, Z., Navasardyan, S., and Shi, H. Streamingt2v: Consistent, dynamic, and extendable long video generation from text. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2568–2577, 2025. 2

Hessel, J., Holtzman, A., Forbes, M., Le Bras, R., and Choi, Y. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 7514–7528, 2021. 4

Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 3, 4

Jia, Y., Wu, X., Hao, L., QinglinZhang, Q., Hu, Y., Zhao, S., and Fan, W. Uni-retrieval: A multi-style retrieval framework for stem's education. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10182–10197, 2025a. 2

Jia, Y., Xie, J., Jivaganesh, S., Hao, L., Wu, X., and Zhang, M. Seeing sound, hearing sight: Uncovering modality bias and conflict of ai models in sound localization. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b. 2

Kondratyuk, D., Yu, L., Gu, X., Lezama, J., Huang, J., Schindler, G., Hornung, R., Birodkar, V., Yan, J., Chiu, M.-C., et al. Videopoet: A large language model for zero-shot video generation. In *International Conference on Machine Learning*, pp. 25105–25124. PMLR, 2024. 2

Kou, T., Liu, X., Zhang, Z., Li, C., Wu, H., Min, X., Zhai, G., and Liu, N. Subjective-aligned dataset and metric for text-to-video quality assessment. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 7793–7802, 2024. 2

Liu, E., Pan, L., Yang, Y., Zhong, Y., Wu, Z., Wu, X., and Liu, L. Storyboard-guided alignment for fine-grained video action recognition. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 2

Liu, Y., Cun, X., Liu, X., Wang, X., Zhang, Y., Chen, H., Liu, Y., Zeng, T., Chan, R., and Shan, Y. Evalcrafter: Benchmarking and evaluating large video generation models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22139–22149, 2024. 2

Manim Community Developers. Manim: Mathematical animation engine, 2021. URL https://www.manim.community. Accessed 2025. 3

Peebles, W. and Xie, S. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023. 1

Sun, Q., Cui, Y., Zhang, X., Zhang, F., Yu, Q., Wang, Y., Rao, Y., Liu, J., Huang, T., and Wang, X. Generative multimodal models are in-context learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14398–14409, 2024a. 2

Sun, Q., Yu, Q., Cui, Y., Zhang, F., Zhang, X., Wang, Y., Gao, H., Liu, J., Huang, T., and Wang, X. Emu: Generative pretraining in multimodality. In *The Twelfth International Conference on Learning Representations*, 2024b. 2

Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.-W., Chen, D., Yu, F., Zhao, H., Yang, J., et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 2

Wu, H., Zhu, H., Zhang, Z., Zhang, E., Chen, C., Liao, L., Li, C., Wang, A., Sun, W., Yan, Q., et al. Towards open-ended visual quality comparison. In *European Conference on Computer Vision*, pp. 360–377. Springer, 2024. 2

Wu, X., Jia, Y., Zhang, Q., Qin, Y., Xiao, L., and Zhao, S. Towards affective evaluation of stem education: Leveraging mllms in project-based learning. *IEEE Transactions on Affective Computing*, 2026. 2

Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 1, 2, 4

Zhao, S., Tuan, L. A., Fu, J., Wen, J., and Luo, W. Exploring clean label backdoor attacks and defense in language models. *IEEE/ACM transactions on audio, speech, and language processing*, 32:3014–3024, 2024. 2

Zhao, S., Lin, Q., Jia, Y., Wu, X., Li, Y., and Tuan, L. A. Unifle: Uniform fusion of multiple lora experts for backdoor defense in large language models. *IEEE Transactions on Dependable and Secure Computing*, 2026. 2

Zheng, Z., Peng, X., Yang, T., Shen, C., Li, S., Liu, H., Zhou, Y., Li, T., and You, Y. Open-sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*, 2024. 1, 2

Supplementary Material

A. VLM Evaluator Prompts

KDR evaluation prompt (per adjacent shot pair). We provide the following prompt to GPT-4o alongside the frame from shot v_{t+1}:

You are evaluating knowledge drift in a STEM instructional video.
The previous shot established the following knowledge entities: `[E_t]`,
Current knowledge state: `[S_t]`.
Examine the provided video frame (current shot) and determine:
(1) Are all previously established entities correctly represented (same symbol form, directional convention, no disappearance)?
(2) Has any entity been incorrectly modified or contradicted?
(3) Has any new entity appeared without formal introduction?
Respond only in JSON:

```
{
    "entities_preserved": bool,
    "entities_incorrectly_modified": [list],
    "unexplained_new_entities": [list],
    "drift_detected": bool,
    "drift_severity": 0--3,
    "explanation": "one sentence".
}
```

PAS evaluation prompt (per shot).

You are evaluating whether a STEM instructional video shot follows its intended pedagogical plan.
Planned shot `[shot_id]/[total]`*: Phase:* `[phase]`,
Intended action: `[action]`,
Expected content: `[description]`.
Examine the provided video frame and determine:
(1) Does the visual content match the intended pedagogical action?
(2) Is the content at the appropriate level of detail for this phase?
(3) Does it logically continue from the previous phase?
Respond only in JSON:

```
{
    "action_matched": bool,
    "phase_appropriate": bool,
    "logical_continuity": bool,
    "alignment_score": 0.0--1.0,
    "explanation": "one sentence".
}
```