TeachMaster: Generative Teaching via Code

Yuheng Wang, Runde Yang, Lin Wu, Jie Zhang, Jingru Fan, Tianle Zhou, Ruoyu Fu, Huatao Li, Ruijie Shi, Siheng Chen, Weinan E, Chen Qian✉
SAI School of Artificial Intelligence, Shanghai Jiao Tong University
yuhengwangagent@gmail.com qianc@sjtu.edu.cn
www.teachmaster.cn

Abstract

The scalability of high-quality online education is hindered by the high costs and slow cycles of manual content creation. Despite advancements in video generation, current approaches often fail to ensure pedagogical structure and precise control due to their pixel-level, black-box nature. In this paper, we propose Generative Teaching, a novel paradigm shifting educators from manual creators to high-level directors who focus on pedagogical intents while agents handle the execution. To realize this vision, we introduce TeachMaster, a multi-agent framework that leverages code as an intermediate semantic medium. Unlike traditional video generation methods, TeachMaster orchestrates a collaborative team of agents, spanning planning, design, and rendering, to automate the production of interpretable, editable, and curriculum-ready educational videos. Experiments validate that TeachMaster significantly boosts production efficiency without compromising structural coherence or visual fidelity, slashing production costs to only 0.3% of traditional online course videos and providing a robust solution for scalable education.

1 Introduction

The global education system faces significant challenges, including the uneven distribution of high-quality educators (P.Stromquist, 2018), insufficient personalization (Bloom, 1984; Kasneci et al., 2023), and lagging content updates (Antoninis et al., 2023; Meng et al., 2024), all of which limit equitable access to learning opportunities. Although the internet has enabled the widespread dissemination of digitized courses, mainstream online education platforms such as Coursera and edX remain largely confined to the distribution of pre-recorded material (Reich and Ruipérez-Valiente, 2019). Content creation heavily relies on manual design, production, modification, and recording (Xalxo et al., 2025; Guo et al., 2014), a process

Figure 1: TeachMaster: A code-centric multi-agent framework for Generative Teaching that converts pedagogical intents into ready-to-teach videos.

characterized by high production costs (averaging approximately $30,000 per course), slow update cycles, and limited scalability (Hollands and Tirthali, 2014b; Raccoon Gang, 2025).

To address these limitations, we introduce "*Generative Teaching*"[1], a novel paradigm where the educator transitions from a manual creator to a high-level director for a suite of specialized generative agents. By merely specifying pedagogical objectives, educators can trigger the autonomous creation of curriculum-ready materials, such as videos. This model abstracts away granular implementation details, shifting the workflow from manual content creation to intent-driven instruction, thereby liberating educators from the burdens of extensive preparation.

From a technical perspective, while end-to-end (E2E) video generation methods (Xing et al., 2025b; Ho et al., 2022) offer direct output, they often neglect pedagogical structure, yielding results that are uneditable and computationally intensive (Wei et al., 2025; Liu et al., 2024; Xie et al., 2024b). Conversely, approaches mimicking human software usage (Niu et al., 2024) rely heav-

[1]We term this intent-driven paradigm "Generative Teaching," or informally "Vibe Teaching," as it allows educators to focus on pedagogical intents while agents handle execution.

ily on large-scale multimodal trajectories (Rawles et al., 2023; Fan et al., 2025) and substantial training costs (Xing et al., 2025a; Xie et al., 2024a). We argue that pixel-level generation is unnecessary for this domain (Chen et al., 2025; Ku et al., 2025). Instead, we leverage the semantic reasoning, world knowledge, and generative capabilities of pre-trained foundation models (Chang et al., 2024; Wei et al., 2022; Naveed et al., 2025), proposing a novel workflow that employs code as an intermediate representation (Pang et al., 2025; Zhu et al., 2025; Surís et al., 2023). Unlike opaque black-box models, this programmatic approach (Avetisyan et al., 2024; Han et al., 2023; Gao et al., 2023) ensures content interpretability, modularity, and precise control, satisfying the rigorous requirements for scalable educational content production.

More concretely, we present TeachMaster, a multi-agent framework that accepts a lecture outline as input and automates the E2E production of educational videos. Acting as a digital production team, TeachMaster orchestrates a collaborative process where agents responsible for content planning, layout design, animation rendering, and speech synthesis work in concert. This synergy ultimately produces coherent, controllable, and scalable educational content. Experimental results across multiple languages and disciplines reveal that TeachMaster demonstrates superior efficiency without significantly compromising on quality compared to human-made content, thereby offering a more effective solution in the comprehensive trade-off between quality and production cost.

Our contributions are summarized below:

- We propose *Generative Teaching*, a novel paradigm that shifts the educator's role from manual creator to high-level director. By prioritizing pedagogical intents over technical implementation, this paradigm empowers generative agents to autonomously handle lesson planning and educational delivery.
- To realize this vision, we introduce TeachMaster, a multi-agent framework that utilizes code as an intermediate semantic medium. This approach automates the generation of educational videos, facilitating the scalable production of high-quality, interpretable learning resources.
- Extensive experiments and real-world deployment across diverse disciplines validate TeachMaster. Performance metrics and user feedback confirm superior structural coherence and cross-modal alignment. TeachMaster reduces production costs to approximately 0.3% of traditional expenses, balancing efficiency and quality.

2 Method

Generating high-quality educational videos is a multi-stage task that requires navigating a multi-modal transformation from abstract pedagogical intents to intricate content synthesis. Therefore, we introduce TeachMaster, a multi-agent system, as illustrated in Figure 2, that deconstructs the process into three sequential stages: content planning, presentation generation and quality validation. First, the system converts user inputs into page-level blueprints, establishing a semantic foundation. Next, these semantics are transformed into code to produce visual elements, which are then coupled with corresponding narration. Finally, the quality validation ensures the effectiveness of the generated videos through code verification, audio-video synchronization and layout optimization.

For clarity, the generation process can be formalized as follows: given a lecture outline k (e.g., a set of keywords) and optional configurations Φ, TeachMaster generates an output set of educational materials. The process is denoted as F:

$$\mathcal{O} = F(k, \Phi, f_{\text{human}}) \tag{1}$$

The output set $\mathcal{O} = \{V_{\text{out}}, L_{\text{out}}\}$ contains the generated video and lecture scripts.

2.1 Content Planning

Aiming for high coherence and informativeness, a *composition agent* transforms raw inputs into a comprehensive manuscript L_{out}. Inspired by the communicative reflection in ChatDev (Qian et al., 2024), it cascades semantic skeletonization (S), content expansion (E) and length refinement (R), where the latter iteratively rectifies the content to match the target video duration t:

$$L_{\text{out}} = R\Big(E\big(S(k)\big), t\Big) \tag{2}$$

Subsequently, a *pagination agent* segments L_{out} into discrete units $\{P_i\}$. To handle long-text constraints, we employ a Chain-of-Agents (CoA) framework (Zhang et al., 2024a) that partitions L_{out} into segments $\{D_i\}$ for collaborative processing and aggregation:

$$F_{\text{CoA}}(L_{\text{out}}) = \bigcup_{i=1}^{n} F_i(D_i) = \{P_i\}_{i=1}^{n} \tag{3}$$

where F_i represents a local *pagination agent*.

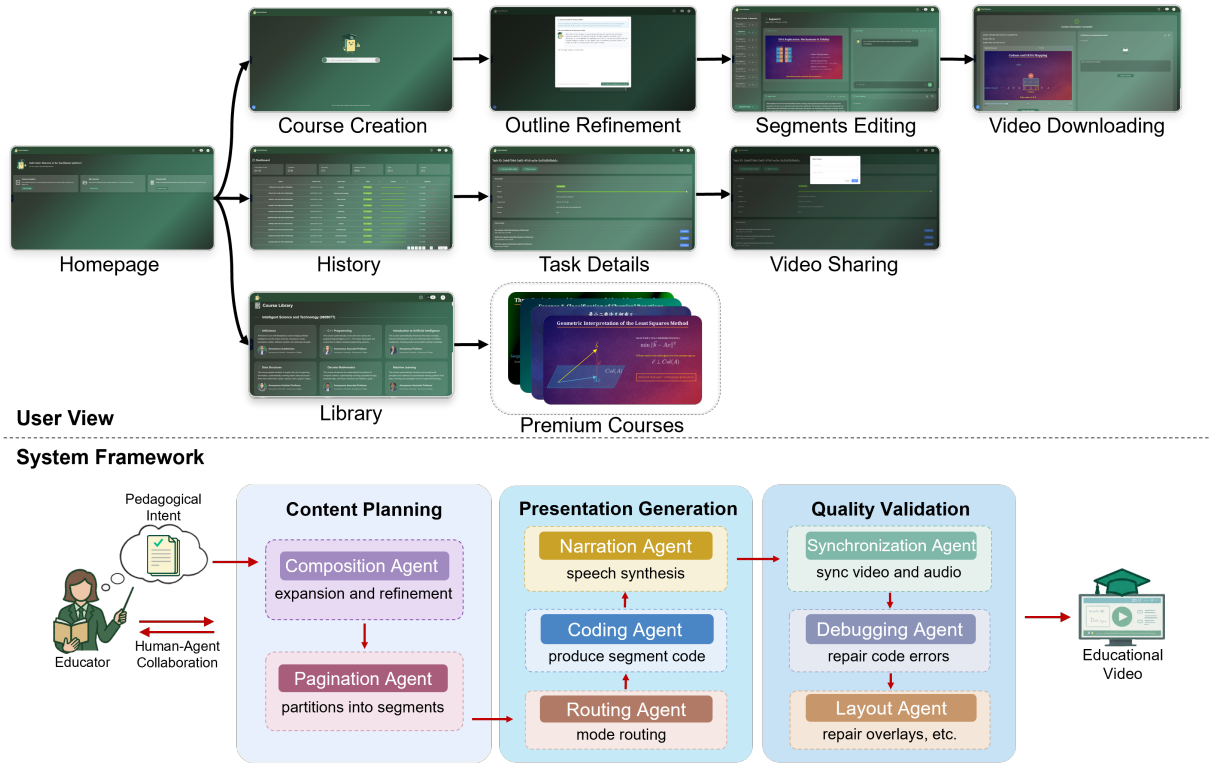


Figure 2: TeachMaster Architecture. The user view (top) offers a seamless platform for course creation and library management, while the underlying framework (bottom) manages a three-stage multi-agent workflow for content planning, presentation generation, and quality validation.

2.2 Presentation Generation

Presentation generation transforms semantics into multimodal expressions. Unlike unified streams struggling to balance visual reasoning and narrative delivery (Bandyopadhyay et al., 2024), TeachMaster decouples the process through a dual-stream mechanism. By integrating code-driven visual synthesis with context-aware audio generation, it achieves precise control over visuals and pacing.

To bridge abstract semantics and visual representation, a *routing agent* selects the optimal synthesis path for each blueprint P_i. While abstract logic is compiled by a standard *coding agent*, content requiring photorealistic assets beyond geometric primitives triggers an *image-enhanced coding agent* to integrate synthesized images as programmable objects. This process is defined as:

$$C_i = F_{\text{vis}}(P_i, \text{Mode}) \tag{4}$$

where $\text{Mode} \in \{\text{Standard}, \text{ImageEnhanced}\}$.

Building upon the visual synthesis, a *narration agent* module synthesizes the linguistic component by conditioning on the current page unit P_i, the preceding lecture script T_{i-1}, and the corresponding visual code C_i. This cross-modal reference ensures continuity in terminology and maintains high visual-verbal symmetry, formally defined as:

$$T_i = F_{\text{narr}}(P_i, T_{i-1}, C_i) \tag{5}$$

Subsequently, a *Text-to-Speech (TTS) agent* F_{tts} processes the script T_i to generate the audio track A_i while simultaneously quantifying speaking rate r_i required for downstream temporal alignment:

$$A_i, r_i = F_{\text{tts}}(T_i) \tag{6}$$

2.3 Quality Validation

Even with successful multimodal generation, automatically produced educational materials may exhibit execution instability, timing misalignment or visually distracting layouts. To ensure pedagogical quality and structural reliability, TeachMaster incorporates three specialized agents to provide multi-level quality validation: a *debugging agent*, a *synchronization agent*, and a *layout agent*.

To address potential syntax or runtime errors inherent in generative code, the *debugging agent* employs an iterative render-and-repair loop. Upon detecting a rendering failure, the system extracts the error trace and prompts the agent to rectify the specific code segment:

$$C_i^{\text{debug}} = F_{\text{debug}}\Big(C_i, \text{Error}(C_i)\Big) \tag{7}$$

If the failure persists beyond a retry threshold τ, a fallback mechanism activates, replacing complex elements with standardized templates to improve production stability.

Subsequently, the *synchronization agent* ensures temporal coherence by aligning visual dynamics with the linguistic flow. Leveraging the event anchors in the debugged code C_i^{debug} and the semantic units in the script T_i, the agent utilizes the calculated speaking rate r_i to determine precise trigger timestamps. It then injects temporal control logic (e.g., waiting statements) into the code:

$$C_i^{\text{sync}} = F_{\text{sync}}(C_i^{\text{debug}}, T_i, r_i) \tag{8}$$

This programmatic adjustment ensures both traceability and reversibility.

The *layout agent* then resolves visual clutter and occlusion by iteratively rectifying geometric conflicts. It first detects overlaps O_i, computes optimal coordinates Ω_i via a heuristic scanning order, and subsequently executes code-level adjustments:

$$\begin{aligned} O_i &= F_{\text{detect}}(C_i^{\text{sync}}), \\ \Omega_i &= F_{\text{retrieve}}(O_i, \text{dir}_{h,v}), \\ C_i^{\text{layout}} &= F_{\text{layout}}(C_i^{\text{sync}}, \Omega_i) \end{aligned} \tag{9}$$

To bridge automated efficiency with professional preference, the system incorporates a comprehensive human-in-the-loop interface offering two interaction schemes: (1) *Interactive Refinement*, enabling natural language-driven code modifications; and (2) *Direct Code Editing* for granular manual code intervention.

$$C_i^{\text{final}} = F_{\text{human}}(C_i^{\text{layout}}) \tag{10}$$

Afterwards, TeachMaster renders each page into a video segment and merges it with the narration audio to produce the complete video:

$$V_{\text{out}} = \bigcup_{i=1}^{n} \text{Render}(C_i^{\text{final}}) \tag{11}$$

3 Evaluation

Metrics. To evaluate the effectiveness of TeachMaster, we established a comprehensive framework encompassing three primary dimensions: *Video Generation Quality*, *Educational Script Quality*, and *Cross-modal Semantic Alignment* (Huang et al., 2024; Brame, 2016). Specifically, we assessed educational videos for visual clarity and pedagogical logic, validated the structural coherence and factual accuracy of the educational scripts, and measured semantic consistency between visual and textual modalities. For quantitative evaluation, we followed established VLM-based assessment paradigms for open-ended generation tasks (Liu et al., 2023; Fu et al., 2024; Zheng et al., 2023), utilizing GPT-5.2 to score all metrics on a 1–10 scale. We conducted a human-AI preference study involving a panel of three human experts across 300 randomly sampled videos. The overall agreement rate between the experts and GPT-5.2 reached 81.71%, demonstrating its validity as a scalable proxy for large-scale evaluations.

Baselines. We benchmark against two distinct references: (1) E2E: Represented by Sora 2 (OpenAI, 2025), a state-of-the-art video generation model capable of autonomously producing full educational content. We use it to demonstrate how TeachMaster's code-centric paradigm ensures pedagogical structure. (2) Human-Crafted: Represented by professional educational videos[2], serving as the gold standard.

We exclude existing AI tutoring and slide generation systems (Dan et al., 2023; Dao et al., 2021) due to a fundamental functional gap. They lack the multimodal synchronization or production scale of TeachMaster. Unlike recent educational agents (Chen et al., 2025; Ku et al., 2025) that focus on short and silent clips, TeachMaster is the first to provide a full-scale, editable solution for industrial-grade course production.

Implementation Details. For visual synthesis, the system renders animations from executable Python scripts via the Manim engine (The Manim Community Developers, 2025). Within the coding agent, TeachMaster features a switchable dual-engine configuration: it either invokes Gemini-3 (Team et al., 2025) via official APIs or leverages a locally deployed Qwen3-32B (Yang et al., 2025) specifically optimized for high-fidelity Manim code generation. The TTS agent is Minimax (Zhang et al., 2025).

To bolster the performance of Qwen3-32B, we curated a specialized dataset comprising 3735 high-quality human-annotated pairs, categorized by difficulty. Training is underpinned by a curriculum learning strategy (Wang et al., 2021). The model was trained on a computing cluster equipped with 8 $\times$ NVIDIA A800 (40GB) GPUs. Fine-tuning was

[2]We curated a dataset of highly-rated educational videos and their official captions from publicly available online video-sharing platforms to serve as high-quality human references for both visual and script quality.

Method	Quality						Efficiency		
	Spat.	Rich.	Logic.	T-I Corr.	Acc.	Overall ↑	P.T. ↓	Dur. ↑	Ratio ↓
Human	**8.22**	**7.31**	**8.38**	**8.29**	**9.24**	**8.29**	795.00	32.50	24.46
Sora 2	7.36	6.36	7.55	7.64	8.96	7.57	**3.20**	0.25	12.80
$\text{TeachMaster}_{\text{Gemini}}$	7.97	6.98	7.97	7.63	8.99	$7.91^{\dagger}$	88.43	**35.97**	**2.46**
$\text{TeachMaster}_{\text{Qwen}}$	7.42	6.42	7.49	7.66	8.94	$7.59^{\dagger}$	112.80	**32.55**	3.47

Table 1: Video Generation Quality & Efficiency Evaluation. Quality metrics include: Spat. (Spatial Clarity and Layout), Rich. (Visual Richness), Logic. (Pedagogical and Narrative Logic), T-I Corr. (Text–Image Correspondence), and Acc. (Factual Accuracy). The Efficiency section compares: P.T. (Total Production Time, mins), Dur. (Total Video Duration, mins), and Ratio (Production Time divided by Duration), which indicates the time cost required to produce one minute of content. **Bold** and underline denote the best and the second-best results, respectively.† indicates significant statistical differences ($p \leq 0.05$) between a baseline and ours.

Method	Struct.	Content Metrics			Overall ↑
	Coh.	Acc.	Comp.	Cons.	
Human	**8.90**	**9.11**	9.05	**8.32**	8.84
Sora 2	3.14	6.57	1.86	6.00	4.39
$\text{TeachMaster}_{\text{Gemini}}$	8.89	9.00	**9.67**	8.22	$\mathbf{8.95}^{\dagger}$
$\text{TeachMaster}_{\text{Qwen}}$	8.50	9.00	8.17	7.67	$\mathbf{8.34}^{\dagger}$

Table 2: Educational Script Quality Evaluation. Comparison of script generation performance. Metrics: Coh. (Narrative Coherence), Acc. (Accuracy), Comp. (Completeness), Cons. (Consistency).

Method	Cov.	Acc.	Sym.	Overall ↑
Human	8.17	7.94	8.28	8.13
Sora 2	6.64	6.59	6.73	6.65
$\text{TeachMaster}_{\text{Gemini}}$	8.63	8.11	8.57	$8.44^{\dagger}$
$\text{TeachMaster}_{\text{Qwen}}$	**8.93**	**8.57**	**8.87**	$\mathbf{8.79}^{\dagger}$

Table 3: Cross-modal Semantic Alignment Evaluation. We evaluate performance across three dimensions: Semantic Coverage (Cov.), Referential Accuracy (Acc.), and Visual–Verbal Symmetry (Sym.), which measures the coordination between visual and auditory channels.

implemented via LoRA ($r = 128, \alpha = 256$) (Hu et al., 2022) and DeepSpeed ZeRO-3 (Rasley et al., 2020) with a learning rate of 1×10^{-5}.

3.1 Qualitative Analysis

As shown in Table 1, both Gemini and Qwen powered variants of TeachMaster significantly outperform the E2E baseline and approach the quality of Human references. The generated videos exhibit superior spatial organization and visual balance, ensuring logical consistency that is crucial for education. Notably, TeachMaster supports flexible duration to meet diverse educational needs, whereas E2E models are constrained to short, uncontrollable clips that often fail to deliver comprehensive educational content.

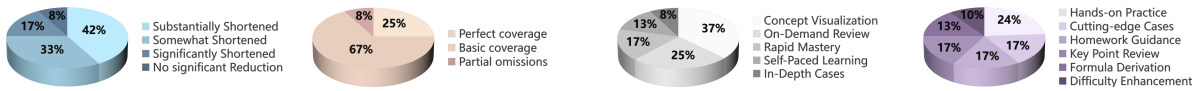
Besides, TeachMaster achieves good performance in script quality, surpassing all baselines (Table 2). The generated scripts are logically coherent and pedagogically rigorous, covering essential knowledge points without redundancy.

A key advantage of our code-centric approach is evident in cross-modal alignment, where TeachMaster outperforms both E2E and Human baselines (Table 3). It excels in semantic coverage and referential accuracy, ensuring zero information loss between modalities. The high visual–verbal symmetry score demonstrates precise coordination between visual and auditory channels, reinforcing learner perception more effectively than even human-curated content.

In terms of efficiency, TeachMaster demonstrates a decisive advantage. It requires only ∼3 minutes to generate one minute of video, which is substantially faster than E2E (>12 minutes) and the human baseline. The human time cost is based on established research into educational video production (Hollands and Tirthali, 2014a; Seitz, 2022). This significant speedup, coupled with high-quality output, highlights TeachMaster's potential for scalable and automated educational video creation.

3.2 Practical Deployment and User Feedback

Deployment TeachMaster is in active deployment on a computing cluster equipped with 8× NVIDIA A800 GPUs (each with 40GB memory). This infrastructure supports seamless switching between dual engines based on user selection. To ensure system stability under multi-user scenarios, we implemented an *asynchronous task queue* mechanism, effectively managing computational resources for simultaneous generation requests.



(a) Educator perception of preparation time reduction. (b) Video coverage of the original instructional intent. (c) Distribution of key advantages of TeachMaster over traditional methods. (d) Student suggestions for future content optimization.

Figure 3: Analysis of Real-World Classroom Feedback.

Leveraging this robust foundation, the system has demonstrated exceptional practical performance. As shown in Figure 1, it has served over 1000 educators across primary, vocational, and QS Global Top 50 universities. To date, the platform has generated more than 30,000 minutes of educational content. Based on cumulative production records spanning more than 40 distinct disciplines (See Figure 4), TeachMaster demonstrates remarkable cross-disciplinary generalization.

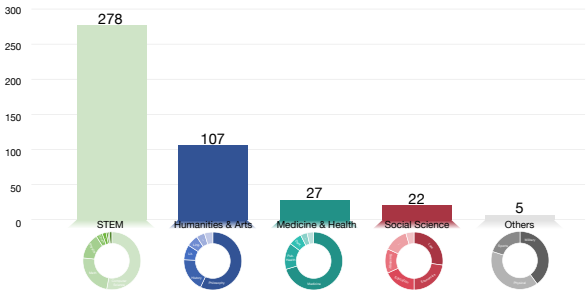


Figure 4: Quantitative distribution of academic materials across major disciplinary categories.

Notably, with this deployment setup, the cost of producing a standard 45-hour semester-long course is approximately $83.70, a mere 0.3% of traditional online course production expenses.

User Feedback As illustrated in Figure 3, we conducted surveys among both educators and students to evaluate the efficacy of TeachMaster in real-world pedagogical settings.

By prioritizing code-level editability and an automated workflow, TeachMaster significantly reduces preparation time, allowing educators to focus on high-value instructional guidance. Empirical data demonstrates that over 75.2% of pages require no manual intervention, while the remainder are finalized within an average of just 1.88 interaction rounds. This human-agent collaborative paradigm effectively overcomes the black-box limitations of conventional AI, ensuring pedagogical rigor while maintaining precise control over specialized content. Simultaneously, the system surpasses traditional lectures in both conceptual clarity and visual engagement. Students reported higher comprehension efficiency for abstract knowledge and appreciated the flexibility of self-paced review, while also expressing a growing demand for advanced modules like cutting-edge case studies and exam-tailored exercises.

3.3 Case Study

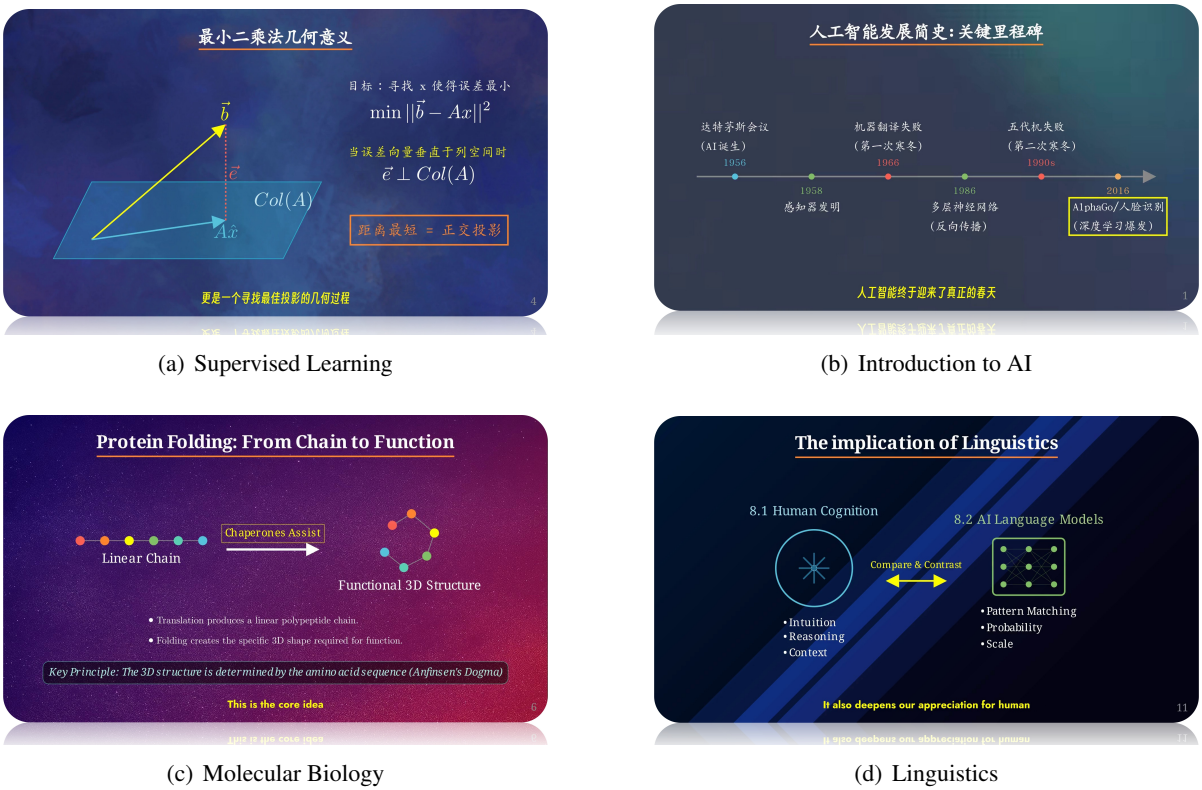
Figure 5 showcases TeachMaster's versatility through generated video clips across diverse academic disciplines. These cases, combined with real-world applications, confirm the system as a universal teaching assistant that seamlessly adapts to subject-specific requirements.

The system demonstrates robust performance in both Chinese and English contexts, exhibiting precise control over concise text generation, adaptive color coding, high-fidelity imagery, and dynamic animations. For instance, in Supervised Learning, geometric projections demystify the mathematical intuition behind the Least Squares method. Similarly, the system visualizes the transition of protein chains into functional 3D structures, facilitating an intuitive understanding of complex natural sciences.

4 Related Work

Large Language Models (LLMs) have been extended from static predictors to autonomous agents capable of planning, tool use and memory (Li et al., 2025). Early studies primarily focused on single-agent systems for specific tasks (Xi et al., 2023; Zhou et al., 2023; Huang et al., 2022; Yang et al., 2024b; Park et al., 2023; Shinn et al., 2023; Bran et al., 2023; Gou et al., 2023; Wang et al., 2025). To address the growing complexity of real-world workflows, subsequent work has explored multi-agent collaboration (Wu et al., 2024; Hong et al., 2023; Du et al., 2023; Chen et al., 2023; Dang et al., 2025; Du et al., 2025; Li et al., 2023).

Meanwhile, AI-Generated Content (AIGC) has evolved from unimodal generation toward integrated multimodal systems, covering text (Zhao et al., 2023; Li et al., 2021), code (Wang et al.,



(a) Supervised Learning

(b) Introduction to AI

(c) Molecular Biology

(d) Linguistics

Figure 5: Examples of TeachMaster-generated bilingual course materials across multiple disciplines. The system transforms structured textual outlines into complete multimodal teaching materials, including animated visuals, narration, voiceovers.

2024a; Yang et al., 2024a), and audiovisual outputs (Wu et al., 2023). Recent systems incorporate verification or refinement mechanisms for text and code generation (Wang et al., 2024b; Gao et al., 2025), while diffusion and transformer based models enable high-quality visual synthesis and voice cloning (Ho et al., 2020; Du et al., 2024), supporting E2E multimedia generation by autonomous agents (Zhang et al., 2026).

In educational contexts, earlier AI systems focused on text-based tasks (Mageira et al., 2022; Grassini, 2023; Dan et al., 2023; Dao et al., 2021; Lee et al., 2023; Yu et al., 2024; Zhang et al., 2024b). While later works explored visual generation, they relied on rigid templates (Imran and Almusharraf, 2024). Recently, agentic frameworks have emerged to automate the production pipeline (Zhang-Li et al., 2024).

5 Conclusion

We introduce TeachMaster, the pioneering framework for the Generative Teaching paradigm. Its code-centric workflow for synthesizing scripts and dynamic animations ensures transparency and editability while lowering production barriers. Within this framework, we establish a novel human-AI collaboration model: positioning AI as the technical foundation for large-scale multimodal synthesis, while empowering educators as posteriors to exercise high-level logical supervision. This approach achieves a deep complementarity between machine efficiency and human educational intuition.

Large-scale real-world deployment validates the system's exceptional balance between production efficiency and instructional quality, achieving a substantial reduction in production costs compared to traditional online courses. Ultimately, Generative Teaching will serve as a catalyst, liberating educators from repetitive content production to focus on the art of personalized instruction, while AI ensures the scalable expansion and precise presentation of knowledge.

Acknowledgements

This work was supported by the Shanghai Municipal Special Program for Basic Research on General AI Foundation Models (Grant No. 2025SHZDZX026D04).

References

Manos Antoninis, Benjamin Alcott, Samaher Al Hadheri, Daniel April, Bilal Fouad Barakat, Marcela Barrios Rivera, Yekaterina Baskakova, Madeleine Barry, Yasmine Bekkouche, Daniel Caro Vasquez, and 1 others. 2023. Global education monitoring report 2023: Technology in education: A tool on whose terms?

Armen Avetisyan, Christopher Xie, Henry Howard-Jenkins, Tsun-Yi Yang, Samir Aroudj, Suvam Patra, Fuyang Zhang, Duncan P. Frost, Luke Holland, Campbell Orme, Jakob J. Engel, Edward Miller, Richard A. Newcombe, and Vasileios Balntas. 2024. Scenescript: Reconstructing scenes with an autoregressive structured language model. In *European Conference on Computer Vision (ECCV)*.

Sambaran Bandyopadhyay, Himanshu Maheshwari, Anandhavelu Natarajan, and Apoorv Saxena. 2024. Enhancing presentation slide generation by LLMs with a multi-staged end-to-end approach. In *Proceedings of the 17th International Natural Language Generation Conference*.

Benjamin S. Bloom. 1984. The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. In *Educational Researcher*.

Cynthia J Brame. 2016. Effective educational videos: Principles and guidelines for maximizing student learning from video content. In *CBE—Life Sciences Education*.

Andres M Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. 2023. Chemcrow: Augmenting large-language models with chemistry tools. In *arXiv preprint arXiv:2304.05376*.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, and 1 others. 2024. A survey on evaluation of large language models. In *ACM transactions on intelligent systems and technology*.

Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chi-Min Chan, Heyang Yu, Yaxi Lu, Yi-Hsin Hung, Chen Qian, and 1 others. 2023. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors. In *International Conference on Learning Representations (ICLR)*.

Yanzhe Chen, Kevin Qinghong Lin, and Mike Zheng Shou. 2025. Code2video: A code-centric paradigm for educational video generation. In *arXiv preprint arXiv:2510.01174*.

Yuhao Dan, Zhikai Lei, Yiyang Gu, Yong Li, Jianghao Yin, Jiaju Lin, Linhao Ye, Zhiyan Tie, Yougen Zhou, Yilei Wang, Aimin Zhou, Ze Zhou, Qin Chen, Jie Zhou, Liang He, and Xipeng Qiu. 2023. Educhat: A large-scale language model-based chatbot system for intelligent education. In *arXiv preprint arXiv:2308.02773*.

Yufan Dang, Chen Qian, Xueheng Luo, Jingru Fan, Zihao Xie, Ruijie Shi, Weize Chen, Cheng Yang, Xiaoyin Che, Ye Tian, Xuantang Xiong, Lei Han, Zhiyuan Liu, and Maosong Sun. 2025. Multi-agent collaboration via evolving orchestration. In *arXiv preprint arXiv:2505.19591*.

Xuan-Quy Dao, Ngoc-Bich Le, and Thi-My-Thanh Nguyen. 2021. Ai-powered moocs: Video lecture generation. In *Conference on Image, Video and Signal Processing (IVSP)*.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. In *arXiv preprint arXiv:2305.14325*.

Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, and 1 others. 2024. Cosyvoice 2: Scalable streaming speech synthesis with large language models. In *arXiv preprint arXiv:2412.10117*.

Zhuoyun Du, Chen Qian, Wei Liu, Zihao Xie, Yifei Wang, Rennai Qiu, Yufan Dang, Weize Chen, Cheng Yang, Ye Tian, Xuantang Xiong, and Lei Han. 2025. Multi-agent collaboration via cross-team orchestration. In *Annual Meeting of the Association for Computational Linguistics (ACL) (Findings)*.

Jingru Fan, Yufan Dang, Jingyao Wu, Huatao Li, Runde Yang, Xiyuan Yang, Yuheng Wang, and Chen Qian. 2025. Appcopilot: Toward general, accurate, long-horizon, and efficient mobile agent. In *arXiv preprint arXiv:2509.02444*.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2024. Gptscore: Evaluate as you desire. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. PAL: program-aided language models. In *International Conference on Machine Learning (ICML)*.

Pengfei Gao, Zhao Tian, Xiangxin Meng, Xinchen Wang, Ruida Hu, Yuanan Xiao, Yizhou Liu, Zhao Zhang, Junjie Chen, Cuiyun Gao, and 1 others. 2025. Trae agent: An llm-based agent for software engineering with test-time scaling. In *arXiv preprint arXiv:2507.23370*.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Minlie Huang, Nan Duan, and Weizhu Chen. 2023. Tora: A tool-integrated reasoning agent for mathematical problem solving. In *International Conference on Learning Representations (ICLR)*.

Simone Grassini. 2023. Shaping the future of education: Exploring the potential and consequences of ai and chatgpt in educational settings. In *Education sciences*.

Philip J. Guo, Juho Kim, and Rob Rubin. 2014. How video production affects student engagement: an empirical study of mooc videos. In *Proceedings of the first ACM conference on Learning @ scale conference*.

Yucheng Han, Chi Zhang, Xin Chen, Xu Yang, Zhibin Wang, Gang Yu, Bin Fu, and Hanwang Zhang. 2023. Chartllama: A multimodal LLM for chart understanding and generation. In *arXiv preprint arXiv:2311.16483*.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. 2022. Video diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Fiona M Hollands and Devayani Tirthali. 2014a. Moocs: Expectations and reality. In *Center for Benefit-Cost Studies of Education, Teachers College, Columbia University*.

Fiona M Hollands and Devayani Tirthali. 2014b. Resource requirements and costs of developing and delivering moocs. In *International Review of Research in Open and Distributed Learning*.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, and 1 others. 2023. Metagpt: Meta programming for a multi-agent collaborative framework. In *International Conference on Learning Representations (ICLR)*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*.

Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. 2022. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *International Conference on Machine Learning (ICML)*.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, and 1 others. 2024. Vbench: Comprehensive benchmark suite for video generative models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Muhammad Imran and Norah Almusharraf. 2024. Google gemini as a next generation ai educational tool: a review of emerging educational technology. In *Smart Learning Environments*.

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, and 1 others. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. In *Learning and individual differences*.

Max Ku, Cheuk Hei Chong, Jonathan Leung, Krish Shah, Alvin Yu, and Wenhu Chen. 2025. Theoremexplainagent: Towards video-based multimodal explanations for LLM theorem understanding. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.

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 *IEEE/CVF International Conference on Computer Vision (ICCV)*.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. Camel: Communicative agents for "mind" exploration of large language model society. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Junyi Li, Tianyi Tang, Wayne Xin Zhao, and Ji-Rong Wen. 2021. Pretrained language models for text generation: A survey. In *arXiv preprint arXiv:2201.05273*.

Moxin Li, Yong Zhao, Wenxuan Zhang, Shuaiyi Li, Wenya Xie, See-Kiong Ng, Tat-Seng Chua, and Yang Deng. 2025. Knowledge boundary of large language models: A survey. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, and 1 others. 2024. Sora: A review on background, technology, limitations, and opportunities of large vision models. In *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. In *Applied Sciences*.

Yongye Meng, Wei Xu, Ziqing Liu, and Zhong-Gen Yu. 2024. Scientometric analyses of digital inequity in education: problems and solutions. In *Humanities and Social Sciences Communications*.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed

Akhtar, Nick Barnes, and Ajmal Mian. 2025. A comprehensive overview of large language models. In *ACM Transactions on Intelligent Systems and Technology*.

Runliang Niu, Jindong Li, Shiqi Wang, Yali Fu, Xiyu Hu, Xueyuan Leng, He Kong, Yi Chang, and Qi Wang. 2024. Screenagent: A vision language model-driven computer control agent. In *International Joint Conference on Artificial Intelligence(IJCAI)*.

OpenAI. 2025. Sora 2 system card.

Wei Pang, Kevin Qinghong Lin, Xiangru Jian, Xi He, and Philip Torr. 2025. Paper2poster: Towards multimodal poster automation from scientific papers. In *arXiv preprint arXiv:2505.21497*.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*.

Nelly P.Stromquist. 2018. The global status of teachers and the teaching profession. In *Education International*.

Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, and 1 others. 2024. Chatdev: Communicative agents for software development. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Raccoon Gang. 2025. How much does it cost to create an online course in 2025?

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Christopher Rawles, Alice Li, Daniel Rodriguez, Oriana Riva, and Timothy Lillicrap. 2023. Androidinthewild: A large-scale dataset for android device control. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Justin Reich and José A. Ruipérez-Valiente. 2019. The mooc pivot. In *Science*.

Steve Seitz. 2022. Graphics in 5 minutes.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Dídac Surís, Sachit Menon, and Carl Vondrick. 2023. Vipergpt: Visual inference via python execution for reasoning. In *IEEE/CVF International Conference on Computer Vision (ICCV)*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2025. Gemini: a family of highly capable multimodal models. In *arXiv preprint arXiv:2312.11805*.

The Manim Community Developers. 2025. Manim - mathematical animation framework (v0.19.0). In *10.5281/zenodo.14699705*.

Ke Wang, Junting Pan, Linda Wei, Aojun Zhou, Weikang Shi, Zimu Lu, Han Xiao, Yunqiao Yang, Houxing Ren, Mingjie Zhan, and Hongsheng Li. 2025. Mathcoder-vl: Bridging vision and code for enhanced multimodal mathematical reasoning. In *Annual Meeting of the Association for Computational Linguistics (ACL) (Findings)*.

Xin Wang, Yudong Chen, and Wenwu Zhu. 2021. A survey on curriculum learning. In *IEEE transactions on pattern analysis and machine intelligence*.

Xingyao Wang, Yangyi Chen, Lifan Yuan, Yizhe Zhang, Yunzhu Li, Hao Peng, and Heng Ji. 2024a. Executable code actions elicit better LLM agents. In *International Conference on Machine Learning (ICML)*.

Xingyao Wang, Boxuan Li, Yufan Song, Frank F Xu, Xiangru Tang, Mingchen Zhuge, Jiayi Pan, Yueqi Song, Bowen Li, Jaskirat Singh, and 1 others. 2024b. Openhands: An open platform for ai software developers as generalist agents. In *arXiv preprint arXiv:2407.16741*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. In *Transactions on Machine Learning Research*.

Jingxuan Wei, Cheng Tan, Qi Chen, Gaowei Wu, Siyuan Li, Zhangyang Gao, Linzhuang Sun, Bihui Yu, and Ruifeng Guo. 2025. From words to structured visuals: A benchmark and framework for text-to-diagram generation and editing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and Philip S. Yu. 2023. Multimodal large language models: A survey. In *IEEE Big Data*.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, and 1 others. 2024. Autogen: Enabling next-gen llm applications via multi-agent conversations. In *First Conference on Language Modeling*.

Pallavi Vijay Xalxo, James Kindo, and Praveen Kachhap. 2025. Online education ecosystem—exploring the challenges and opportunities. In *European booktitle of Education*.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yi Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, and 1 others. 2023. The rise and potential of large language model based agents: A survey. In *arXiv preprint arXiv:2309.07864*.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh Jing Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, Yitao Liu, Yiheng Xu, Shuyan Zhou, Silvio Savarese, Caiming Xiong, Victor Zhong, and Tao Yu. 2024a. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Zhifei Xie, Daniel Tang, Dingwei Tan, Jacques Klein, Tegawend F. Bissyand, and Saad Ezzini. 2024b. Dreamfactory: Pioneering multi-scene long video generation with a multi-agent framework. In *arXiv preprint arXiv:2408.11788*.

Jinbo Xing, Menghan Xia, Yuxin Liu, Yuechen Zhang, Yong Zhang, Yingqing He, Hanyuan Liu, Haoxin Chen, Xiaodong Cun, Xintao Wang, Ying Shan, and Tien-Tsin Wong. 2025a. Make-your-video: Customized video generation using textual and structural guidance. In *IEEE Transactions on Visualization and Computer Graphics*.

Zhen Xing, Qijun Feng, Haoran Chen, Qi Dai, Han Hu, Hang Xu, Zuxuan Wu, and Yu-Gang Jiang. 2025b. A survey on video diffusion models. In *ACM Computing Surveys*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. In *arXiv preprint arXiv:2505.09388*.

John Yang, Carlos E. Jimenez, Alexander Wettig, Kilian Lieret, Shunyu Yao, Karthik Narasimhan, and Ofir Press. 2024a. Swe-agent: Agent-computer interfaces enable automated software engineering. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Zhiyu Yang, Zihan Zhou, Shuo Wang, Xin Cong, Xu Han, Yukun Yan, Zhenghao Liu, Zhixing Tan, Pengyuan Liu, Dong Yu, Zhiyuan Liu, Xiaodong Shi, and Maosong Sun. 2024b. Matplotagent: Method and evaluation for llm-based agentic scientific data visualization. In *Annual Meeting of the Association for Computational Linguistics (ACL) (Findings)*.

Jifan Yu, Zheyuan Zhang, Daniel Zhang-li, Shangqing Tu, Zhanxin Hao, Rui Miao Li, Haoxuan Li, Yuanchun Wang, Hanming Li, Linlu Gong, Jie Cao, Jiayin Lin, Jinchang Zhou, Fei Qin, Haohua Wang, Jianxiao Jiang, Lijun Deng, Yisi Zhan, Chaojun Xiao, and 14 others. 2024. From mooc to maic: Reshaping online teaching and learning through llm-driven agents. In *arXiv preprint arXiv:2409.03512*.

Bowen Zhang, Congchao Guo, Geng Yang, Hang Yu, Haozhe Zhang, Heidi Lei, Jialong Mai, Junjie Yan, Kaiyue Yang, Mingqi Yang, and 1 others. 2025. Minimax-speech: Intrinsic zero-shot text-to-speech with a learnable speaker encoder. In *arXiv preprint arXiv:2505.07916*.

Daoan Zhang, Wenlin Yao, Xiaoyang Wang, Yebowen Hu, Jiebo Luo, and Dong Yu. 2026. A versatile multimodal agent for multimedia content generation. In *arXiv preprint arXiv:2601.03250*.

Yusen Zhang, Ruoxi Sun, Yanfei Chen, Tomas Pfister, Rui Zhang, and Sercan Ö. Arik. 2024a. Chain of agents: Large language models collaborating on long-context tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Zheyuan Zhang, Daniel Zhang-Li, Jifan Yu, Linlu Gong, Jinchang Zhou, Zhanxin Hao, Jianxiao Jiang, Jie Cao, Huiqin Liu, Zhiyuan Liu, Lei Hou, and Juanzi Li. 2024b. Simulating classroom education with llm-empowered agents. In *arXiv preprint arXiv:2406.19226*.

Daniel Zhang-Li, Zheyuan Zhang, Jifan Yu, Joy Lim Jia Yin, Shangqing Tu, Linlu Gong, Haohua Wang, Zhiyuan Liu, Huiqin Liu, Lei Hou, and Juanzi Li. 2024. Awaking the slides: A tuning-free and knowledge-regulated ai tutoring system via language model coordination. In *arXiv preprint arXiv:2409.07372*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, and 1 others. 2023. A survey of large language models. In *arXiv preprint arXiv:2303.18223*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, and 1 others. 2023. Webarena: A realistic web environment for building autonomous agents. In *arXiv preprint arXiv:2307.13854*.

Zeyu Zhu, Kevin Qinghong Lin, and Mike Zheng Shou. 2025. Paper2video: Automatic video generation from scientific papers. In *arXiv preprint arXiv:2510.05096*.