

From Content Generation to Learning Support: Pedagogy-Guided Generative Video Tutors for STEM Learning

Xinchen Ma¹, Shuimu Wang¹, Gaole He^{2*}, Yanbin Zhang¹,
Chunyang Wang¹, Yunshi Lan^{1*}, Weining Qian¹

¹Department of Data Science and Engineering, East China Normal University

²School of Computing, National University of Singapore

xinchen.ma@stu.ecnu.edu.cn, hegaole@nus.edu.sg, yslan@dase.ecnu.edu.cn

Abstract

Generative AI enables scalable production of educational videos, but current systems largely focus on producing visually coherent content rather than supporting learning. As a result, generated videos often lack explicit pedagogical structure, reliable quality control, and mechanisms for assessing learner understanding or addressing misconceptions. In this work, we introduce **PIVOT** (Pedagogy-guided Instructional Video Tutoring), a generative video tutoring framework for STEM learning via learning-centered instructional support.¹ Inspired by conventional teaching workflows, our framework integrates pedagogy into the full generation pipeline: it first uses instructional principles to guide storyboard generation, then produces verified multimodal videos through code-centric generation and a pedagogical verification harness, and finally connects videos with assessment and misconception-aware remediation. Experiments and expert evaluations across four STEM domains show that our framework produces educational videos with pedagogically aligned content, clear and engaging presentation, coherent instructional flow, and perceived effectiveness for learning. These findings suggest a human-centered perspective on educational content generation: generative systems should be evaluated and designed not only by what they produce, but also by how they support teaching practices, learner understanding, and corrective feedback.

1 Introduction

Preparing instructional materials and recording lectures is a time-consuming task for educators, and providing adaptive feedback to support multiple students can be an additional burden (Shi et al., 2025). The rise of generative AI has opened new opportunities for creating high-quality educational

videos that can foster student engagement and facilitate understanding (Ye et al., 2025; Pal Chowdhury et al., 2024; Graesser et al., 2004). Recent work has explored generating visually coherent educational videos from learning goals (Chen et al.; Ku et al., 2025), demonstrating the feasibility of translating instructional intents into multimodal learning materials at scale.

While these systems ensure content fidelity and visual fluency, they primarily focus on content generation rather than learning support. Consequently, several important aspects of educational video design remain largely unaddressed (Hu et al., 2025; Zerkouk et al., 2025). First, videos often lack explicit **pedagogical guidance**, such as clear learning objectives, prerequisite activation, and structured explanations, which can make it difficult for learners to follow and internalize new knowledge. Second, ensuring that visualizations, narration, and examples are **factually correct and cognitively clear** remains challenging, as errors or ambiguous layouts can impede learning or lead to misconceptions. Third, existing approaches typically do not incorporate **formal assessment of learning outcomes** or mechanisms for **corrective feedback**, limiting insight into whether students have grasped the intended learning goals. Together, these limitations highlight gaps in current generative video research regarding pedagogical support, content quality, and learning-centered evaluation.

To address these challenges, we introduce **PIVOT**, a three-stage pedagogy-guided generative video tutoring framework for STEM education that supports active learning. The first two stages build upon prior code-centric video generation frameworks (Chen et al.; Ku et al., 2025; Chen et al., 2025) and introduce two key improvements. First, pedagogical guidance is incorporated during the instructional storyboard generation. The system structures learning objectives, activates prerequisite knowledge, and provides step-by-step explanations

*Corresponding authors.

¹Code is available at our [GitHub repository](#).

along with visual representation plans and worked examples. This approach ensures that the instructional content is aligned with established principles of instructional design, scaffolding, and multimedia learning. (Cohn et al., 2026; Baek et al., 2026; Mayer, 2005) Second, video quality is enhanced through a pedagogical verification harness, ensuring that rendered animations are factually correct, visually clear, and pedagogically aligned, thereby improving the robustness and interpretability of the generated instructional videos.

In the final stage, learner assessment is incorporated to evaluate learning outcomes and address misconceptions. For each quiz question derived from the instructional storyboard, the system provides feedback on all answer options, while incorrect responses trigger a targeted remediation video that contrasts the incorrect choice with the correct reasoning. These contrastive explanations clarify underlying misconceptions and reinforce new knowledge, providing tailored guidance that addresses each learner's specific errors. By dynamically generating instructional content in response to learner performance, this stage connects AI-generated videos with post-video formative assessment. In this way, the framework shifts educational video generation from passive content delivery toward active learning support and learner-centered instructional interaction.

Extensive evaluation demonstrates that the proposed PIVOT framework can produce educational videos that effectively support STEM learning, ensuring both high-quality instructional content and meaningful learner assessment. Expert evaluation conducted with experienced high school and university instructors confirms that the videos are pedagogically sound and facilitate conceptual understanding. The contributions of this work can be summarized as follows:

- We introduce **PIVOT**, a pedagogy-guided instructional video tutoring framework that integrates storyboard-based planning, verified multimodal video generation, and assessment-driven remediation to support STEM learning.
- We validate the framework through experiments and expert assessments, demonstrating that it reliably generates videos that support student comprehension and address misconceptions.
- The framework represents a step toward AI-supported STEM learning guided by pedagogy, transforming AI-generated videos from passive

content into interactive, pedagogically informed learning experiences.

2 Related Work

2.1 Text-to-Video Generation

Recent advances in text-to-video generation models, such as Sora (Liu et al., 2024), Kling (Team et al., 2025), and other large-scale generative video systems (Google DeepMind, 2025; Zheng et al., 2025; Bar-Tal et al., 2024), have substantially expanded the research scope of AI-assisted education by enabling automatic synthesis of dynamic visual content from textual descriptions. These systems demonstrate strong capabilities in cinematic video synthesis and multimodal storytelling. However, applying general-purpose video generation models to STEM education remains challenging due to issues in temporal consistency, layout stability, and accurate rendering of symbolic content such as formulas, diagrams, and algorithms (Chen et al., 2024). Such limitations reduce the pedagogical reliability required for instructional videos.

2.2 Code-centric Educational Video Generation

Early LLM-based (Large Language Model) educational applications primarily focused on language-centric tasks, including question answering, exercise generation, and conversational tutoring (Mageira et al., 2022; Grassini, 2023; Lee et al., 2023; Zhang et al., 2025; Bi et al., 2026; Baek et al., 2026; Dickey and Bejarano, 2024). With the rapid development of multimodal generation capabilities, recent research has gradually shifted toward richer forms of educational content generation, including visual explanations, instructional animations, and educational videos.

2.3 Pedagogical Guidance and Adaptive Educational Generation

Recent research in Intelligent Tutoring Systems (ITS) and AI-assisted education has explored various forms of pedagogical guidance (Pal Chowdhury et al., 2024; Borchers and Shou, 2025; Schmucker et al., 2024; Graesser et al., 2004; Al-Shaikh et al., 2024; Calo and Maclellan, 2024; Yang et al., 2025), including adaptive learning, learner modeling, formative assessment, corrective feedback, and scaffolded instructional support (Zerkouk et al., 2025; Cohn et al., 2026; Ye et al., 2025). Prior work has shown that peda-

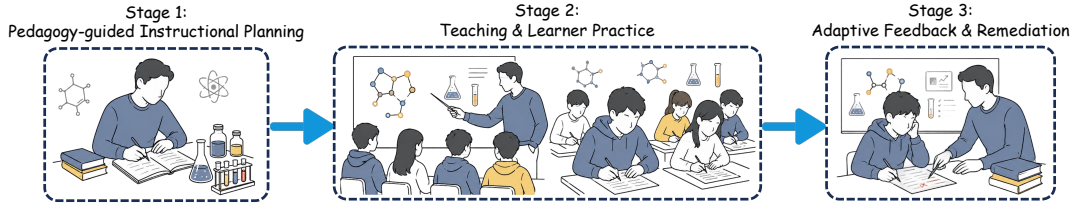


Figure 1: Real-world instructional workflow.

gogically informed tutoring systems can improve learner engagement and conceptual understanding through adaptive instruction based on learner performance.

Despite recent advances, existing educational video generation systems mainly focus on multimedia synthesis with limited support for pedagogical planning, assessment, and misconception-aware feedback. In contrast, our framework integrates pedagogical guidance, formative assessment, and remediation into a unified pipeline, transforming educational video generation into learner-centered instructional interaction.

3 PIVOT: Pedagogy-guided Instructional Video Tutoring

3.1 System Architecture

Designing effective educational videos demands careful alignment with pedagogical principles, including lesson planning, instructional delivery, learner practice, and adaptive feedback. In conventional STEM education, instructors typically follow a structured workflow (as shown in Figure 1): they first prepare lesson plans that organize instructional goals, prerequisite knowledge, worked examples, and assessments. They then present lessons to students through demonstrations or lectures. Students engage in exercises to consolidate understanding, and instructors provide targeted feedback to address misconceptions and reinforce learning.

Inspired by this instructional process, our framework is structured into three stages (shown in Figure 2). The first stage, **Pedagogy-guided Instructional Planning**, emulates lesson preparation by generating structured storyboards that organize learning objectives, prerequisite activation, stepwise explanations, visual representation plans, worked examples and formative assessment probes. This stage ensures that educational content is pedagogically coherent and scaffolded, providing a solid foundation for downstream video generation. The second stage, **Video Generation with**

Pedagogical Verification Harness, transforms the storyboard into shot-level animations. Building upon prior code-centric video generation frameworks (Chen et al.; Ku et al., 2025), this stage integrates a multimodal verification harness to improve visual clarity, factual correctness, and cognitive clarity, enhancing the quality and interpretability of generated videos. The third stage, **Adaptive Assessment and Remediation**, parallels post-class student practice and feedback, generating quizzes and targeted explanatory videos to identify misunderstandings and reinforcing conceptual understanding. By integrating pedagogy across the pipeline, our framework turns educational video generation from passive content production into learner-centered support.

3.2 Pedagogy-guided Instructional Planning

To generate high-quality educational videos, we first integrate pedagogical principles into the creation of instructional materials. Simply generating video frames without structured content risks producing visually coherent but educationally ineffective material. By constructing a structured instructional storyboard, we ensure that content is not only organized but also **pedagogically meaningful**, providing learners with a clear path through complex concepts. This intermediate representation allows the system to embed instructional scaffolding, manage cognitive load, and prepare content that is interpretable by downstream video generation modules, ultimately guiding the learner toward progressively deeper understanding.

Given a learner-specified topic, the storyboard includes components grounded in educational theory, each serving a clear pedagogical purpose:

- **Learning Objectives:** Define the target concepts and expected outcomes, clarifying the purpose of each lesson and focusing both the content generation and the learner’s attention (Bloom, 1956).
- **Prerequisite Activation:** Recall relevant prior knowledge to reduce cognitive load and support

PIVOT (Pedagogy-guided Instructional Video Tutoring)

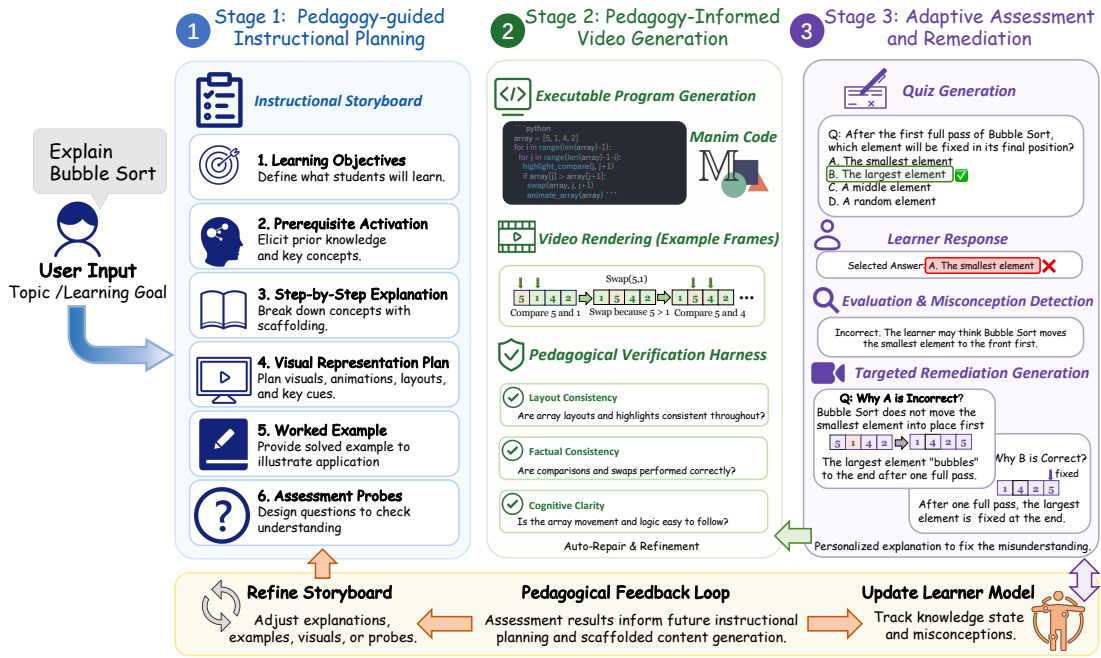


Figure 2: Overview of the PIVOT Framework.

integration of new concepts, preparing learners to grasp advanced material (Sweller, 1988).

- **Concept Explanations**: Present ideas progressively through stepwise reasoning, scaffolding complex concepts and enabling incremental knowledge acquisition (Wood et al., 1976; Cohn et al., 2026).
- **Visual Representation Plans**: Specify diagrams, formulas, highlights, and animations according to multimedia learning principles, enhancing comprehension, retention, and cognitive clarity (Mayer, 2005).
- **Worked Examples**: Demonstrate procedural steps and intermediate reasoning, helping learners internalize problem-solving strategies and reinforcing conceptual understanding (Renkl, 2014).
- **Diagnostic Assessment Probes**: Include formative questions to detect misconceptions early, guiding adaptive feedback and reinforcing learning (Black and Wiliam, 1998)

This careful preparation helps bridge educational theory and AI-generated content, laying a solid foundation for producing videos that are coherent, scaffolded, and pedagogically effective. The generated storyboard further serves as the semantic backbone for executable Manim² code generation,

narration synthesis, assessment construction, and multimodal verification³.

3.3 Video Generation with Pedagogical Verification Harness

Building on the storyboard from Stage 1, the system transforms structured instructional content into shot-level executable videos. Each shot contains pedagogically grounded components, including local learning goals, narration, visual descriptions, visual elements (e.g., formulas, diagrams, highlights), and duration constraints. Dividing the content into shot-level units allows precise control over each instructional segment, facilitates iterative verification and repair, and ensures that pedagogical objectives are faithfully represented. By using this intermediate representation, the system ensures that videos preserve instructional structure and scaffolded learning.

Pedagogical Verification Harness. To ensure that generated videos faithfully implement the instructional storyboard, we employ a multimodal verification harness. The harness is designed to detect potential issues that could compromise educational fidelity or visual clarity, covering layout, cognitive clarity and factual consistency. As summarized in Table 1, each component guides iterative repair,

²<https://github.com/ManimCommunity/manim>

³Examples of the storyboard are provided in Appendix A.

Component	Function
Layout Consistency	Detects overlap, spacing inconsistency, readability issues, and excessive information density.
Factual Consistency	Verifies formulas, symbolic transitions, algorithmic procedures, and instructional explanations.
Cognitive Clarity	Verifies logical consistency, instructional progression, and coherence of explanatory content.

Table 1: Pedagogical verification components.

ensuring that pedagogical intent is preserved while improving interpretability and robustness.

Verification-guided Rendering and Repair. To further maintain instructional and visual quality, the system applies constrained layout prompting (see Appendix A.4) and automated code repair. These mechanisms reduce common generation failures such as overlapping elements, unreadable text, and inconsistent scene transitions, preserving the instructional structure of each shot. When verification failures are detected, the generated Manim code is iteratively repaired using low-temperature prompting conditioned on execution traces and rendering feedback. GPT-5.2 serves as the primary generation and repair model, while persistent failures are transferred to Claude Sonnet 4.6 for additional debugging and executable repair.

After verification, a Text-to-Speech (TTS) module generates synchronized narration aligned with the rendered animations. For remediation videos, generation is further conditioned on learner responses (*i.e.*, quiz questions, candidate options, incorrect answers, and corresponding correct reasoning), enabling adaptive, misconception-aware explanations. By grounding video generation in the pedagogically structured storyboard and enforcing multi-level verification, this stage ensures that the generated videos remain aligned with the intended instructional objectives.

3.4 Adaptive Assessment and Remediation

To provide learner-centered feedback, address misconceptions, and reinforce conceptual understanding, the framework incorporates an adaptive assessment and remediation module. This module generates quizzes and targeted follow-up explanations based on the instructional storyboard from Stage 1, forming a closed feedback loop that aligns instructional content, learner responses, and peda-

gogical objectives while actively addressing misunderstandings.

Pedagogically Grounded Quiz Generation. Quiz generation leverages the structured storyboard, summarizing shot-level content including learning objectives, explanations, key concepts, worked examples, and visual demonstrations. This design ensures that generated questions remain semantically aligned with the instructional video and intended learning goals. To encourage conceptual understanding rather than memorization, our framework does not reuse examples, equations, or numerical values from the generated educational video. By grounding assessments in the pedagogical structure of the lesson, the framework ensures alignment with instructional goals and supports learners in building robust conceptual knowledge.

Misconception-aware Remediation Generation. Learner responses provide adaptive pedagogical signals to guide remediation. When a learner answers incorrectly, the system generates a targeted explanatory video contrasting the incorrect reasoning with the correct solution. The remediation generation is conditioned on the quiz question, candidate options, learner response, and the correct answer, ensuring that feedback addresses specific misconceptions. By integrating learner signals with the instructional storyboard, the framework produces remediation that preserves semantic consistency, instructional fidelity, and pedagogical coherence, supporting learners in correcting misunderstandings and reinforcing knowledge.

Through adaptive pedagogy-guided generation, the system connects instructional content, video synthesis, and learner interaction to support effective learning.

4 Experiments

For evaluation, we select two representative baseline systems: **TheoremExplainAgent** (Ku et al., 2025), and **Code2Video** (Chen et al.). These systems represent state-of-the-art approaches in automatic educational video generation, primarily focusing on content synthesis without explicit pedagogical guidance or learner-adaptive feedback. To ensure fair comparison, all methods are evaluated on the same set of 40 STEM instructional topics spanning mathematics, physics, chemistry, and computer science, with 10 knowledge points selected for each subject (see Appendix B.1).

4.1 Automatic Evaluation

To evaluate generated instructional videos, we conduct automatic assessment using large Vision-Language Models (VLMs). Video frames sampled from the central timeline are independently evaluated by Gemini-2.5-Flash (Imran and Almusharraf, 2024) under deterministic decoding ($T = 0$). Each frame is scored on two dimensions: **Layout**, measuring spatial clarity and visual organization, and **Richness**, evaluating visual diversity and instructional content coverage. These dimensions primarily evaluate visual presentation quality, while higher-level pedagogical effectiveness is assessed separately through human evaluation. In Table 2, final video scores are obtained by averaging frame-level results across all sampled frames.

Method	Layout	Richness	Overall
TheoremExplainAgent	9.00	4.23	6.62
Code2video	8.18	4.22	6.20
PIVOT	9.10	6.06	7.58

Table 2: Video quality evaluation results (1–10).

In Table 3, we further evaluate instructional narration quality using GPT-4o (OpenAI et al., 2024). Narration fields are extracted from generated scripts and independently scored along two pedagogically grounded dimensions: **Accuracy and Depth**, assessing conceptual correctness and explanatory quality, and **Logical Flow**, evaluating instructional coherence and progression. The final transcript score is computed as their arithmetic mean⁴. Overall, our method achieves the best performance on both video and narration quality evaluation, demonstrating improved visual organization, instructional richness, conceptual accuracy, and logical coherence.

Method	Accuracy	Logical	Overall
TheoremExplainAgent	0.85	0.82	0.84
Code2video	0.78	0.72	0.75
PIVOT	0.87	0.91	0.89

Table 3: Narration quality evaluation results (0–1).

4.2 Ablation Analysis

We further conduct an ablation study to analyze the contribution of three core verification components in the proposed framework: **Layout Consistency**,

⁴Detailed evaluation prompts are provided in Appendix B.2.

Factual Consistency, and **Cognitive Clarity**. As summarized in Table 4, these modules mainly affect visual organization, conceptual correctness, and instructional coherence, respectively. We evaluate their impact on both visual quality and narration quality.

Ablation	Layout	Accuracy	Logical
w/o Layout Verification	7.87	–	–
w/o Factual Verification	–	0.85	–
w/o Cognitive Verification	–	–	0.89
Full Framework (PIVOT)	9.10	0.87	0.91

Table 4: Ablation study of core verification components in the proposed framework.

First, removing the **Layout Consistency** module significantly degrades the Layout score from 9.10 to 7.87. This result demonstrates that layout verification plays an important role in maintaining spatial clarity and visual cleanliness. Without explicit layout control, generated scenes are more likely to contain overlapping elements and inconsistent organization, reducing instructional readability. This module mainly affects visual quality rather than narration generation.

Second, removing **Factual Consistency** decreases the Accuracy & Depth score from 0.87 to 0.85, indicating that factual verification improves conceptual correctness and explanatory reliability. Without this module, generated narrations are more likely to contain imprecise explanations or incomplete reasoning. For example, in chemical equation balancing, the model may incorrectly describe the law of conservation of mass as conservation of atom counts, potentially introducing learner misconceptions.

Third, removing the **Cognitive Clarity** module decreases the Logical Flow score from 0.91 to 0.89. The resulting narrations are more likely to exhibit abrupt topic transitions, weaker instructional progression, and less coherent explanation structures. This finding indicates that explicit logical verification improves instructional coherence and helps the generated videos better follow pedagogically structured teaching sequences rather than mechanically listing isolated reasoning steps.

4.3 Human Evaluation

We conducted a questionnaire-based expert evaluation with 32 STEM instructors to assess the educational quality and perceived learning effects of the generated videos. The expert panel included

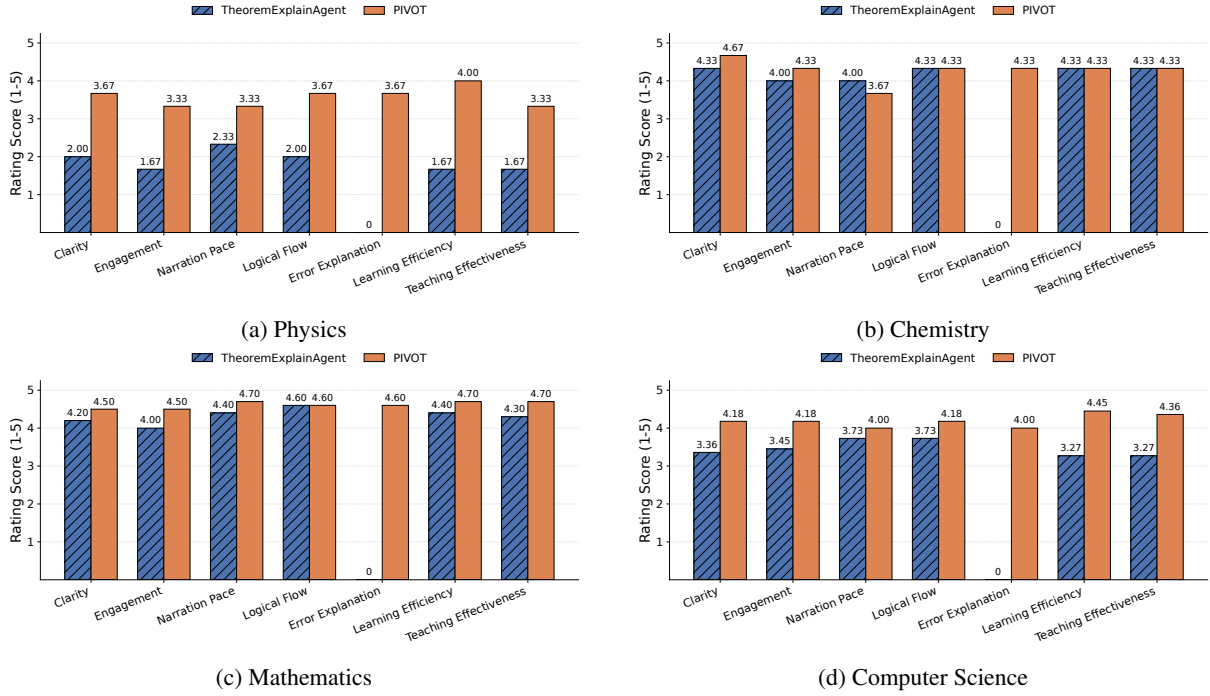


Figure 3: Expert evaluation results across four STEM domains.

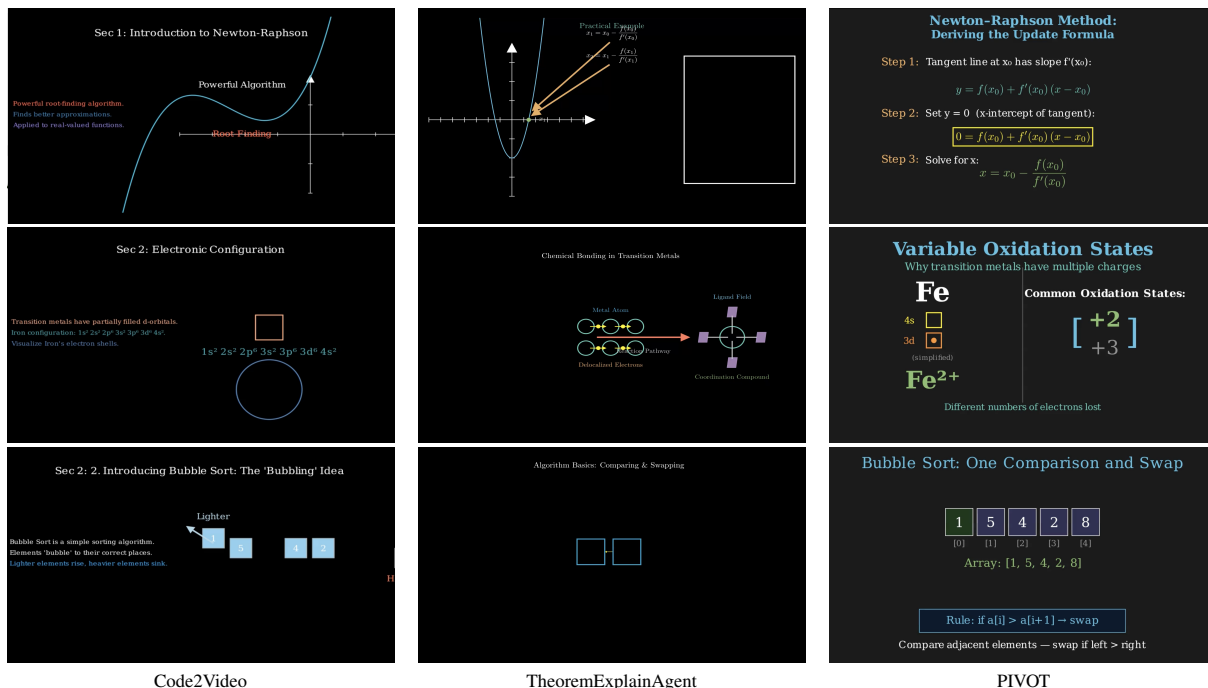
high-school teachers in mathematics, physics, and chemistry, as well as university instructors in computer science, covering domains where abstract concepts often require careful explanation and visual representation (see Appendix B.3). Because these instructors regularly design lessons, present complex concepts, and identify student misconceptions, their judgments provide an appropriate basis for evaluating whether the generated videos are pedagogically sound and potentially useful for learning. The panel consisted of 3 physics teachers, 12 mathematics teachers, 5 chemistry teachers, and 12 computer science instructors.

We compare two instructional paradigms. Based on the previous automatic evaluation results, TheoremExplainAgent (Ku et al., 2025) is selected as the strongest baseline for expert evaluation. In the baseline setting, participants only watch a single instructional video generated by TheoremExplainAgent without any subsequent assessment or corrective feedback stage. In contrast, our framework follows a learner-centered instructional workflow. Participants first watched a pedagogically structured instructional video and reviewed a related assessment question. We then presented a plausible incorrect answer to simulate a learner misconception, followed by a remediation video generated to address that error. Participants (*i.e.*, experts) evaluated whether the remediation effectively clari-

fied the misconception and supported conceptual understanding.

To reduce presentation bias, the display order of the baseline videos and our generated videos is randomly shuffled during the evaluation process. In total, 32 teachers and educational experts participated in the study. All questionnaire items are evaluated using a five-point Likert scale ranging from “**Strongly Disagree**” to “**Strongly Agree**.” The evaluation measures multiple instructional dimensions, including clarity, engagement, narration pace, logical flow, learning efficiency, and overall teaching effectiveness. Detailed questionnaire dimensions are provided in Appendix B.3.

As shown in Figure 3, our framework consistently outperforms the baseline across nearly all evaluation dimensions. The improvements are particularly pronounced in Physics and Computer Science: Physics shows an average gain of +1.67 points over the baseline across comparable dimensions, while Computer Science improves by +0.76 points. In these domains, learners’ misconceptions are often tied to intermediate procedural states that are difficult to resolve through standalone instructional videos. By incorporating assessment and targeted remediation, the framework can explicitly diagnose incorrect reasoning patterns and provide more effective corrective explanations. In contrast, improvements in Mathematics and Chemistry are



Code2Video

TheoremExplainAgent

PIVOT

Figure 4: Qualitative comparison of educational videos generated by different methods.

comparatively smaller because many evaluated topics are more symbolic and rule-driven, where procedural steps can already be conveyed relatively well through a single instructional video.

Notably, the largest improvement is observed in **Error Explanation**, suggesting that misconception-aware remediation improves both conceptual understanding and potential learning effectiveness by correcting learner misunderstandings. Moreover, our framework achieves higher **Clarity** scores, indicating that pedagogy-guided instructional planning improves the readability, coherence, and interpretability of generated educational videos. Furthermore, 96.9% of participating experts reported that the workflow of “instructional video → assessment → targeted remediation” provides a more effective learning experience than instructional videos alone.

These results demonstrate that integrating pedagogical guidance, formative assessment, and adaptive remediation can substantially improve learner engagement and instructional effectiveness, further supporting our goal of transforming AI-generated educational videos from passive multimedia content into active learning support systems.

4.4 Case Study

To provide comparison across different instructional generation paradigms, Figure 4 presents representative educational videos generated by **Code2Video**, **TheoremExplainAgent**, and our

framework on mathematics, chemistry, and computer science topics, including the Newton-Raphson Method, Transition Metals, and Bubble Sort. Compared with prior methods, our framework produces more visually organized and pedagogically structured instructional videos, with clearer symbolic visualization, more coherent instructional layouts, and richer multimodal presentation ⁵.

5 Conclusion

We presented PIVOT, a pedagogy-guided instructional video tutoring framework that reframes STEM educational video generation as a learner-centered workflow. By integrating instructional planning, pedagogically verified video generation, and assessment-based remediation, the framework connects content generation with pedagogical structure and learner feedback. Our evaluations show that the framework can generate high-quality educational videos that are pedagogically aligned and perceived as effective, while the integrated assessment and remediation process provides meaningful support for learner understanding and misconception correction. More broadly, our findings point to a human-centered direction for educational content generation: future systems should be designed not only to produce high-quality artifacts, but also to

⁵Additional assessment questions and bilingual case examples are provided in Appendix C.

support teaching practices, learner understanding, and corrective feedback.

Limitations

Although the proposed framework demonstrates promising capabilities in pedagogy-guided educational video generation and adaptive remediation, several limitations remain.

First, the current system primarily performs interaction-level adaptation conditioned on immediate learner responses rather than maintaining a fully personalized long-term learner model. While the framework can generate misconception-aware remediation videos for incorrect answers, it does not yet explicitly model long-term knowledge tracing, learner memory, curriculum progression, or individualized difficulty estimation across extended learning sessions.

Second, the multi-stage storyboard generation, executable rendering, verification, and remediation pipeline introduces substantially higher computational cost than direct multimedia synthesis approaches. Improving generation efficiency while preserving pedagogical quality remains an important direction for future work.

Ethics and Impact Statement

All participants involved in the human evaluation study participated voluntarily with informed consent. The evaluation was conducted anonymously, and no personally identifiable information was collected or disclosed. Participant responses were used solely for academic research purposes and were not shared with third parties.

The study was conducted entirely online through a questionnaire-based platform, and participants could withdraw at any time without penalty. The evaluation process did not involve sensitive personal data, psychological intervention, or any procedure that could reasonably introduce ethical, physical, or social risks to participants.

More broadly, our framework is designed to support educational content creation and learner assistance in STEM education. While the system aims to improve instructional accessibility and learning support, generated educational content may still contain factual inaccuracies or pedagogical limitations. Therefore, human oversight from educators remains important when deploying such systems in real educational settings.

Acknowledgments

The authors would like to thank the anonymous reviewers for their insightful comments. This work is supported by the National Natural Science Foundation of China (Project No. 62507015, 62537001, 62137001) and the Chenguang Program of Shanghai Education Development Foundation and Shanghai Municipal Education Commission (Project No. 24CGA26).

References

- Rana AlShaikh, Norah Al-Malki, and Maida Almasre. 2024. The implementation of the cognitive theory of multimedia learning in the design and evaluation of an ai educational video assistant utilizing large language models. *Heliyon*, 10(3).
- Injun Baek, Yearim Kim, and Nojun Kwak. 2026. Pedaco-gen: Scaffolding pedagogical agency in human-ai collaborative video authoring. In *Proceedings of the Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems*, pages 1–11.
- Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, et al. 2024. Lumiere: A space-time diffusion model for video generation. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11.
- Shuzhen Bi, Mingzi Zhang, Zhuoxuan Li, Xiaolong Wang, Keqian Li, and Aimin Zhou. 2026. Eduillustrate: Towards scalable automated generation of multimodal educational content. *arXiv preprint arXiv:2604.05005*.
- Paul Black and Dylan Wiliam. 1998. Assessment and classroom learning. *Assessment in Education: principles, policy & practice*, 5(1):7–74.
- Benjamin Samuel Bloom. 1956. *Taxonomy of educational objectives: Affective domain*, volume 2. Longmans, Green.
- Conrad Borchers and Tianze Shou. 2025. Can large language models match tutoring system adaptivity? a benchmarking study. In *International Conference on Artificial Intelligence in Education*, pages 407–420. Springer.
- Tommaso Calo and Christopher Maclellan. 2024. Towards educator-driven tutor authoring: generative ai approaches for creating intelligent tutor interfaces. In *Proceedings of the Eleventh ACM Conference on Learning@ Scale*, pages 305–309.
- Hao Chen, Tianyu Shi, Pengran Huang, Zeyuan Li, Jiahui Pan, Qianglong Chen, and Lewei He. 2025. Visualedu: A benchmark for assessing coding and

- visual comprehension through educational problem-solving video generation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 16363–16394.
- Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. 2024. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7310–7320.
- Yanzhe Chen, Kevin Qinghong Lin, and Mike Zheng Shou. Code2video: A code-centric paradigm for educational video generation. In *NeurIPS 2025 Fourth Workshop on Deep Learning for Code*.
- Clayton Cohn, Surya Rayala, Namrata Srivastava, Joyce Horn Fonteles, Shruti Jain, Xinying Luo, Divya Mereddy, Naveeduddin Mohammed, and Gautam Biswas. 2026. A theory of adaptive scaffolding for llm-based pedagogical agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 1757–1765.
- Ethan Dickey and Andres Bejarano. 2024. Gaide: A framework for using generative ai to assist in course content development. In *2024 IEEE Frontiers in Education Conference (FIE)*, pages 1–9. IEEE.
- Google DeepMind. 2025. Veo 3: Generative video model. <https://deepmind.google/technologies/veo/>.
- Arthur C Graesser, Shulan Lu, George Tanner Jackson, Heather Hite Mitchell, Mathew Ventura, Andrew Olney, and Max M Louwerse. 2004. Autotutor: A tutor with dialogue in natural language. *Behavior Research Methods, Instruments, & Computers*, 36(2):180–192.
- Simone Grassini. 2023. Shaping the future of education: Exploring the potential and consequences of ai and chatgpt in educational settings. *Education sciences*, 13(7):692.
- Xiangen Hu, Sheng Xu, Richard Tong, and Art Graesser. 2025. Generative ai in education: From foundational insights to the socratic playground for learning. *arXiv preprint arXiv:2501.06682*.
- Muhammad Imran and Norah Almusharraf. 2024. Google gemini as a next generation ai educational tool: a review of emerging educational technology. *Smart Learning Environments*, 11(1):22.
- Max Ku, Cheuk Hei Chong, Jonathan Leung, Krish Shah, Alvin Yu, and Wenhui Chen. 2025. Theoremexplainagent: Towards video-based multimodal explanations for llm theorem understanding. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6663–6684.
- Dong Won Lee, Chaitanya Ahuja, Paul Pu Liang, Sanika Natu, and Louis-Philippe Morency. 2023. Lecture presentations multimodal dataset: Towards understanding multimodality in educational videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20087–20098.
- Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, et al. 2024. Sora: A review on background, technology, limitations, and opportunities of large vision models. *arXiv preprint arXiv:2402.17177*.
- Kleopatra Mageira, Dimitra Pittou, Andreas Papasalouros, Konstantinos Kotis, Paraskevi Zangogianni, and Athanasios Daradoumis. 2022. Educational ai chatbots for content and language integrated learning. *Applied Sciences*, 12(7):3239.
- Richard E Mayer. 2005. Cognitive theory of multimedia learning. *The Cambridge handbook of multimedia learning*, 41(1):31–48.
- OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. [Gpt-4 technical report](#). *Preprint*, arXiv:2303.08774.
- Sankalan Pal Chowdhury, Vilém Zouhar, and Mrinmaya Sachan. 2024. Autotutor meets large language models: A language model tutor with rich pedagogy and guardrails. In *Proceedings of the Eleventh ACM Conference on Learning@ Scale*, pages 5–15.
- Alexander Renkl. 2014. Toward an instructionally oriented theory of example-based learning. *Cognitive science*, 38(1):1–37.
- Robin Schmucker, Meng Xia, Amos Azaria, and Tom Mitchell. 2024. Ruffle&riley: Insights from designing and evaluating a large language model-based conversational tutoring system. In *International Conference on Artificial Intelligence in Education*, pages 75–90. Springer.
- Yao Shi, Rongkeng Liang, and Yong Xu. 2025. Educationq: Evaluating llms’ teaching capabilities through multi-agent dialogue framework. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32799–32828.
- John Sweller. 1988. Cognitive load during problem solving: Effects on learning. *Cognitive science*, 12(2):257–285.
- Kling Team, Jialu Chen, Yuanzheng Ci, Xiangyu Du, Zipeng Feng, Kun Gai, Sainan Guo, Feng Han, Jingbin He, Kang He, et al. 2025. Kling-omni technical report. *arXiv preprint arXiv:2512.16776*.

- David Wood, Jerome S Bruner, and Gail Ross. 1976. The role of tutoring in problem solving. *Journal of child psychology and psychiatry*, 17(2):89–100.
- Qihao Yang, Yu Yang, Sixu An, Tianyong Hao, and Guandong Xu. 2025. Llm-based collaborative agents with pedagogy-guided interaction modeling for timely instructive feedback generation in task-oriented group discussions. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 9972–9980.
- Jingheng Ye, Shen Wang, Deqing Zou, Yibo Yan, Kun Wang, Hai-Tao Zheng, Ruitong Liu, Zenglin Xu, Irwin King, Philip S Yu, et al. 2025. Position: Llms can be good tutors in english education. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 17527–17546.
- Meriem Zerkouk, Miloud Mihoubi, and Belkacem Chikhaoui. 2025. [A comprehensive review of ai-based intelligent tutoring systems: Applications and challenges](#). *Preprint*, arXiv:2507.18882.
- Zheyuan Zhang, Daniel Zhang-Li, Jifan Yu, Linlu Gong, Jinchang Zhou, Zhanxin Hao, Jianxiao Jiang, Jie Cao, Huiqin Liu, Zhiyuan Liu, et al. 2025. Simulating classroom education with llm-empowered agents. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10364–10379.
- Zangwei Zheng, Xiangyu Peng, Yuxuan Lou, Chenhui Shen, Tom Young, Xinying Guo, Binluo Wang, Hang Xu, Hongxin Liu, Mingyan Jiang, et al. 2025. Open-sora 2.0: Training a commercial-level video generation model in \$200 k. *arXiv preprint arXiv:2503.09642*.

A Appendix

A.1 Instructional Planning Prompt

The prompt in Figure 5 is used to guide the LLM to generate a pedagogically structured instructional storyboard before executable Manim code generation.

A.2 Generated Storyboard Example

For the learner-specified topic “Bubble Sort”, the instructional planning module generates the corresponding pedagogical storyboard, as illustrated in Figure 6.

A.3 Shot-level Storyboard Example

As shown in Figure 7, the high-level instructional plan is further decomposed into shot-level storyboard units. Each shot contains a local learning goal, narration, visual specification, expected duration, and visual elements that guide downstream Manim generation.

A.4 Constrained Layout Prompt

Figure 8 presents the constrained layout prompt used to improve spatial organization, visual readability, and instructional consistency during executable educational video generation.

B Evaluation

B.1 Instructional Topics Used for Video Generation and Evaluation

The knowledge points in Figure 9 were used for instructional video generation and evaluation in our experiments.

B.2 Automatic Evaluation

Figure 11 and Figure 10 present the prompts used for automatic evaluation of generated instructional videos and narrations.

B.3 Expert evaluation

The expert evaluation was conducted fully online using a questionnaire-based assessment platform. We randomly recruited 32 STEM instructors from public educational discussion forums and teacher communities, including high-school teachers in mathematics, physics, and chemistry, as well as university instructors in computer science. This recruitment strategy helps improve diversity in teaching backgrounds and reduces potential sampling

bias associated with a single institution or local region, providing a reasonably representative sample for the user study.

All participants independently evaluated the generated instructional videos without interaction with the authors or other participants. The questionnaire included multiple pedagogically grounded evaluation dimensions, including instructional clarity, engagement, narration pace, logical flow, learning efficiency, and overall teaching effectiveness. Table 5 summarizes the abbreviations of questionnaire items used in the evaluation. To improve transparency and reproducibility, screenshots of the online evaluation interface and questionnaire examples are provided in Figure 12.

C Qualitative case studies

C.1 Examples of Assessment Questions

As shown in Figure 13, the generated post-instruction assessment questions for transition metals and gradient descent are designed to identify learners’ potential misconceptions and reasoning errors.

C.2 Bilingual Case Analysis

The generated workflow includes pedagogically structured instructional videos together with assessment-driven remediation videos. The framework supports controllable multimodal instructional generation, including symbolic visualization, high-fidelity diagrams, and dynamic instructional animations. Moreover, the generated remediation videos remain aligned with the corresponding assessment questions and learner misunderstanding patterns, enabling misconception-aware corrective explanations rather than generic answer summaries.

The top row of Figure 14 presents a Chinese chemistry example on transition metals, while the bottom row presents an English computer science example on gradient descent optimization. In particular, Figure 14(c) demonstrates that our framework can further integrate externally generated visual assets, where a Nano Banana-generated gradient descent visualization is incorporated into the executable Manim rendering pipeline to improve visual interpretability and instructional expressiveness. Figure 15 shows additional examples of Chinese instructional videos generated by PIVOT framework.

Metric	Questionnaire Item
Clarity	The explanation content is clear and easy to understand.
Engagement	The presentation style is engaging and keeps my attention.
Narration Pace	The narration pace is appropriate.
Logical Flow	The instructional structure and logical flow are clear.
Error Explanation	The error-explanation video clearly explains each answer option.
Learning Efficiency	This teaching method improves my learning efficiency.
Teaching Effectiveness	This teaching method improves overall teaching effectiveness.

Table 5: Mapping between evaluation metrics and questionnaire items used in the human evaluation study.

Instructional Planning Prompt

Given a learner-specified topic, generate a structured instructional storyboard for an educational animation video.

The output should not be executable code. Instead, it should describe the instructional design of the video. The storyboard must organize the topic into pedagogically meaningful components, including learning objectives, prerequisite activation, concept explanation, visual representation plan, worked examples, and diagnostic assessment probes.

For each topic, produce a JSON object with the following structure:

```
{
  "title": "...",
  "topic": "...",
  "difficulty_level": "...",
  "learning_objective": "...",
  "prerequisite_activation": "...",
  "concept_explanations": [
    {
      "concept": "...",
      "explanation_goal": "...",
      "scaffolding_strategy": "..."
    }
  ],
  "visual_plan": "...",
  "worked_example": "...",
  "assessment_probe": "...",
  "shots": [
    {
      "shot": 1,
      "goal": "...",
      "narration": "...",
      "visual_description": "...",
      "visual_elements": ["...", "..."],
      "duration": "..."
    }
  ]
}
```

Design requirements:

1. Each shot should focus on one local learning goal.
2. The storyboard should progress from prerequisite knowledge to core concepts, then to examples and assessment.
3. Visual plans should specify formulas, diagrams, highlights, arrows, labels, and animations where appropriate.
4. Narration should explain why the concept works, not only describe what appears on screen.
5. The assessment probe should target a likely misconception or reasoning error.
6. Output only valid JSON.

Figure 5: Prompt template used for instructional planning.

Generated Storyboard Example

```
{
  "title": "Bubble Sort: Moving the Largest Element Step by Step",
  "topic": "Bubble Sort",
  "difficulty_level": "Basic",
  "learning_objective": "Learners will understand how bubble sort repeatedly compares adjacent elements, swaps inverted pairs, and places the largest remaining element at the end of the unsorted region after each pass.",
  "prerequisite_activation": "Recall the concepts of arrays, element indices, comparison operators, loops, and swapping two values.",
  "concept_explanations": [
    {
      "concept": "Adjacent comparison",
      "explanation_goal": "Show that bubble sort only compares neighboring elements rather than searching the whole array at once.",
      "scaffolding_strategy": "Begin with one pair of elements, display the comparison condition, and animate either swap or no-swap."
    },
    {
      "concept": "One full pass",
      "explanation_goal": "Explain why the largest value must move to the rightmost unsorted position after one pass.",
      "scaffolding_strategy": "Track the same large element across successive adjacent comparisons until it reaches the end."
    },
    {
      "concept": "Shrinking unsorted region",
      "explanation_goal": "Show how each completed pass reduces the range that still needs to be sorted.",
      "scaffolding_strategy": "Use a locked or greyed-out region to mark elements that no longer participate in later passes."
    }
  ],
  "visual_plan": "Represent the array as a row of numbered rectangles. Use yellow outlines to highlight the current adjacent pair, red arrows for swaps, green locks for elements that are fixed, and a grey overlay for the sorted suffix. Display loop variables i and j near the array to connect the animation with the algorithm.",
  "worked_example": "Use the array [5, 1, 4, 2, 8]. Demonstrate the first pass: compare 5 and 1, swap; compare 5 and 4, swap; compare 5 and 2, swap; compare 5 and 8, no swap. Conclude that 8 is fixed at the end.",
  "assessment_probe": "After the first pass of bubble sort on [3, 1, 4, 2], which element is guaranteed to be in its final position? Explain why."
}
```

Figure 6: Pedagogical Storyboard of Bubble Sort.

Shot-level Storyboard Example

```
[
  {
    "shot": 1,
    "goal": "Introduce the intuition of bubble sort.",
    "narration": "Bubble sort works by repeatedly comparing two neighboring values. If they are in the wrong order, we swap them. After enough adjacent swaps, larger values gradually move toward the right side of the array.",
    "visual_description": "Show the title 'Bubble Sort' and an array [5, 1, 4, 2, 8]. A small moving marker travels from left to right above adjacent pairs, indicating the direction of comparison.",
    "visual_elements": [
      "title text",
      "array rectangles",
      "comparison arrow",
      "moving marker",
      "rightward direction cue"
    ],
    "duration": "12 seconds"
  },
  ...
  {
    "shot": 4,
    "goal": "Connect the animation to nested loops.",
    "narration": "The outer loop counts how many passes have been completed. The inner loop performs adjacent comparisons only inside the remaining unsorted region.",
    "visual_description": "Place pseudocode on the left and the animated array on the right. Highlight loop variable i in blue and j in orange. Grey out the sorted suffix.",
    "visual_elements": [
      "pseudocode block",
      "i highlight",
      "j pointer",
      "grey sorted suffix",
      "range bracket"
    ],
    "duration": "12 seconds"
  },
  ...
]
```

Figure 7: Pedagogical Shot-Level Storyboard of Bubble Sort.

Constrained Layout Prompt

You are a Manim animation programming expert and visual designer. Your task is to generate beautiful and professional Manim animation code based on an educational video storyboard.

[Most Important] Visual Layout Quality Requirements (must be strictly followed):

1. Screen Layout
 - Place the title near the top of the screen, around $UP * 3$.
 - Keep the main content centered.
 - Place explanatory text near the bottom of the screen, around $DOWN * 3$.
 - Leave sufficient margins on the left and right; do not place elements too close to the edges.
2. Object Spacing
 - Maintain sufficient spacing between elements, using `buff=0.5` or larger.
 - Do not allow formulas, diagrams, labels, or text to overlap.
 - Use `.arrange()` to automatically align and space multiple elements.
3. Font Size Hierarchy
 - Main title: `font_size=56`
 - Subtitle: `font_size=40`
 - Body text: `font_size=32`
 - Notes or labels: `font_size=24`

...

[Key Requirement] Animation Sequencing Control to Prevent Overlap:

- Before introducing new content, first clear the previous content with ``FadeOut``.
- Alternatively, use ``Transform`` or ``ReplacementTransform`` to replace old content.
- Do not stack multiple unrelated elements on top of each other.
- Very important: Do not use ``FadeOut`` to clear the screen at the end of a shot.

...

Layout Checklist:

- [] Are elements centered or reasonably aligned?
- [] Is there sufficient spacing between elements?
- [] Is the font size hierarchy clear?
- [] Are the colors visually coordinated?
- [] Are there any overlapping elements?

Figure 8: Constrained Layout Prompt for Instructional Video Generation.

Knowledge Points

Mathematics

Fermat's Theorem, Newton--Raphson Method, Arithmetic Sequence, Congruent Triangle Theorem, Eccentricity, Euler's Formula, Geometric Sequence, Heron's Formula, Inverse Function, Laws of Exponents

Computer Science

Bubble Sort, Deadlock, Huffman Coding, IEEE Floating-Point Representation, ALOHA Protocol, Properties of Binary Trees, Pigeonhole Principle, Karnaugh Map, Run-Length Encoding, Signal-to-Noise Ratio

Chemistry

Transition Metals, Chemical Equation Balancing, Combustion Analysis, Crystallization, Distillation, First Law of Thermodynamics, Law of Conservation of Mass, Noble Gases, Oxidation-Reduction Reactions, Titration

Physics

Avogadro Constant, Color and Wavelength, Coulomb's Law, Dalton's Law of Partial Pressures, Universal Gravitation, Kinetic Energy, Newton's First Law of Motion, Newton's Second Law of Motion, Ohm's Law, Archimedes' Principle

Figure 9: Knowledge Points in Experiment.

Transcript Evaluation Prompt

You are a specialist in evaluating theorem explanation videos, known for giving clear and objective feedback. You will be given the transcript of a video. Your task is to evaluate and score the content of the video in several dimensions.

Evaluation Criteria:

1. Accuracy and Depth

- Does the narration explain the theorem accurately?
- Does the video provide intuitive and/or rigorous explanations for why the theorem holds?

2. Logical Flow

- Does the video follow a clear and logical structure?
- Does the video present a coherent buildup of ideas?

Figure 10: Narration Evaluation Prompt.

Visual Layout Evaluation Prompt

Please evaluate the image in terms of Spatial Clarity and Layout quality. Assign a score from 0 to 10 based on the following dimensions:

Evaluation Dimensions:

1. Visual Cleanliness (40%)

- Is the background clean and free of distracting or cluttered elements?

2. Occlusion and Overlap (30%)

- Do text, diagrams, or graphical elements overlap or block each other?

3. Layout Comfort and Readability (30%)

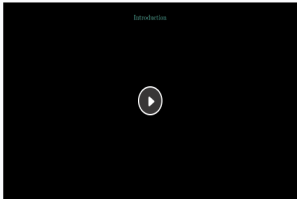
- Is the layout well organized?
- Is whitespace used appropriately?
- Is the visual flow clear and easy to follow?

...

Figure 11: Visual Evaluation Prompt.

化学·教学视频对比 (本次讲解的知识点为“过渡金属”)

Q1 请观看关于“过渡金属”相关的讲解视频B



Q2 视频B的讲解内容清晰、易于理解

☐ 非常不同意

☐ 不同意

☐ 一般

☐ 同意

☐ 非常同意

Q3 视频B的呈现方式具有吸引力, 能够保持我的注意力

☐ 非常不同意

☐ 不同意

☐ 一般

(a) Chinese questionnaire

Mathematics Teaching Video Comparison (This video explains the "Newton-Raphson Method")

Q1 Please watch video B explaining the "Newton-Raphson method".



Q2 Video B's explanations are clear and easy to understand.

☐ strongly disagree

☐ disagree

☐ neutral

☐ agree

☐ completely agree

Q3 Video B was presented in an engaging way that kept my attention.

☐ strongly disagree

(b) English questionnaire

Figure 12: Online questionnaire interface used in the expert evaluation study (Chinese and English versions).

Which of the following statements about transition metals is incorrect?

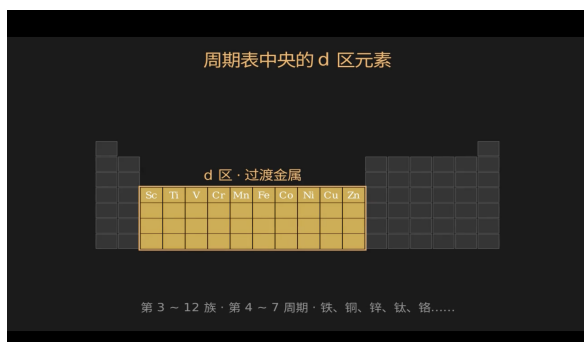
- A. Transition metals are mainly located in the d-block of the periodic table, corresponding to Groups 3--12.
- B. Transition metals often exhibit multiple oxidation states. For example, manganese can exist in oxidation states such as +2, +3, +4, +6, and +7.
- C. The colors of transition metal compounds are related to electron transitions among partially filled d orbitals.
- D. The valence electrons of transition metals come only from the outermost ns orbital, while the $(n-1)d$ orbitals do not participate in chemical reactions.

...

You are minimizing a differentiable loss function and you are currently at a point where the slope (gradient) is positive. To reduce the loss using gradient descent, which update behavior is most appropriate, and what would you expect if the learning rate is set far too large?

- A. Move in the negative gradient direction; an overly large learning rate can overshoot the minimum and cause oscillation or divergence.
- B. Move in the positive gradient direction; an overly large learning rate will converge faster without risk.
- C. Do not move because a positive gradient indicates you are already at the minimum; an overly large learning rate has no effect.
- D. Move in a random direction to escape local minima; an overly large learning rate mainly slows progress but remains stable.

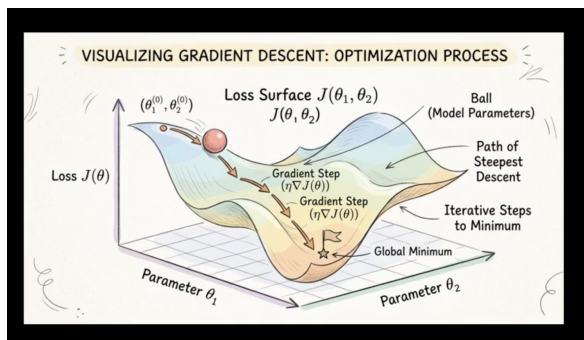
Figure 13: Post-instruction assessment questions generated from instructional videos on transition metals and gradient descent.



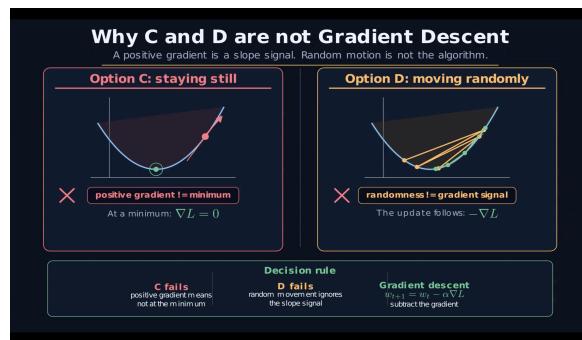
(a) Chinese chemistry instructional video



(b) Chinese chemistry remediation video



(c) English gradient descent instructional video



(d) English gradient descent remediation video

Figure 14: Qualitative case studies generated under bilingual educational settings.

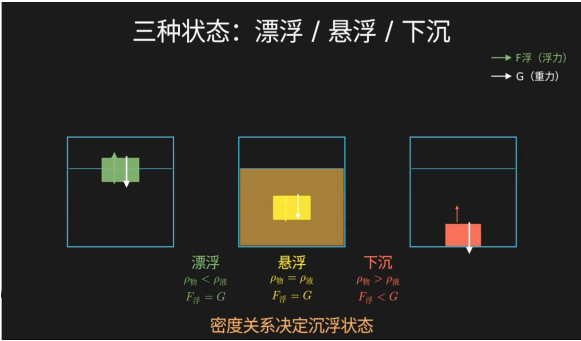
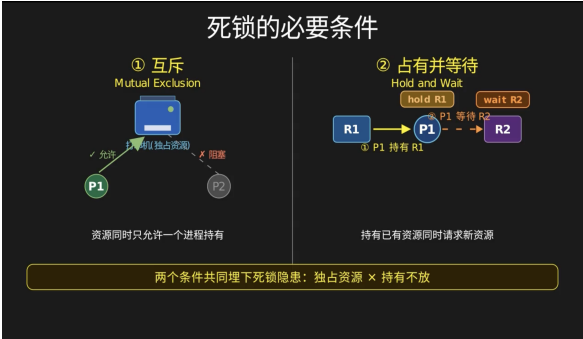


Figure 15: Representative Chinese instructional videos generated by PIVOT framework.