Preacher: Paper-to-Video Agentic System

Jingwei Liu[1,2*] Ling Yang[6†] Hao Luo[2,3] Fan Wang[2] Hongyan Li[1,4,5 ‡] Mengdi Wang[6 ‡]
[1]School of Intelligence Science and Technology, Peking University
[2]DAMO Academy, Alibaba group, 310023, Hangzhou, China
[3]Hupan Lab, 310023, Hangzhou, China
[4] National Key Laboratory of General Artificial Intelligence, Peking University
[5] PKU-Wuhan Institute of Artificial Intelligence
[6] Department of Electrical and Computer Engineering, Princeton University

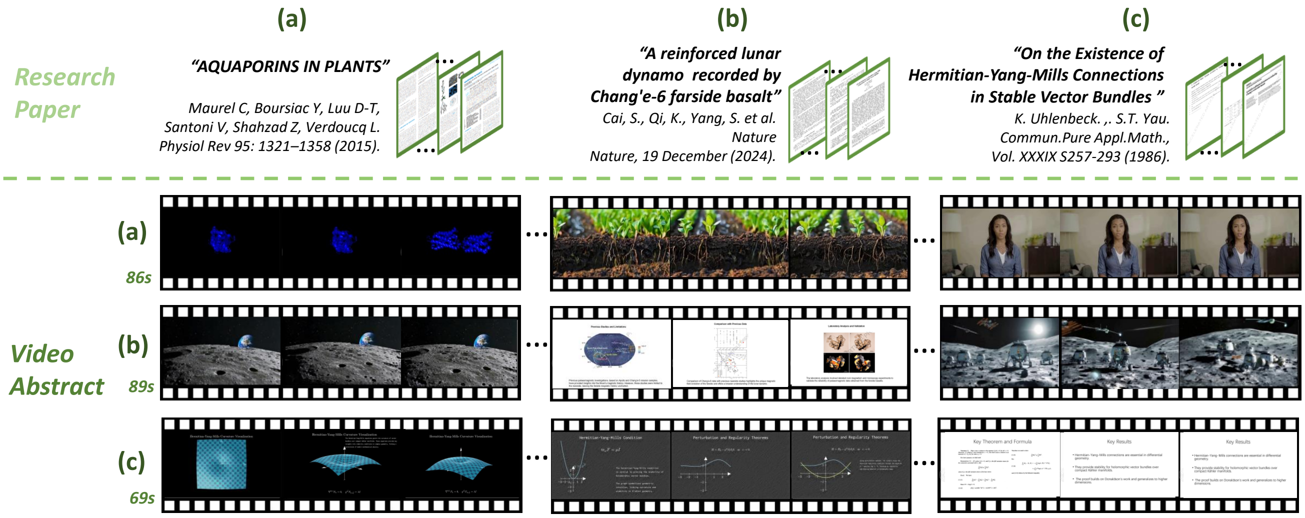

(a)
(b)
(c)
Research Paper
"AQUAPORINS IN PLANTS"
Maurel C, Boursiac Y, Luu D-T, Santoni V, Shahzad Z, Verdoucq L. Physiol Rev 95: 1321–1358 (2015).
"A reinforced lunar dynamo recorded by Chang'e-6 farside basalt"
Cai, S., Qi, K., Yang, S. et al. Nature Nature, 19 December (2024).
"On the Existence of Hermitian-Yang-Mills Connections in Stable Vector Bundles "
K. Uhlenbeck. ,. S.T. Yau. Commun.Pure Appl.Math., Vol. XXXIX S257-293 (1986).
Video Abstract
(a)
86s
(b)
89s
(c)
69s

Figure 1. Preacher can generate long video abstract conditioning on input paper with diverse topics.

Abstract

The paper-to-video task converts a research paper into a structured video abstract, distilling key concepts, methods, and conclusions into an accessible, well-organized format. While state-of-the-art video generation models demonstrate potential, they are constrained by limited context windows, rigid video duration constraints, limited stylistic diversity, and an inability to represent domain-specific knowledge. To address these limitations, we introduce Preacher, the first paper-to-video agentic system. Preacher employs a top-down approach to decompose, summarize, and reformulate the paper, followed by bottom-up video generation, synthesizing diverse video segments into a coherent abstract. To align cross-modal representations, we define key scenes and introduce a Progressive Chain of Thought (P-CoT) for granular, iterative planning. Preacher successfully generates high-quality video abstracts across five research fields, demonstrating expertise beyond current video generation models. Code will be released at: https://github.com/Gen-Verse/Paper2Video

1. Introduction

According to Scopus data [*], over three million scientific papers have been published since 2022, with an annual rise. As the volume of academic publications continues to grow, the need for effective dissemination and visibility has become increasingly critical. Among various dissemination strategies, video abstracts [18] offer a compelling means of

[*]Work done during an internship at DAMO Academy.
[†]Contributed equally. Ling Yang, yangling0818@163.com
[‡]Corresponding Authors: Hongyan Li, Mengdi Wang

[*]https://www.elsevier.com/products/scopus

communicating research findings by integrating visual and auditory elements, thereby enhancing comprehension and extending outreach. Studies have shown that papers accompanied by video abstracts receive 15% more citations [4, 14, 75]. However, producing video abstracts remains resource-intensive, requiring both domain-specific expertise and professional video production skills, making it a costly process.

Given the recent advancements in generative artificial intelligence for video generation [34, 35, 37, 50], developing an end-to-end video abstract generation model presents a compelling alternative to the high costs of manual production. While current methods can generate long-form videos exceeding 60 seconds [35], they remain unsuitable for video abstract generation. First, contemporary methods exhibit inadequate capability in directly processing research papers containing embedded multimodal elements and long contexts. Second, video generation frameworks trained on large-scale real-world video datasets [44, 45] exhibit rigid, homogeneous visual style, making them ill-suited for capturing the specialized representational demands of diverse academic disciplines.

To address these issues, we introduce Preacher, a novel paper-to-video agentic system integrating large multimodal models (LMMs), and specialized generative models. We introduce several key technologies in Preacher: **(i)** We construct a *top-down and bottom-up* structure to support the complicated modality transition. In the top-down structure, Preacher decomposes and reformulates the paper as “key scenes”, structured textual representations that encapsulate essential content while including visual descriptions to guide subsequent video generation. Serving as an intermediate bridge between textual and visual modalities, these key scenes ensure accurate content representation. In the bottom-up structure, key scenes are sequentially transformed into video segments, which are then assembled into a coherent video abstract. This structure enables precise collaboration between LMMs and generative models, effectively mitigating context window limitations while ensuring high-quality video generation. **(ii)** To enhance key scene planning and counteract the performance degradation of LMMs when handling long contexts [31, 32] or low-level detailed planning [50, 66], we introduce a progressive chain of thought (P-CoT). This method enables incremental fine-grained planning, improving coherence and scene accuracy. **(iii)** Preacher integrates video generation models with distinct styles, alongside Python-based professional visualization tools, allowing for the adaptive presentation of specialized content in the most appropriate video format. By aligning content planning with style selection, Preacher ensures that domain-specific concepts are effectively conveyed in academically relevant visual representations.

Through the top-down and bottom-up structure, multi-agent collaboration is effectively facilitated, generating high-quality video abstracts. To conduct a systematic evaluation, we employ an LMM to comprehensively evaluate the generated video abstracts across multiple dimensions, including accuracy, professionalism, aesthetic quality, and alignment with the input paper. We separately evaluate key scene planning and video generation quality, enabling direct comparison with alternative approaches. Preacher was tested on papers from five research fields and compared against state-of-the-art LMMs and video generation frameworks. Empirical results indicate that Preacher outperforms existing methods in both planning and generation, further substantiating its efficacy and applicability.

Our main contributions are as follows:

- We introduce Preacher, the first agentic system to autonomously convert papers into video abstracts.
- We develop a top-down and bottom-up structure to augment agent collaboration, and introduced key scenes bridging the gap between disparate modalities, with P-CoT enabling fine-grained key scene planning.
- We validate Preacher across five research fields, demonstrating its capability as an end-to-end solution that mitigates the high costs of manual video production and enhances knowledge dissemination.

2. Related Work

2.1. Automatic Knowledge Summary

With the advancements in LMMs [56, 67], including enhanced text comprehension and expanding context lengths [10, 13, 28], research has focused on leveraging LMMs for automated knowledge extraction and summarization [29, 43, 62]. [24] propose an end-to-end review-generation pipeline with preprocessing, modeling, and evaluation stages. Similarly, AutoSurvey [58] utilizes LMMs to retrieve and synthesize existing literature, while Tian et al. [49] introduced techniques such as clustering, dimensionality reduction, and stepwise prompting to enhance knowledge extraction from research papers. Agentic systems have also been explored for automated paper reviewing [25, 73]. However, existing methods primarily output textual summaries, which often fail to effectively convey key visual elements such as figures, charts, and experimental workflows, limiting the accessibility and impact of research findings. To address this limitation, we propose automatically generated video abstracts as a more intuitive and comprehensive alternative to traditional textual summaries.

2.2. Conditional Video Generation

Conditional video generation has been a core topic in machine learning research. Early models were constrained to 16-frame outputs [20], with subsequent approaches incorporating text-to-image diffusion models [41, 55, 65, 66] to extend generation length [26, 61]. Beyond text-based conditioning, image-conditioned generation has emerged as a complementary approach. VideoComposer [57] integrates images as control signals into the diffusion process, and VideoCrafter2 [6] leverages CLIP-derived textual and visual embeddings for cross-attention. However, these methods primarily produce simple motions and struggle with frame consistency in extended sequences, which are further improved in StreamingT2V [19] and VideoTetris [50]. Recent efforts have addressed these limitations by adopting regression-based conditioning, leveraging previous frames for improved temporal coherence in long-form video synthesis [50, 64, 69, 70].

While closed-source models remain state-of-the-art in performance [33, 35, 37], enabling generation at scales of tens of seconds, they cannot process research papers as direct inputs and fail to accommodate the stylistic diversity required for video abstracts. To bridge this gap, we integrate LMMs with a suite of heterogeneous video generation tools, forming a collaborative framework capable of processing research papers as input and producing long-form video abstracts in varied, contextually appropriate styles.

2.3. Agentic Systems

Recent advancements in LMM-based agentic system have demonstrated reasoning and planning capabilities approaching human-level performance, aligning with the expectations for autonomous agents—systems capable of perceiving environments, making decisions, and executing actions. Compared to single-agent approaches [1, 51], agentic systems harness collective intelligence and specialized expertise, enabling them to address complex challenges, including advanced programming tasks [11, 21, 30] and planning in physical environments [8, 17, 22, 46]. Several studies explore agentic systems to enhance the capabilities of generative models [2, 15, 59]. In video generation, DreamFactory [63] employs multi-agent collaboration and keyframe iteration to ensure consistency and style in long-form videos, while Mora [68] integrates human-in-the-loop feedback to refine output quality. SPAgent [53] autonomously orchestrates tools for video generation and editing through a structured three-step framework. Unlike existing approaches, our methodology advances agentic systems by introducing enhanced collaborative mechanisms, enabling the execution of cross-modal tasks that exceed the capabilities of a single agent.

3. Preliminary

Let P represent a complete and standardized academic paper, consisting of text, equations, figures, and tables. A video $V$is represented as a sequence of frames:$V = F_1, F_2, ..., F_T$, where each F_t corresponds to an image at time step t. Video abstracts may even incorporate multiple styles as a special kind of video [18] [†]. For clarity, we define V specifically as a video abstract:$V = V_1^{s_1}, ..., V_i^{s_n}, ..., V_H^{s_N}$, where $s_n \in \mathbb{S}$ and $\mathbb{S}$ is the space of all possible video abstract styles, and $V_i^{s_n}$ represents a segment of a video abstract with a specific style.

Formally, we aim to learn a generative model G that maps P into video abstract V within the video space $V = G(P)$. We construct an agentic system, decomposing G into a set of agents $\mathcal{A}$, each dedicated to a distinct subtask. These agents collaborate with each other, ensuring the generation of stylistically diverse video abstracts.

4. Preacher

Preacher is a paper-to-video agentic system integrating LMMs, LMMs, and diverse generative models. Sec. 4.1 outlines the architecture of the system and the specialization of the agents. Sec. 4.2 details the key scene planning and presents the progressive chain of thought to improve planning accuracy. Finally, Sec. 4.3 introduces how Preacher utilizes key scenes to generate video abstracts.

4.1. Overview of Preacher

Top-Down and Bottom-Up Structure Most existing cross-modal agentic systems employ a unified multi-step pipeline for cross-modal tasks [53, 68, 71, 74]. However, their performance is heavily dependent on existing text-to-visual generation models, making them ineffective for processing highly complex inputs. Inspired by prior research [11, 21, 38], we decompose and summarize input papers before feeding them into generative models. While these summaries improve compatibility, the resulting videos are often low-quality and semantically hollow. This limitation arises from insufficient detail in the summaries, preventing accurate reconstruction of key content, and from granularity constraints that hinder contemporary models' ability to fully leverage CLIP-based cross-modal mechanisms [39].

To address these challenges, we introduce a top-down and bottom-up framework, inspired by the U-Net architecture [42]. In the top-down phase, the paper undergoes decomposition and summarization into multiple raw scenes, each encapsulating core content while omitting fine details, serving as anchors for content segmentation. Analogous to the U-Net encoder, where spatial resolution is reduced while feature depth increases, we perform structured planning on these downsampled raw scenes, enriching them with higher-dimensional information.

Unlike prompt augmentation [3], Preacher's planning process continuously references the original paper, ensuring that the generated content maintains precise semantic alignment with the source material. The planning results termed key scenes, not only enhance compatibility with Preacher's

[†] https://www.animateyour.science/post/8-ways-to-make-a-video-abstract

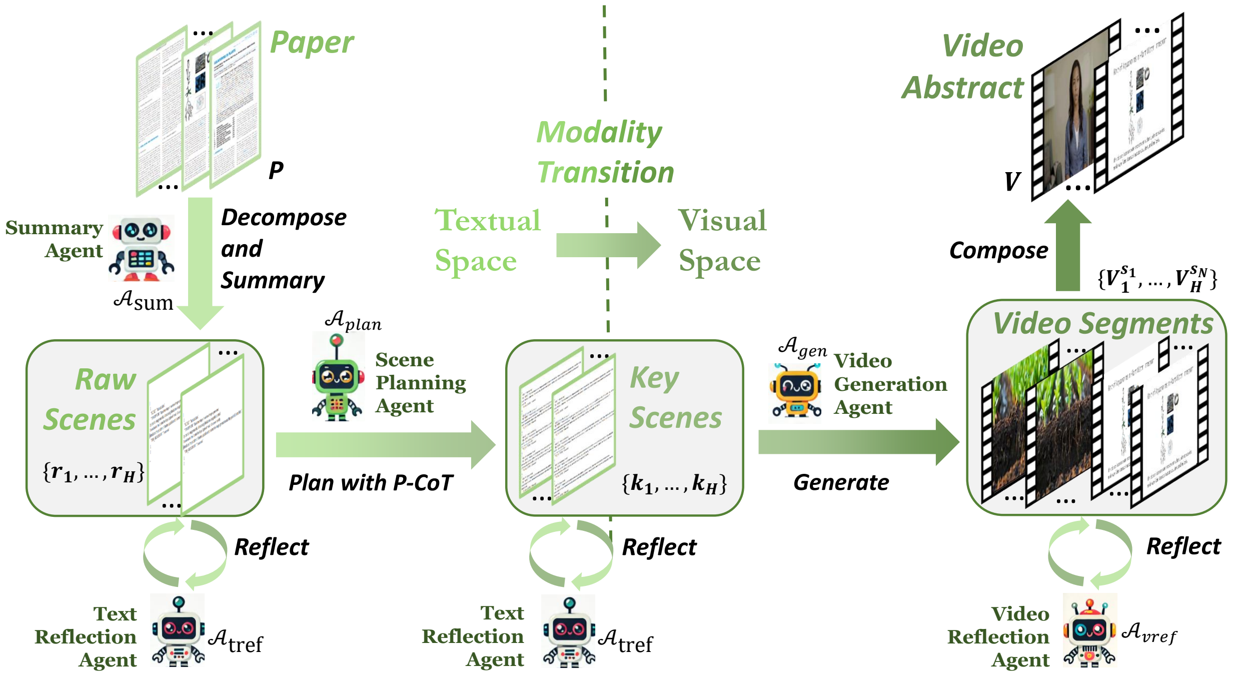


Figure 2. **Overview of Preacher**. The summary agent $\mathcal{A}_{\text{sum}}$ decomposes and summarizes the paper into H raw scenes. Subsequently, the planning agent $\mathcal{A}_{\text{plan}}$ and video generation agents $\mathcal{A}_{\text{gen}}$ then iteratively process these scenes, generating H corresponding video segments, which are subsequently assembled into a complete video abstract. For clarity, $\mathcal{A}_{\text{form}}$ is omitted, with the detailed workflow provided in Appendix B.3.

generation tools but also embed rich semantic information, providing multi-dimensional guidance for subsequent generative processes Sec. 4.2.

Within the bottom-up structure, agents equipped with video generation tools reconstruct the key scenes, generating both video and corresponding audio. Each video segment is synthesized from these elements, and upon completion, all segments are integrated into the final video abstract.

Agent Specialization Agent specialization allows agents to collaborate on tasks that a single agent cannot complete. We have six agents in the Preacher: the Summary Agent $\mathcal{A}_{\text{sum}}$, the Format Agent $\mathcal{A}_{\text{form}}$, the Scene Planning Agent $\mathcal{A}_{\text{plan}}$, the Text Reflection Agent $\mathcal{A}_{\text{tref}}$, the Video Reflection Agent $\mathcal{A}_{\text{vref}}$, the Video Generation Agent $\mathcal{A}_{\text{gen}}$. A brief description of agents follows, with more details in Sec. 5.1.

- Summary Agent $\mathcal{A}_{\text{sum}}$: This agent employs LMMs, such as GPT-4o [1] and Gemini [48], to understand, decompose and summary the input paper.
- Format Agent $\mathcal{A}_{\text{form}}$: This agent employs LLMs, such as Llama [51, 52] to format the output from $\mathcal{A}_{\text{sum}}$, ensuring the output of $\mathcal{A}_{\text{sum}}$ is correctly structured as raw scenes.
- Scene Planning Agent $\mathcal{A}_{\text{plan}}$: This agent employs LMMs, same as $\mathcal{A}_{\text{sum}}$, and its task is to provide a more detailed plan for each raw scene.
- Rule-based Reflection Agents $\mathcal{A}_{\text{tref}}$ and $\mathcal{A}_{\text{vref}}$: There are two reflection agents in Preacher: $\mathcal{A}_{\text{tref}}$ and $\mathcal{A}_{\text{vref}}$. They are both based on LMMs.
- Video Generation Agent $\mathcal{A}_{\text{gen}}$: $\mathcal{A}_{\text{gen}}$ is composed of LMMs and video generation tools, designed to generate videos with key scenes. $\mathcal{A}_{\text{gen}}$ is equipped with variable video generation tools: the Python package, text-to-image models [40, 41, 60], text-to-video models [33, 35, 37, 41, 60], talking heads generation models[7, 16, 47].

4.2. Automatic Planning of Key Scenes

Progressive CoT Planning As illustrated in Fig. 3, key scenes comprise essential elements, including duration, video style, audio content, video prompts, and corresponding sources (e.g., specific sections, figures, or equations from the original paper). Serving as an intermediary between the top-down and bottom-up structures, key scenes facilitate seamless cross-modal representation alignment. To ensure effective planning, we employ a multi-agent collaboration framework to systematically refine key scenes.

$$\{r_1, r_2, ..., r_H\} \leftarrow \mathcal{A}_{\text{form}}(\mathcal{A}_{\text{sum}}(P)) \quad (1)$$

$$k_i \leftarrow \mathcal{A}_{\text{plan}}(r_i, P), i = 1, 2, ..., H \quad (2)$$

The quality of a video abstract is highly contingent on the effective planning of key scenes. However, LMMs exhibit degraded performance in low-level planning tasks, particularly when handling long-context dependencies [32]. A common issue is the generation of partially inappropriate scenes, which, despite evaluation and re-execution, may

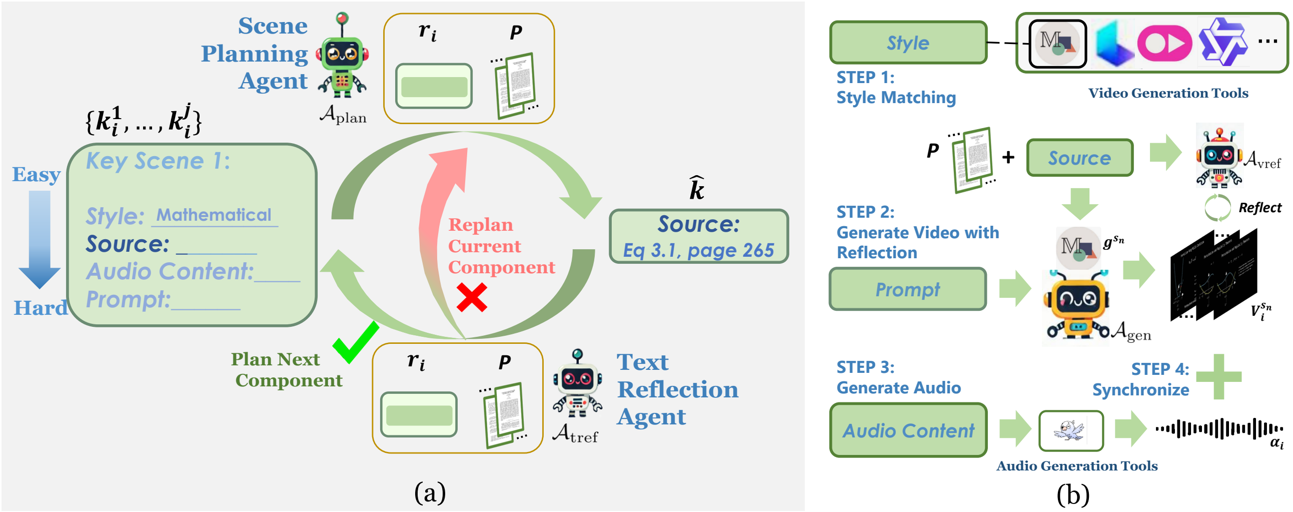


Figure 3. (a) A schematic representation of the progressive chain of thought. The key scenes consist of multiple components requiring systematic planning. The Scene Planning Agent $\mathcal{A}_{\text{plan}}$ devises a structured plan for each component, which is then evaluated by the Text Reflection Agent $\mathcal{A}_{\text{tref}}$. Based on the reflection outcome, $\mathcal{A}_{\text{plan}}$ either advances to the next component using the existing plan or revises the current component, iterating this process until all components are effectively planned. (b) The Generation Agent $\mathcal{A}_{\text{gen}}$ utilizes the key scenes to synthesize video segments. The elements enclosed within the green frame represent the structured components of the key scenes.

correct prior errors while inadvertently introducing new ones. Additionally, after multiple rounds of re-planning, LMMs may deviate from the original task objective due to accumulated contextual drift from repeated reflections.

To address these limitations, we introduce the Progressive Chain of Thought (P-CoT), a specialized CoT framework that incorporates reflection mechanisms to enhance planning coherence. As illustrated in Fig. 3(a), when planning across J components, tasks are assigned to $\mathcal{A}_{\text{plan}}$ sequentially in a simple-to-complex order, with one component planned at a time.

$$\hat{k} \leftarrow \begin{cases} \mathcal{A}_{\text{plan}}(r_i, P) & \text{if } j = 1, \\ \mathcal{A}_{\text{plan}}(r_i, P, \{k_i^n | 1 \leq n \leq j-1\}) & \text{else.} \end{cases} \qquad (3)$$

where $\{k_i^n : 1 \leq n \leq j-1\}$ are the approved components in the i_{th} key scene and $\hat{k}$ is the plan for the current component. The agent focuses on the $\hat{k}j_i$ until it has been approved by $\mathcal{A}_{\text{tref}}$. Once $\mathcal{A}_{\text{tref}}(\hat{k})$ is approved, it is fixed and passed to $\mathcal{A}_{\text{plan}}$ to plan the subsequent components:

$$k_i^j \leftarrow \hat{k}, j \leftarrow j+1 \quad \text{if } \mathcal{A}_{\text{tref}}(\hat{k}) \rightarrow \text{Approved} \qquad (4)$$

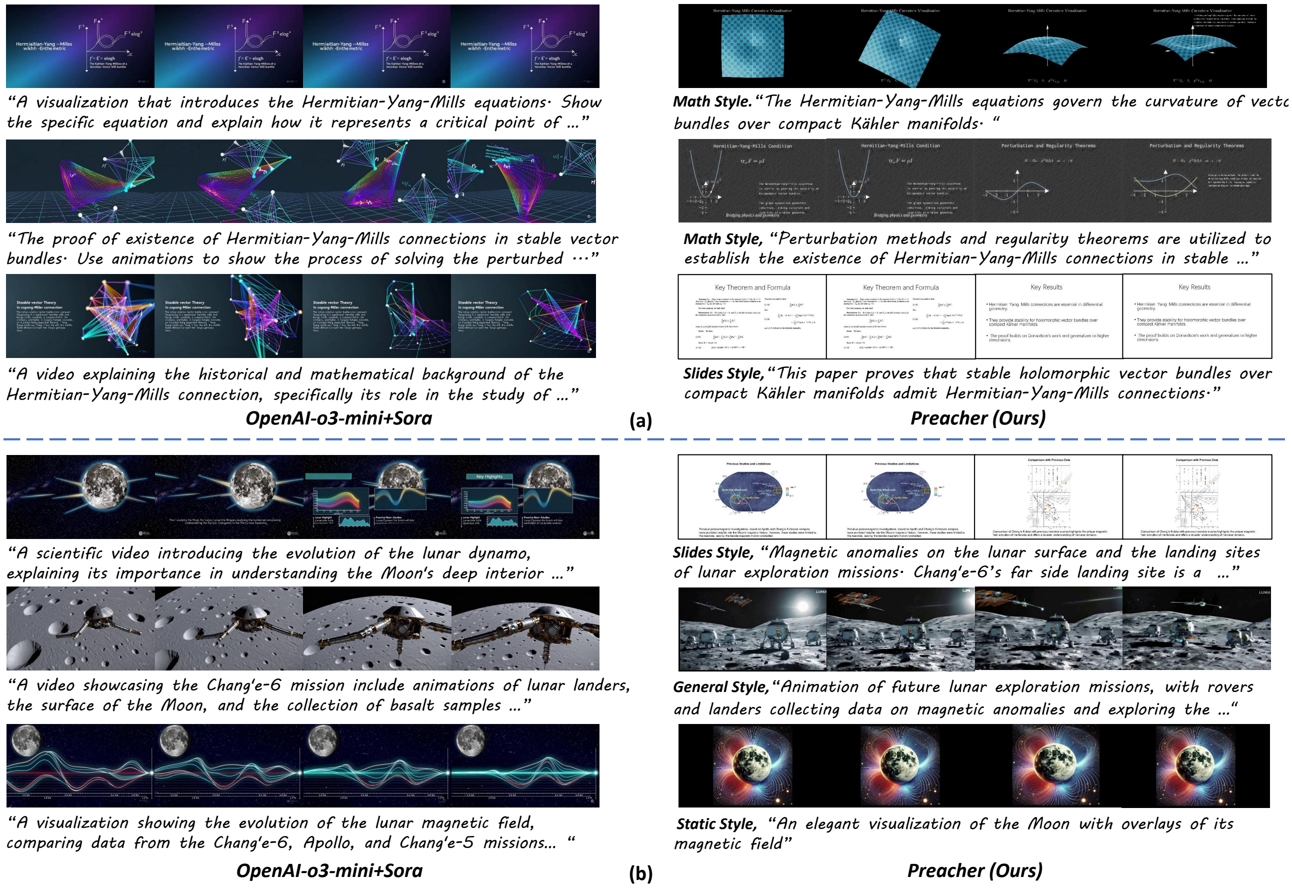
If disapproved, $\mathcal{A}_{\text{tref}}$ provides reflection to $\mathcal{A}_{\text{plan}}$ for re-planning Eq. (3). This iterative process continues until all components within the key scene are approved. The progressive complexity approach mitigates the challenges of intricate scene planning while addressing inconsistencies arising from iterative plannings.

Structured Communication between Agents While natural language communication between agents offers convenience, it is inherently unstable, as LMMs may introduce ambiguities or incomplete responses [21]. In Preacher, incompleteness in natural language-driven planning can substantially impair the effectiveness of the subsequent video generation agent. $\mathcal{A}_{\text{gen}}$.

To address this issue, we implement a structured fill-in task format, where the Format Agent $\mathcal{A}_{\text{form}}$ populates predefined dictionaries with the appropriate content. As illustrated in Fig. 2, both raw scenes and key scenes are stored as structured *json* files, ensuring consistency and reliability. Additionally, human users retain the flexibility to manually create or modify *json* files, either substituting the top-down structure or refining existing scene plans as needed.

4.3. Generating Professional Video Abstracts

While existing video generation models [35] are proficient in generating conventional scenes and motions, they encounter challenges in producing content that requires specialized knowledge, such as mathematical concepts or the structural representation of specific molecules. To mitigate this challenge, we have integrated multiple video generation tools in $\mathcal{A}_{\text{gen}}$. Upon acquiring key scene with style s_n, $\mathcal{A}_{\text{gen}}$ initially selects the appropriate video generation tool g^{s_n}. Six video styles are supported in Preacher: “talking heads,” “general,” “static concept,” “molecular visualization,” “slides,” and “mathematics.” More details about these video styles can be found in Appendix B.1. If the style of the video in the key scene is “molecular visualization”, “slides” or “mathematics”, we utilize LMMs to generate the corresponding Python code and execute it. Given the inherent susceptibility of this process to execution failures, $\mathcal{A}_{\text{tref}}$ iteratively reviews and refines the generated code to enhance its executability, ensuring successful script execution:



“A visualization that introduces the Hermitian-Yang-Mills equations· Show the specific equation and explain how it represents a critical point of …”

“The proof of existence of Hermitian-Yang-Mills connections in stable vector bundles· Use animations to show the process of solving the perturbed …”

“A video explaining the historical and mathematical background of the Hermitian-Yang-Mills connection, specifically its role in the study of …”

OpenAI-o3-mini+Sora

Math Style. “The Hermitian-Yang-Mills equations govern the curvature of vectc bundles over compact Kähler manifolds· “

Math Style, “Perturbation methods and regularity theorems are utilized to establish the existence of Hermitian-Yang-Mills connections in stable …”

Slides Style, “This paper proves that stable holomorphic vector bundles over compact Kähler manifolds admit Hermitian-Yang-Mills connections·”

Preacher (Ours)

(a)

“A scientific video introducing the evolution of the lunar dynamo, explaining its importance in understanding the Moon's deep interior …”

“A video showcasing the Chang'e-6 mission include animations of lunar landers, the surface of the Moon, and the collection of basalt samples …”

“A visualization showing the evolution of the lunar magnetic field, comparing data from the Chang'e-6, Apollo, and Chang'e-5 missions… “

OpenAI-o3-mini+Sora

Slides Style, “Magnetic anomalies on the lunar surface and the landing sites of lunar exploration missions· Chang'e-6's far side landing site is a …”

General Style, “Animation of future lunar exploration missions, with rovers and landers collecting data on magnetic anomalies and exploring the …“

Static Style, “An elegant visualization of the Moon with overlays of its magnetic field”

Preacher (Ours)

(b)

Figure 4. Comparison of the output videos generated by OpenAI o3-mini [23] + Sora [35] and Preacher. The upper and lower sections present frames in video abstract generated from (a) “On the Existence of Hermitian-Yang-Mills Connections in Stable Vector Bundles” [54] and (b) *“A reinforced lunar dynamo recorded by Chang’e-6 farside basalt”[5]*, respectively. A selection of frames has been chosen for demonstration.

$$V_i^{s_n} = g^{s_n}(\tau), \tau, s_n \in k_i \tag{5}$$

where τ denotes the code or the prompt, and g^{s_n} represents the video generation model or the execution of Python code.

To enhance the quality of the video, $\mathcal{A}_{\text{vref}}$ conducts a thorough evaluation of the generated video segment. The evaluation criteria include: (i) Accuracy, (ii) Professionalism, (iii) Alignment between the video content and the paper. If the video segment does not meet the required standards, $\mathcal{A}_{\text{vref}}$ will directly modify τ and initiate the regeneration process Eq. (5).

Once the video segment is generated, $\mathcal{A}_{\text{gen}}$ will generate the corresponding audio α_i and integrate it with the video segment. This process is repeated H times and video segments are concatenated to form the final video:

$$V \leftarrow \bigoplus_{i=1}^{H} \tilde{V}_i^{s_n}, \tilde{V}_i^{s_n} \leftarrow \text{syn}(V_i^{s_n}, \alpha_i) \tag{6}$$

where $\text{syn}(\cdot)$ and $\bigoplus$ represents the synchronization and video composition, respectively.

5. Experiments

5.1. Experimental Setup

Benchmark. To assess the effectiveness of Preacher, we constructed a benchmark dataset comprising 40 research papers spanning five distinct fields: Mathematics, Molecular Biology, Geology, Machine Learning, and Climate Science. These papers were randomly selected using GPT-4o [72], and the complete list is provided in Appendix A.

As no directly comparable baseline exists, we establish an end-to-end paper-to-video generation pipeline by integrating an LMM with a video generation model. Specifically, OpenAI-o3-mini-high [36] serves as the scene decomposition module, segmenting the input paper into multiple key scenes, while state-of-the-art video generation models synthesize 5-second video segments from these scenes. We evaluate multiple video generation models, including the open-source methods StreamingT2V [19], VideoTetris [50] and Wan-2.1-t2v-14B [60], as well as the closed-source

Table 1. Performance comparisons on forty videos in terms of ten metrics. We report mean values and standard error. The best is in bold, while the second best is underlined.

METHOD	GPT EVALUATION				HUMAN EVALUATION				CLIP ↑	AE ↑
	Accuracy ↑	Professionalism ↑	Aesthetic ↑	Alignment ↑	Accuracy ↑	Professionalism ↑	Aesthetic ↑	Alignment ↑		
OpenAI-o3-mini [36] + StreamingT2V [19]	3.35(0.98)	4.03(0.87)	4.00(0.77)	3.60(0.91)	3.13(0.93)	3.83(0.96)	3.10(1.21)	3.87(0.86)	0.23(0.04)	4.99(0.67)
OpenAI-o3-mini + Wan 2.1-14B [60]	3.75(0.43)	4.53(0.48)	4.15(0.41)	4.33(0.69)	3.63(0.91)	4.45(0.49)	**4.33**(0.51)	4.23(0.69)	0.29(0.07)	5.29(0.47)
OpenAI-o3-mini + Kling 1.6 [27]	3.70(0.61)	4.18(0.79)	3.98(0.73)	4.05(0.83)	3.40(1.06)	4.23(0.87)	4.13(0.69)	3.78(0.89)	0.26(0.07)	5.18 (0.63)
OpenAI-o3-mini + Sora[35]	4.33(0.94)	4.45 (0.49)	**4.18**(0.67)	4.30(0.73)	3.88(0.86)	4.50(0.67)	4.30(0.49)	4.38(0.59)	**0.31**(0.06)	**5.31**(0.53)
Preacher (**Ours**)	**4.50**(0.55)	**4.63**(0.44)	4.17(0.69)	**4.35**(0.98)	**4.80**(0.46)	**4.78**(0.46)	4.25(0.58)	**4.75**(0.43)	0.26(0.09)	5.20(0.83)

models OpenAI Sora [35] and Kling 1.6 [27]. To evaluate the Preacher's ability to plan key scenes, we also employed other LMMs to directly plan key scenes and use GPT-4o as the judge.

Evaluation Metrics. We utilize GPT-4 to evaluate the quality of the final video, with GPT-4 providing scores ranging from 1 to 5 in the following aspects: (i) Accuracy: Correctness of the video content, free from errors. (ii) Professionalism: Use of domain-specific knowledge and expertise. (iii) Aesthetic Quality: Visual appeal, design, and overall presentation. (iv) Alignment with the Paper: Semantic Alignment with the paper. Additionally, we use the CLIP text-image similarity score (CLIP) [39] and Aesthetic Score (AE) [45] to evaluate the consistency with the prompt and aesthetic quality. For key scene evaluation, we introduce similar metrics: Accuracy, Professionalism, Compatibility, and Alignment. Here, Compatibility measures the feasibility of directly generating scenes, reflecting the effectiveness of the planning process. All metrics are computed individually, and the results are averaged across all videos for overall evaluation. For quantitative analysis, we sample 60 frames per video to ensure consistency across evaluations.

Implementation Details. Preacher primarily integrates existing APIs and Python scripting, with no GPU requirement. We use Gemini-2.0-flash [48] as the LMM in $\mathcal{A}_{\text{sum}}$ and $\mathcal{A}_{\text{plan}}$, as Gemini's API allows the direct upload of an entire encoded PDF as context. GPT-4o is utilized for $\mathcal{A}_{\text{form}}$, $\mathcal{A}_{\text{tref}}$, and $\mathcal{A}_{\text{vref}}$, where the PDF is processed through an assistant pipeline [‡]. For $\mathcal{A}_{\text{gen}}$, we employ specialized Python libraries to generate professionally styled videos, specifically using: manim for mathematical animations, python-pptx for slide-based visualizations, and Pymol for molecular visualization. Furthermore, we employ Wan-2.1-t2i-turbo [60] as the text-to-image approach, CosyVoice2 [12] as the text-to-speech approach, Luma [33] as the text-to-video approach, and Tavus [47] as the talking-head generation approach. Specific details regarding video styles and implementation methods can be found in Appendix B.1.

5.2. Main Results

Tab. 1 compares Preacher with OpenAI o3-mini + state-of-the-art video generation models. Preacher outperforms existing methods in six out of ten metrics, notably in accuracy, professionalism, and alignment with the paper. Human evaluations further confirm Preacher's superiority, as LMMs struggle to distinguish professional content in videos. Preacher's use of domain-specific styles (e.g., mathematical visualizations, slide-based formats) may reduce scores in aesthetic quality and CLIP similarity, but this trade-off preserves scholarly integrity.

Table 2. Performance comparisons on key scenes on four metrics. We report mean values and standard error. The best is in bold, while the second best is underlined.

METHOD	Accuracy ↑	Professionalism ↑	Compatibility ↑	Alignment ↑
GPT-4o [72]	4.05(0.81)	4.30(0.71)	4.13(0.43)	4.40(0.51)
OpenAI-o3-mini [36]	4.05(0.73)	4.53(0.28)	4.20(0.56)	4.43(0.41)
Gemini-2.0-flash [48]	3.90(0.97)	4.40(0.79)	4.09(0.61)	4.35(0.49)
DeepSeek-R1[9]	4.45(0.54)	**4.68** (0.49)	3.70(1.07)	4.05(0.81)
Preacher (**Ours**)	**4.70**(0.35)	4.63(0.34)	**4.38**(0.66)	**4.50**(0.31)

Tab. 2 evaluates Preacher's key scene planning, where it leads in three out of four metrics. Chain-of-thought reasoning improves accuracy and professionalism but often results in overly complex scene plans, reducing compatibility with generative models.

Fig. 4 compares Preacher-generated video segments with those from OpenAI-o3-mini+Sora. OpenAI-o3-mini summarizes papers but lacks structured scene planning, leading to excessively complex textual descriptions that generative models struggle to process. General video generation models optimize for visual continuity and aesthetics but lack the domain-specific adaptability required for research content. In Fig. 4(a), existing methods fail to adequately convey the concept of "vector bundles" and are unable to present the critical proof from the input paper. In Fig. 4(b), although the prompt includes the crucial concept of the "lunar magnetic field," the excessive information in the prompt leads to incorrect generation, preventing the accurate representation of this important concept.

By integrating progressive planning, multi-stage reflection mechanisms, and diverse video generation tools, Preacher ensures precise content representation, preventing the propagation of erroneous information in video abstracts.

5.3. More Analysis

Ablation Study To assess the contribution of each mechanism in Preacher, we conducted comprehensive ablation

[‡] https://platform.openai.com/docs/assistants/overview

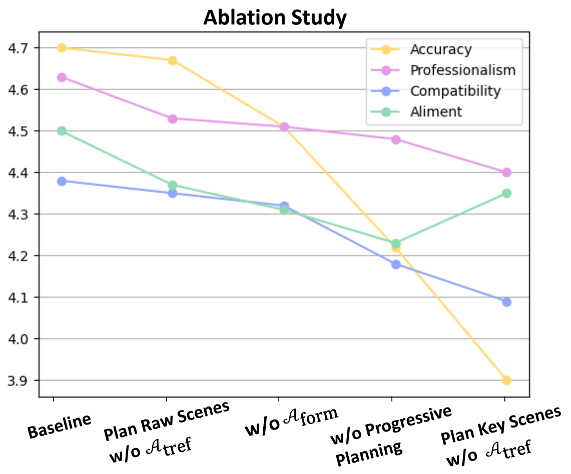


Figure 5. Ablation Study on Preacher. “w/o Progressive Planning” refers to the planning of all components in key scenes at one time.

studies, as shown in Fig. 5. Using Preacher as the baseline, we sequentially removed different mechanisms and evaluated the impact on key scene planning, following the same metrics outlined in Sec. 5.2.

Results in Fig. 5 indicate that accurate key scene planning relies on the synergistic interaction of all mechanisms. Removing any component significantly reduces accuracy, while professionalism and compatibility exhibit lower sensitivity to such omissions. Notably, excluding the reflection mechanism during key scene planning improves alignment with the input paper. This is due to multi-round reflection causing scene drift, where iterative refinements lead to deviations from the original content. The progressive generation mechanism in Preacher mitigates this by iteratively incorporating the input paper and approved key scene components, ensuring that subsequent planning remains contextually anchored and prevents divergence.

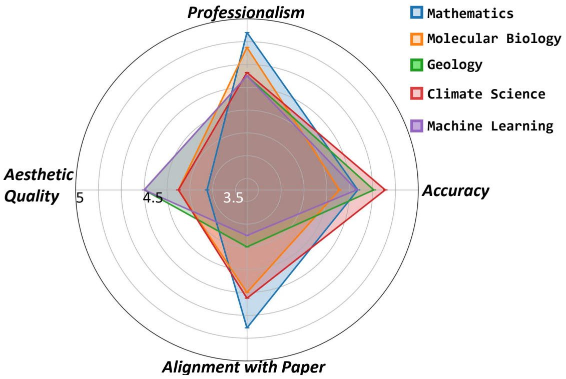


Figure 6. Performance of Preacher with paper from different research fields.

Performance on Papers from Different Research Domains Preacher generates key scenes with diverse video styles, tailored to different research domains to ensure content alignment and effective knowledge dissemination. As shown in Fig. 6, these styles produce distinct visual effects, reflecting the unique requirements of various academic disciplines. While high evaluation scores are generally observed across styles, achieving simultaneous excellence in both professionalism and aesthetics remains challenging. This trade-off likely arises from Preacher’s prioritization of content accuracy, which inherently limits the complexity of visual composition and stylistic embellishments. Moreover, certain research fields, such as mathematics and molecular biology, require precise and schematic representations, further constraining the integration of elaborate visual effects. However, as text comprehension capabilities in video generation models continue to improve, allowing for a more balanced integration of scientific rigor and visual appeal.

6. Conclusions and Limitations

Conclusions We introduce Preacher, the first paper-to-video agentic system. By leveraging a top-down and bottom-up agentic architecture, Preacher facilitates enhanced collaboration between agents. Through progressive chain-of-thought planning, Preacher systematically plans key scenes, generating high-quality video abstracts enriched with domain-specific expertise. Our evaluation across multiple research domains demonstrates Preacher’s effectiveness in representing and communicating domain-specific knowledge. In future work, we seek to broaden Preacher’s scope and applicability by integrating more video generation tools with diverse stylistic capabilities, ensuring adaptability to diverse disciplines and presentation formats.

Limitations As the first method to achieve paper-to-video generation, Preacher has several limitations. First, its multi-agent collaboration necessitates over an hour for end-to-end processing, with token consumption for inter-agent communication. Second, the absence of high-fidelity text-to-animation models restricts Preacher’s ability to generate animation-style content, limiting its visual versatility. Lastly, when processing papers in fields like artificial intelligence, key scenes are confined to “slides” and “talking heads” due to the abstract nature of such papers, which primarily comprise methodological descriptions and experimental analyses rather than concrete visualizable concepts.

7. Acknowledgments

This work is supported by the Damo Academy through Damo Academy Research Intern Program. This work is also supported by the National Natural Science Foundation of China (No.62172018, No.62102008) and Wuhan East Lake High-Tech Development Zone National Comprehensive Experimental Base for Governance of Intelligent Society.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 3, 4

[2] Hanane Allioui, Mazin Abed Mohammed, Narjes Benameur, Belal Al-Khateeb, Karrar Hameed Abdulkareem, Begonya Garcia-Zapirain, Robertas Damaševičius, and Rytis Maskeliūnas. A multi-agent deep reinforcement learning approach for enhancement of covid-19 ct image segmentation. *Journal of personalized medicine*, 12(2):309, 2022. 3

[3] Rumeysa Bodur, Binod Bhattarai, and Tae-Kyun Kim. Prompt augmentation for self-supervised text-guided image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8829–8838, 2024. 3

[4] Tristan Bonnevie, Aurore Repel, Francis-Edouard Gravier, Joel Ladner, Louis Sibert, Jean-François Muir, Antoine Cuvelier, and Marc-Olivier Fischer. Video abstracts are associated with an increase in research reports citations, views and social attention: a cross-sectional study. *Scientometrics*, 128 (5):3001–3015, 2023. 2

[5] Shuhui Cai, Kaixian Qi, Saihong Yang, Jie Fang, Pingyuan Shi, Zhongshan Shen, Min Zhang, Huafeng Qin, Chi Zhang, Xiaoguang Li, et al. A reinforced lunar dynamo recorded by chang'e-6 farside basalt. *Nature*, pages 1–3, 2024. 6

[6] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. 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, 2024. 3

[7] Jiahao Cui, Hui Li, Yao Yao, Hao Zhu, Hanlin Shang, Kaihui Cheng, Hang Zhou, Siyu Zhu, and Jingdong Wang. Hallo2: Long-duration and high-resolution audio-driven portrait image animation. *arXiv preprint arXiv:2410.07718*, 2024. 4

[8] Ishita Dasgupta, Christine Kaeser-Chen, Kenneth Marino, Arun Ahuja, Sheila Babayan, Felix Hill, and Rob Fergus. Collaborating with language models for embodied reasoning. In *NeurIPS 2022 Foundation Models for Decision Making Workshop*. 3

[9] DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. 7

[10] Yiran Ding, Li Lyna Zhang, Chengruidong Zhang, Yuanyuan Xu, Ning Shang, Jiahang Xu, Fan Yang, and Mao Yang. Longrope: extending llm context window beyond 2 million tokens. JMLR.org, 2025. 2

[11] Yihong Dong, Xue Jiang, Zhi Jin, and Ge Li. Self-collaboration code generation via chatgpt. *ACM Transactions on Software Engineering and Methodology*, 33(7):1–38, 2024. 3

[12] Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, et al. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *CoRR*, 2024. 7

[13] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 2

[14] Miguel Ferreira, Betina Lopes, António Granado, Helena Freitas, and João Loureiro. Audio-visual tools in science communication: the video abstract in ecology and environmental sciences. *Frontiers in Communication*, 6:596248, 2021. 2

[15] Rinon Gal, Adi Haviv, Yuval Alaluf, Amit H Bermano, Daniel Cohen-Or, and Gal Chechik. Comfygen: Prompt-adaptive workflows for text-to-image generation. *arXiv preprint arXiv:2410.01731*, 2024. 3

[16] Yue Gao, Yuan Zhou, Jinglu Wang, Xiao Li, Xiang Ming, and Yan Lu. High-fidelity and freely controllable talking head video generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5609–5619, 2023. 4

[17] T Guo, X Chen, Y Wang, R Chang, S Pei, NV Chawla, O Wiest, and X Zhang. Large language model based multi-agents: A survey of progress and challenges. In *33rd International Joint Conference on Artificial Intelligence (IJCAI 2024)*. IJCAI; Cornell arxiv, 2024. 3

[18] Ria Gupta, Mrudula Joshi, and Latika Gupta. An integrated guide for designing video abstracts using freeware and their emerging role in academic research advancement. *Journal of Korean Medical Science*, 36(9), 2021. 1, 3

[19] Roberto Henschel, Levon Khachatryan, Daniil Hayrapetyan, Hayk Poghosyan, Vahram Tadevosyan, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Streamingt2v: Consistent, dynamic, and extendable long video generation from text. *arXiv preprint arXiv:2403.14773*, 2024. 3, 6, 7

[20] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022. 2

[21] Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, et al. Metagpt: Meta programming for a multi-agent collaborative framework. In *The Twelfth International Conference on Learning Representations*. 3, 5

[22] Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, et al. Inner monologue: Embodied reasoning through planning with language models. In *Conference on Robot Learning*, pages 1769–1782. PMLR, 2023. 3

[23] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024. 6

[24] Hanlei Jin, Yang Zhang, Dan Meng, Jun Wang, and Jinghua Tan. A comprehensive survey on process-oriented automatic text summarization with exploration of llm-based methods. *arXiv preprint arXiv:2403.02901*, 2024. 2

[25] Yiqiao Jin, Qinlin Zhao, Yiyang Wang, Hao Chen, Kaijie Zhu, Yijia Xiao, and Jindong Wang. Agentreview: Exploring peer review dynamics with llm agents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1208–1226, 2024. 2

[26] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15954–15964, 2023. 3

[27] Kling. https://klingai.kuaishou.com/. 7

[28] Kuang-Huei Lee, Xinyun Chen, Hiroki Furuta, John Canny, and Ian Fischer. A human-inspired reading agent with gist memory of very long contexts. In *Proceedings of the 41st International Conference on Machine Learning*. JMLR.org, 2025. 2

[29] Hugo Letiche and Michael Lissack. Literature reviews with llm-based tools. *Available at SSRN 5110658*. 2

[30] Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for" mind" exploration of large language model society. *Advances in Neural Information Processing Systems*, 36: 51991–52008, 2023. 3

[31] Jiaqi Li, Mengmeng Wang, Zilong Zheng, and Muhan Zhang. Loogle: Can long-context language models understand long contexts? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16304–16333, 2024. 2

[32] Tianle Li, Ge Zhang, Quy Duc Do, Xiang Yue, and Wenhu Chen. Long-context llms struggle with long in-context learning. *CoRR*, 2024. 2, 4

[33] Luma. https://lumalabs.ai/dream-machine. 3, 4, 7

[34] Midjourney. https://www.midjourney.com/home. 2

[35] OpenAI. Sora: Creating video from text. https://openai.com/sora, 2024. 2, 3, 4, 5, 6, 7

[36] OpenAIO3-mini. https://openai.com/index/openai-o3-mini/. 6, 7

[37] PIKA. https://pika.art/home. 2, 3, 4

[38] Chen Qian, Xin Cong, Cheng Yang, Weize Chen, Yusheng Su, Juyuan Xu, Zhiyuan Liu, and Maosong Sun. Communicative agents for software development. *CoRR*, 2023. 3

[39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 3, 7

[40] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. 4

[41] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 4

[42] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 3

[43] Dmitry Scherbakov, Nina Hubig, Vinita Jansari, Alexander Bakumenko, and Leslie A Lenert. The emergence of large language models (llm) as a tool in literature reviews: an llm automated systematic review. *arXiv preprint arXiv:2409.04600*, 2024. 2

[44] Christoph Schuhmann, Robert Kaczmarczyk, Aran Komatsuzaki, Aarush Katta, Richard Vencu, Romain Beaumont, Jenia Jitsev, Theo Coombes, and Clayton Mullis. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. In *NeurIPS Workshop Datacentric AI*, number FZJ-2022-00923. Jülich Supercomputing Center, 2021. 2

[45] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022. 2, 7

[46] Chan Hee Song, Jiaman Wu, Clayton Washington, Brian M Sadler, Wei-Lun Chao, and Yu Su. Llm-planner: Few-shot grounded planning for embodied agents with large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2998–3009, 2023. 3

[47] tavus. https://https://www.tavus.io/. 4, 7

[48] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 4, 7

[49] Yangjie Tian, Xungang Gu, Aijia Li, He Zhang, Ruohua Xu, Yunfeng Li, and Ming Liu. Overview of the nlpcc2024 shared task 6: Scientific literature survey generation. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 400–408. Springer, 2024. 2

[50] Ye Tian, Ling Yang, Haotian Yang, Yuan Gao, Yufan Deng, Xintao Wang, Zhaochen Yu, Xin Tao, Pengfei Wan, Di ZHANG, et al. Videotetris: Towards compositional text-to-video generation. *Advances in Neural Information Processing Systems*, 37:29489–29513, 2024. 2, 3, 6

[51] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 3, 4

[52] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 4

[53] Rong-Cheng Tu, Wenhao Sun, Zhao Jin, Jingyi Liao, Jiaxing Huang, and Dacheng Tao. Spagent: Adaptive task decomposition and model selection for general video generation and editing. *arXiv preprint arXiv:2411.18983*, 2024. 3

[54] K. Uhlenbeck and S. T. Yau. On the existence of hermitian-yang-mills connections in stable vector bundles. *Communications on Pure and Applied Mathematics*, 1985. 6

[55] Fu-Yun Wang, Ling Yang, Zhaoyang Huang, Mengdi Wang, and Hongsheng Li. Rectified diffusion: Straightness is not your need in rectified flow. In *International Conference on Learning Representations*, 2025. 2

[56] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 2

[57] Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. *Advances in Neural Information Processing Systems*, 36:7594–7611, 2023. 3

[58] Yidong Wang, Qi Guo, Wenjin Yao, Hongbo Zhang, Xin Zhang, Zhen Wu, Meishan Zhang, Xinyu Dai, Qingsong Wen, Wei Ye, et al. Autosurvey: Large language models can automatically write surveys. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. 2

[59] Zhenyu Wang, Aoxue Li, Zhenguo Li, and Xihui Liu. Genartist: Multimodal llm as an agent for unified image generation and editing. *Advances in Neural Information Processing Systems*, 37:128374–128395, 2025. 3

[60] Tongyi Wanxiang. https://github.com/Wan-Video/Wan2.1. 4, 6, 7

[61] Wenming Weng, Ruoyu Feng, Yanhui Wang, Qi Dai, Chunyu Wang, Dacheng Yin, Zhiyuan Zhao, Kai Qiu, Jianmin Bao, Yuhui Yuan, et al. Art-v: Auto-regressive text-to-video generation with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7395–7405, 2024. 3

[62] Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Lidia Sam Chao, and Derek Fai Wong. A survey on llm-generated text detection: Necessity, methods, and future directions. *Computational Linguistics*, pages 1–66, 2025. 2

[63] Zhifei Xie, Daniel Tang, Dingwei Tan, Jacques Klein, Tegawend F Bissyand, and Saad Ezzini. Dreamfactory: Pioneering multi-scene long video generation with a multi-agent framework. *arXiv preprint arXiv:2408.11788*, 2024. 3

[64] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Gongye Liu, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Dynamicrafter: Animating open-domain images with video diffusion priors. In *European Conference on Computer Vision*, pages 399–417. Springer, 2025. 3

[65] Ling Yang, Jingwei Liu, Shenda Hong, Zhilong Zhang, Zhilin Huang, Zheming Cai, Wentao Zhang, and Bin Cui. Improving diffusion-based image synthesis with context prediction. *Advances in Neural Information Processing Systems*, 36:37636–37656, 2023. 2

[66] Ling Yang, Zhaochen Yu, Chenlin Meng, Minkai Xu, Stefano Ermon, and Bin Cui. Mastering text-to-image diffusion: Recaptioning, planning, and generating with multimodal llms. In *Forty-first International Conference on Machine Learning*, 2024. 2

[67] Ling Yang, Ye Tian, Bowen Li, Xinchen Zhang, Ke Shen, Yunhai Tong, and Mengdi Wang. Mmada: Multimodal large diffusion language models. *arXiv preprint arXiv:2505.15809*, 2025. 2

[68] Zhengqing Yuan, Ruoxi Chen, Zhaoxu Li, Haolong Jia, Lifang He, Chi Wang, and Lichao Sun. Mora: Enabling generalist video generation via a multi-agent framework. *arXiv preprint arXiv:2403.13248*, 2024. 3

[69] Yan Zeng, Guoqiang Wei, Jiani Zheng, Jiaxin Zou, Yang Wei, Yuchen Zhang, and Hang Li. Make pixels dance: High-dynamic video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8850–8860, 2024. 3

[70] Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and Jingren Zhou. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv preprint arXiv:2311.04145*, 2023. 3

[71] Hao Zheng, Xinyan Guan, Hao Kong, Jia Zheng, Hongyu Lin, Yaojie Lu, Ben He, Xianpei Han, and Le Sun. Pptagent: Generating and evaluating presentations beyond text-to-slides. *arXiv preprint arXiv:2501.03936*, 2025. 3

[72] Zhiling Zheng, Oufan Zhang, Ha L Nguyen, Nakul Rampal, Ali H Alawadhi, Zichao Rong, Teresa Head-Gordon, Christian Borgs, Jennifer T Chayes, and Omar M Yaghi. Chatgpt research group for optimizing the crystallinity of mofs and cofs. *ACS Central Science*, 9(11):2161–2170, 2023. 6, 7

[73] Ruiyang Zhou, Lu Chen, and Kai Yu. Is llm a reliable reviewer? a comprehensive evaluation of llm on automatic paper reviewing tasks. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9340–9351, 2024. 2

[74] Kaiwen Zhu, Jinjin Gu, Zhiyuan You, Yu Qiao, and Chao Dong. An intelligent agentic system for complex image restoration problems. *CoRR*, 2024. 3

[75] Qianjin Zong, Yafen Xie, Rongchan Tuo, Jingshi Huang, and Yang Yang. The impact of video abstract on citation counts: evidence from a retrospective cohort study of new journal of physics. *Scientometrics*, 119:1715–1727, 2019. 2