Paper2Video:
Automatic Video Generation from Scientific Papers

Anonymous Author(s)
Affiliation
Address
email

Abstract

Academic presentation videos have become an essential medium for research communication, yet producing them remains highly labor-intensive, often requiring hours of slide design, recording, and editing for a short 2 to 10 minutes video. Unlike natural video, presentation video generation involves distinctive challenges: long-context inputs from research papers, dense multimodal information (text, figures, tables), and the need to coordinate multiple aligned channels such as slides, subtitles, speech, and human talker. To address these challenges, we introduce **Paper2Video**, the first benchmark of 101 research papers paired with author-created presentation videos, slides, and speaker metadata. We further design three tailored evaluation metrics—Meta Similarity, PresentArena, and *PresentQuiz*—to measure how videos convey the paper's information to audience. Building on this foundation, we propose **PaperTalker**, the first multi-agent framework for academic presentation video generation. It integrates Beamer slide generation with effective layout refinement by a novel *Tree Search Visual Choice*, cursor grounding, subtitling, speech synthesis, and talking-head rendering, while parallelizing slide-wise generation for efficiency. Experiments on Paper2Video demonstrate that the presentation videos produced by our approach are more faithful and informative than existing baselines, establishing a practical step toward automated and ready-to-use academic video generation. Our datasets, agent, and codes will be *fully open-sourced* to power the community.

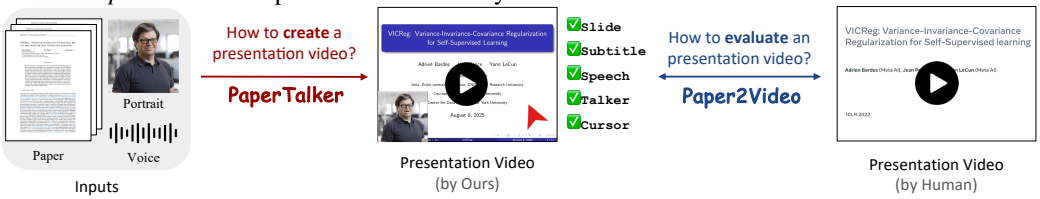


Figure 1: This work solves two core problems for academic presentations: **Left:** *how to create a presentation video from a paper?* PaperTalker – an agent integrates slide, subtitling, cursor grounding, speech synthesis, and talking-head video rendering. **Right:** *how to evaluate a presentation video?* Paper2Video – a benchmark with well-designed metrics to evaluate presentation quality.

1 Introduction

Academic presentation videos are widely used in research communication, serving as a crucial and effective means to bridge researchers, as many conferences require them as an essential material for submission. However, the manual creation of such a video is highly labor-intensive, requiring slide design, subtitle writing, per-slide recording, and careful editing, which on average may take several hours to produce a 2 to 10 minutes video for a scientific paper. Despite some prior works on slide and poster generation [20, 26, 36], automatic academic presentation video generation is a superproblem of them, a practical yet more challenging direction.

Unlike natural video generation [9, 28], presentation video exhibits distinctive characteristics, including multi-sensory integration, multi-figure conditioning, and high text density, which highlight

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

the limitations of current natural video generation models [18]. Specifically, academic presentation video generation faces several crucial challenges: *a.* It originates from long-context papers that contain dense text as well as multiple figures and tables; *b.* It requires the coordination of multiple aligned channels, including slide generation, subtitling, text-to-speech, cursor control, and talking head generation; *c.* It lacks well-defined evaluation metrics: what constitutes a good presentation video, particularly in terms of knowledge conveyance and audience accessibility. Even for the SOTA end-to-end video–audio generation model Veo3 [9], notable limitations remain in video length, clarity of dense on-screen text, and multi-modal long-document condition. In this work, we try to solve these two core problems as shown in Figure 1.

To enable comprehensive evaluation of academic presentation video generation, we present the **Paper2Video** Benchmark, comprising 101 paired research papers and author-recorded presentation videos from recent conferences, together with original slides and speaker identity metadata. Based on this benchmark, we develop a suite of metrics to comprehensively evaluate generation quality from multiple dimensions: **(i)** Meta Similarity — We employ VLM to evaluate the alignment of generated slides and subtitles with human-designed counterparts. **(ii)** PresentArena — We use a VideoLLM as a proxy audience to perform double-order pairwise comparisons between generated and human-made videos. Notably, the purpose of a presentation is to *effectively convey the information contained in the paper*. To evaluate this, we introduce **(iii) PresentQuiz**, which treats VideoLLMs as the audience and requires them to answer paper-derived questions after watching the videos.

To effectively generate ready-to-use academic presentation videos, we propose **PaperTalker**, the first multi-agent framework that enables academic presentation video generation from research papers and speaker identity. It integrates three key modules: **(i)** Slide Generation. Instead of adopting the commonly used format (*e.g.*, pptx, XML) from a template slide as in [36], we employ LaTeX Beamer code for slide generation from sketch, given its formal suitability for academic use and higher efficiency. Specifically, we employ a state-of-the-art Coder to generate code and introduce an effective **focused debugging** strategy, which iteratively narrows the scope and resolves compilation errors using feedback that indicates the relevant rows. To address the insensitivity of LLMs to fine-grained numerical adjustments, we propose a novel method called **Tree Search Visual Choice**. This approach systematically explores parameter variations to generate multiple branches, which are then concatenated into a single figure. A VLM is then tasked with selecting the optimal branch, thereby effectively improving element layouts such as figure arrangement and font size. **(ii)** Subtitling and Cursor Grounding. We generate subtitles and cursor prompts for each sentence on the slides. Then we achieve cursor spatial-temporal alignment using **Computer-use grounding model** [16, 21] models and WhisperX [1] respectively. **(iii)** Speech Synthesis and Talking-head Rendering. We synthesize personalized speech via text-to-speech models [4] and produce talking-head videos [7] for author presentations. Inspired by human recording practice and the independence between each slide, we **parallelize generation** across slides, achieving a speedup of more than $4\times$. Our multi-agent framework is implemented within the `CAMEL`[1], promoting simplicity and enabling scalability. We will open-source all our data, codebase to empower the research community.

To summarize, our contributions are as follows:

- We present Paper2Video, the first high-quality benchmark of 101 papers with author-recorded presentation videos, slides, and speaker metadata, together with evaluation metrics: Meta Similarity, PresentArena, and PresentQuiz.
- We propose PaperTalker, the first multi-agent framework for academic presentation video generation. It introduces three key modules: **(i)** tree search visual choice for fine-grained slide generation; **(ii)** a GUI-grounding model coupled with WhisperX for spatial-termporal aligned cursor grounding; and **(iii)** slide-wise parallel generation to improve efficiency.
- Experiments on Paper2Video validate the superiority of PaperTalker. PaperTalker achieves more than a 10% higher PresentArena score than Veo3 and another multi-agent methods.

2 Related Works

2.1 Video Generation

Recent advances in video diffusion models [13, 14, 29, 32, 35] have substantially improved *natural* video generation in terms of length, quality, and controllability. However, these **end-to-end** diffusion

[1] https://github.com/camel-ai/camel

Table 1: **Comparison of Paper2Video with existing benchmarks.** Top: existing natural video generation; Button: recent Agents for research works.

Benchmarks	Inputs	Outputs	Subtitle	Slides	Cursor	Speaker	
						Face	Voice
Natural Video Generation							
VBench [13]	Text	Short Vid.	✗	✗	✗	✗	✗
VBench++ [14]	Text&Image	Short Vid.	✗	✗	✗	✗	✗
Talkinghead [27]	Audio&Image	Short Vid.	✗	✗	✗	✓	✓
MovieBench [32]	Text&Audio&Image	Long Vid.	✓	✗	✗	✓	✓
Visual Agent for Research							
Paper2Poster [20]	Paper	Poster	✗	✗	✗	✗	✗
PPTAgent [36]	Doc.&Template	Slide	✗	✓	✗	✗	✗
PresentAgent [23]	Doc.&Template	Audio&Long Vid.	✓	✓	✗	✗	✗
Paper2Video (Ours)	Paper&Image&Audio	Audio&Long Vid.	✓	✓	✓	✓	✓

models still struggle to produce long videos [9, 30] (*e.g.*, several minutes), handle multiple shots, and support conditioning on multiple images [18]. Moreover, most existing approaches generate only video without aligned audio, leaving a gap for real-world applications. To address these limitations, recent works leverage **multi-agent** collaboration to generate multi-shot, long video–audio pairs and enable multi-image conditioning. Specifically, for natural videos, MovieAgent [33] adopts a hierarchical CoT planning strategy and leverages LLMs to simulate the roles of a director, screenwriter, storyboard artist, and location manager, thereby enabling long-form movie generation. Alternatively, PresentAgent [23] targets presentation video generation but merely combines PPTAgent [36] with text-to-speech to produce narrated slides. However, it lacks personalization (*e.g.*, mechanical speech and absence of a presenter) and fails to generate academic-style slides (*e.g.*, missing opening and outline slides), thereby limiting its applicability in academic contexts. Our work addresses these limitions and enable ready-to-use academic presentation video generation.

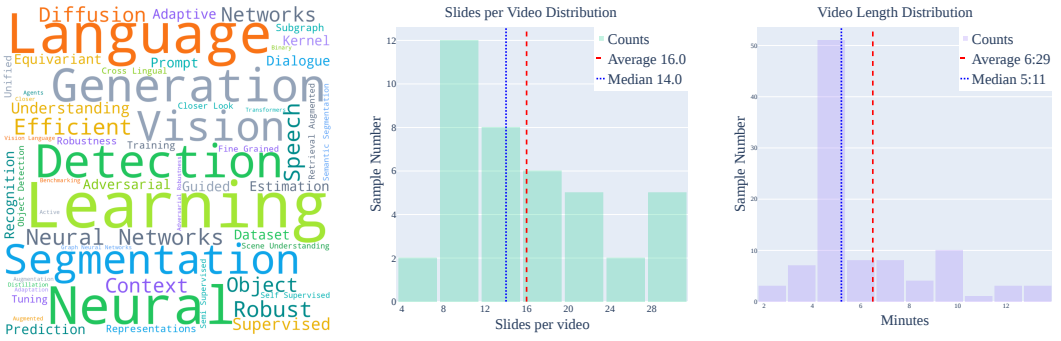
2.2 AI for Research

Many useful tasks have been explored under the umbrella of AI for Research (AI4Research) [3], which aims to support the full scholarly workflow spanning text [8], static visuals [20], and dynamic video [23]. With the breakthrough of LLMs in text generation and the Internet search ability, extensive efforts have been devoted to academic writing [2] and literature surveying [10, 11, 15, 17], substantially improving research efficiency. Besides, some works [25, 34] benchmarks AI agents' end-to-end ability to replicate SOTA ML papers, while others leavage agents to enable idea proposal [24] and data-driven scientific inspiration [6, 19]. To further enhance productivity, a growing number of work focuses on the automatic visual design of figures [31], slides [36], posters [20], and charts [12]. However, very few studies have investigated video generation for scientific purposes, leaving this area relatively underexplored. Our work belongs to one of the pioneering efforts in this direction, initiating systematic study on academic presentation video generation.

3 Paper2Video

3.1 Task Definition

Given a research paper and the author's identity information, our goal is to automatically synthesize an academic presentation video that faithfully conveys the paper's core contributions in an audience-friendly manner. We identify that a perfect presentation video is usually required to integrate four coordinated components: ***(i)*** **slides** contain well-organized, visually oriented expressive figures and tables rather than dense text; ***(ii)*** **synchronized subtitles and speech** are semantically aligned with the slides, include supplementary details, and are delivered in a clear, formal voice; ***(iii)*** **presenter** should exhibit natural yet professional facial expressions, ideally accompanied by appropriate gestures; and ***(iv)*** **a cursor indicator** aserves as an attentional anchor, helping the audience focus and follow the narrative. This task poses several distinctive challenges: ***a.*** **Multi-modal Long-Context Understanding.** Research papers span many pages with dense text, equations, figures, and tables; generating formal, well-structured slides with fine-grained layout requires content selection, cross-modal grounding, and fine-gained layout design. ***b.*** **Multi-turn Agent Tasks.** It is challenging to solve this task with a single end-to-end model, as it requires multi-channel generation and alignment (*e.g.*, slides,



(a) Word cloud of topics (b) Slides per video (c) Video length

Figure 2: **Statistics of Paper2Video benchmark.** It spans diverse topics, with presentations comprising 4–28 slides and lasting 2–12 min, providing a valuable benchmark for the automatic generation and evaluation of academic presentation videos.

cursors, and presenter). An efficient, well-designed agent system is therefore needed to address this task. ***c*. Personalized Presenter Synthesis.** The human presenter shapes credibility and audience engagement. However, achieving high-quality, identity-preserving, and lip-synchronous talking-head video remains time-consuming, and even more challenging when jointly modeling voice, face and gesture; ***d*. Spatial-Termporal-Grounding.** The cursor is a crucial visual cue for audience to follow the narrative. However, producing trajectories synchronized with narration and slide content demands discourse-level planning and precise alignment between linguistic units and visual anchors.

3.2 Data Curation

Data Source. We use AI conference papers as our data source for two reasons: (i) they offer high-quality, diverse contents across subfields with rich text, figure and table; and (ii) the field's rapid growth and open-sharing culture provide plentiful, polished author-recorded presentations and slides on YouTube and SlidesLive. However, complete metadata are often unavailable (*e.g.*, presentation videos, presenter images, and voice samples). We thus manually select papers with relatively complete metadata and supplement missing fields by sourcing presenter images from authors' websites.

In total, we curate 101 peer-reviewed conference papers from the past three years: 41 from machine learning (*e.g.*, NeurIPS, ICLR, ICML), 40 from computer vision (*e.g.*, CVPR, ICCV, ECCV), and 20 from (*e.g.*, ACL, EMNLP, NAACL). Each instance includes the paper's full LaTeX project and a matched, author-recorded presentation video comprising the slide stream and a talking-head stream with speaker identity (*e.g.*, portrait and voice sample). For 40% of the data, we additionally collect the original slide files (PDF), enabling direct, reference-based evaluation of slide generation.

Data Statics. As illustrated in Figure 2, we report the distributions of paper topics, slides per presentation, and video durations in Paper2Video. The topics spans a broad range of disciplines, evidencing substantial diversity. On average, presentations contain 16 slides and last 6 min29 s with some samples reaching up to 12 minutes. Although Paper2Video comprises 101 curated presentations, the benchmark is designed to evaluate long-horizon agentic task rather than mere video generation.

3.3 Evaluation Metrics

Unlike natural video generation, academic presentation videos serve a highly specialized role: they are not merely about visual fidelity but about communicating scholarship. This makes it difficult to directly apply conventional metrics from video synthesis (*e.g.*, FVD, IS, or CLIP-based similarity). Instead, their value lies in how well they disseminate research, amplify scholarly visibility.

From this perspective, we argue that a high-quality academic presentation video should be judged along three complementary dimensions (see Figure 3): **For the audience:** the video must faithfully convey the paper's central ideas, such as its motivation, problem formulation, and key contributions; and it should present these ideas in a manner that is easy to follow, allowing viewers from diverse backgrounds to understand the work without being overwhelmed. **For the author:** the video should foreground the authors' intellectual contribution and identity, functioning as an academic "signature"

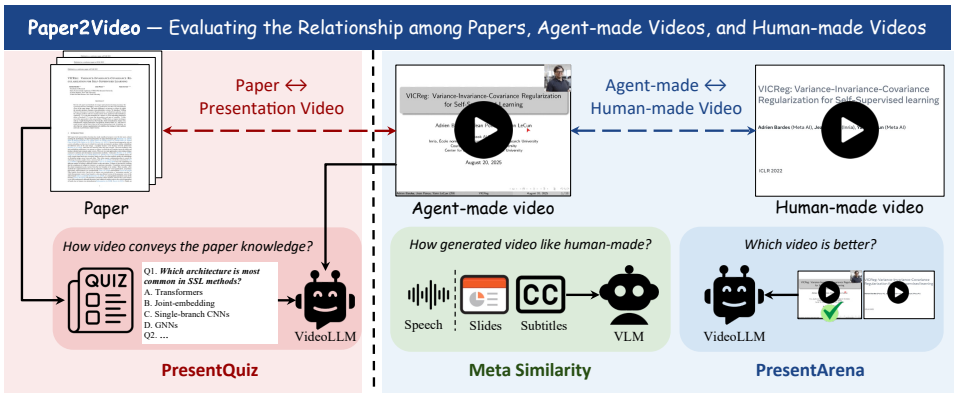


Figure 3: **Overview of evaluation metrics.** We propose three metrics that systematically evaluate academic presentation video generation from the perspective of the relationship between the generated video and **(i)** the original paper and **(ii)** the human-made video.

that enhances the work's reach and impact. To systematically capture these goals, we introduce tailored evaluation metrics specifically designed for academic presentation videos.

Meta Similarity – *How video like human-made?* As we have the ground-truth human-made presentation videos with original slides, we evaluate how well the generated intermediates (*e.g.*, slides and speech) aligned with the ones created by authors which serves as the pseudo ground-truth. For each slide, we pair the slide image with its corresponding subtitles and submit both the generated pair and the reference pair to the VLMs to obtain a similarity score on a five-point scale. To further assess vocal timbre, we uniformly sample a ten-second segment from the presentation audio, encode the generated and reference audio with a speaking embedding model [22], and compute the cosine similarity between the embeddings.

PresentArena – *Which video is better?* Similar with human audience watching the presentation, we employ the VideoLLMs as the proxy audiences to conduct pairwise comparisons of presentation videos. For each pair, the model is queried twice in opposite orders: (A, B) and (B, A). This procedure reduces hallucinations and position bias. The two judgments are then aggregated by averaging to obtain a more stable preference estimation.

PresentQuiz – *How video conveys the paper knowledge?* Following prior work [20], we evaluate information coverage using a multiple choice quiz on the presentation video. We first generate a set of questions with four options and the corresponding correct answers from the source paper. Then we ask the VideoLLMs to watch the presentation and answer each question. Overall accuracy serves as the metric, with higher accuracy indicating better information coverage. However, since generated video durations vary and longer videos can trivially include more content, we introduce a length penalty to overlong generations. In practice, consider that researchers usually intend to use concrete video to present the content thus we define a penalized weight α and aim to penalize the generated videos with too long duration.

$$\alpha = \exp\left(-\frac{\max\{0,\, L^{\text{gen}} - L^{\text{gt}}\}}{L^{\text{gt}}}\right), \tilde{\mathbf{s}} = \mathbf{s} \cdot \alpha. \tag{1}$$

where the L^{gt} and L^{gen} denote the ground-truth and generated video durations, and let s be the original accuracy score. The penalized score, $\tilde{s}$, implies that a good presentation video should try to convey paper's information within limited duration.

4 PaperTalker

Overview. To address the identified challenges, we present PaperTalker, a multi-agent framework that automatically generates presentation videos from academic papers. As illustrated in Figure 4, the pipeline comprises four stages. Starting from a paper, we synthesize slides by Beamer code and iteratively refine them with compilation feedback to correct grammar and optimize layout. The finalized slides are processed by a VLM to generate subtitles and sentence-level visual-focus cues, which are then grounded into on-screen cursor trajectories and synchronized with the narration. Finally, given a short voice sample and the portrait of the speaker, text-to-speech and talking-head modules render a realistic, personalized presentation video. For clarity, we denote p, a, v for the slides, speech audio, and human presentation video, respectively, and l, a_o, v_o for the input LaTeX project, the speaker's voice sample, and portrait.

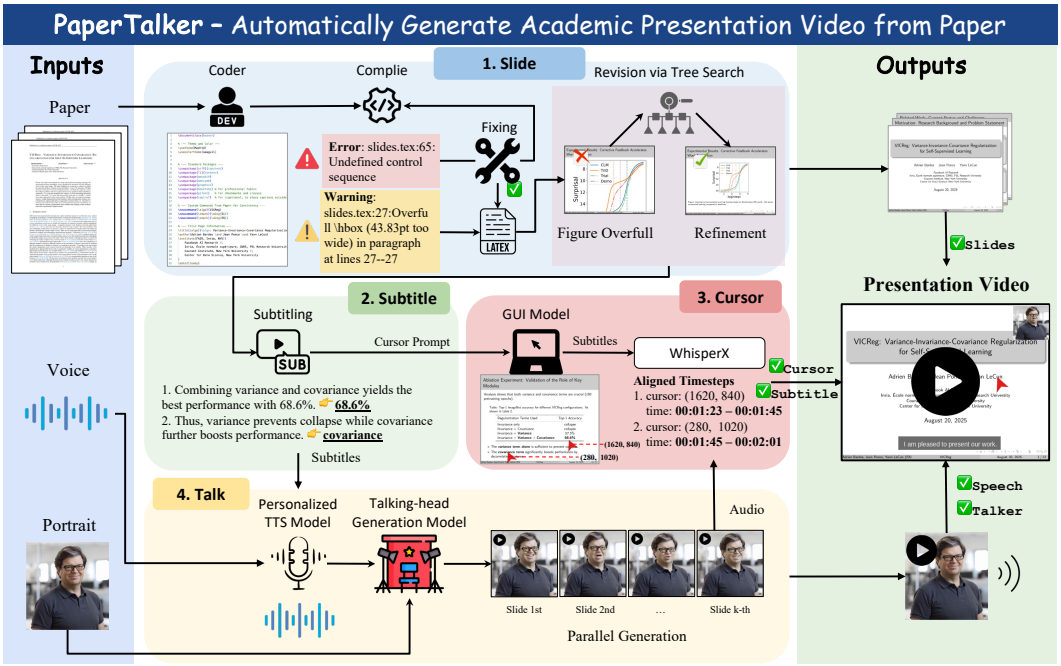


Figure 4: **Overview of PaperTalker.** Our pipeline comprises three key modules: **(i)** tree search visual choice for fine-grained slide layout optimization; **(ii)** a GUI-grounded model paired with WhisperX for spatiotemporally aligned cursor grounding; and **(iii)** slide-wise parallel generation for efficiency.

4.1 Slide Creation

A prerequisite for producing a presentation video is the creation of the slides. Despite there have some existing works [36], we target the generation of academic slides with fine-grained layouts and formal structure from scratch. Rather than selecting a template and iteratively editing it with VLMs, we generate slides directly from a paper's LaTeX project by prompting the model to write Beamer code. We adopt Beamer for three reasons: **(i)** LaTeX's declarative typesetting automatically arranges text block and figures from their parameters without explicitly planing the placement; **(ii)** Beamer is compact and expressive, representing the same content in fewer lines than XML-based formats; and **(iii)** Beamer provides well-designed, formally configured styles (*e.g.*, page numbers, section headers, hyperlinks) that are well suited to academic slide design.

As $l = (t, \mathcal{F})$ denote the project of paper, where t is the text source code and $\mathcal{F} = \{f_i\}_{i=1}^{n}$ are the figure paths. The LLM first produces a draft slide code c conditioned on $(t, \mathcal{F})$. We then compile this code to collect diagnostics, errors e and warnings w. We use these errors info to elicit a repaired, compilable version c^* . This procedure ensures that the generated Beamer code is grammatically correct, and effectively leverages and faithfully covers its content, details shown in Appendix Algorithm 1.

Although LaTeX can automatically layout the content of slides, the generated slides could sometimes still suffer from inappropriate layouts (*e.g.*, mainly overflow) due to the unsuitable figure or text front size. However, as the compilation warnings w signal potential layout issues, we are able to first use them to identify slides that require refinement.

Tree Search Visual Choice. After localizing the slides that require refinement, the key challenge is how to adjust their layouts effectively. As LLMs/VLMs fail to perceive visual feedback like human designers, we observe that prompting the them to directly tune numeric layout parameters (*e.g.*, font sizes, margins, figure scales) is ineffective: the models are largely insensitive to small numeric changes, yielding unstable and inefficient refinement, consistent with limitations of the parameter-editing strategy in PPTAgent [36]. To address this limitation, we introduce a *visual-selection* module for overflowed slides. The module first constructs the neighborhoods of layout variants for the current slide by rule-based adjusting the figure and text parameters, renders each variant to an image, and then uses the VLMs as judge to score candidates and select the best candidate with the best layout. Specifically, for text-only slides, we sweep the font size; for slides with figures, we first vary the figure scaling factors (*e.g.*, 1.25, 0.75, 0.5, 0.25) and then reduce the font size, details shown in Figure 5 and

Appendix Algorithm 2. These edits are straightforward in LaTeX Beamer, whose structured syntax automatically reflows content under parameter changes. This decouples discrete layout search from semantic reasoning and reliably resolves overfull cases with minimal time and token.

4.2 Subtitle Creation

As the speech should follow the slides, given the generated slides p, we rasterize them into images and pass them to a VLM, which produces sentence-level subtitles $b_{j,i}$ and its visual-focus prompt $d_{j,i}$ for each slide. The visual-focus prompt serves as an intermediate representation linking speech a to the cursor c, enabling precise temporal and spatial alignment of cursor with the narration so that to improve audience guidance which will be discussed in Section 4.4.

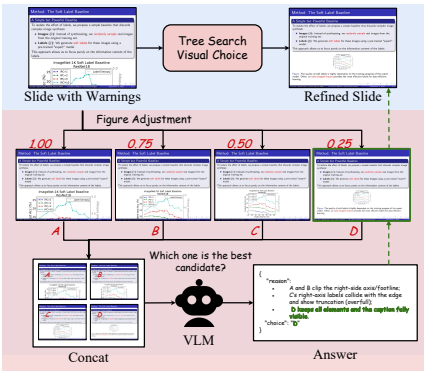


Figure 5: **Tree Search Visual Choice.** It combines a rule-based proposal mechanism with VLM-based scoring to select the optimal candidate.

4.3 Presenter Animation

The presenter video is vital for audience engagement and conveying the researcher's scholarly identity (*e.g.*, face and voice). Given the subtitles, the author's portrait v_o, and a short voice sample a_o, our objective is to synthesize a presenter video that delivers the slide content in the author's voice, with faithful identity preservation and lip–audio synchronization.

Subtitle-to-Speech. Given subtitles and speaker's voice sample, we use F5-TTS [5] model to generate speech audio per slide,

$$a_i = \mathrm{TTS}\big(\{\, b_{j,i} \,\}_{j=1}^{m_i} | a_o\big)\,, \quad i = 1, \ldots, n, \tag{2}$$

where m_i is the number of the sentence in slide i and n is the number of the slides.

Parallel Talkinghead Generation. To balance fidelity and efficiency, we use Hallo2 [7] for head-only synthesis and employ FantasyTalking [30] when upper-body articulation is required considering its higher computational cost. A persistent challenge is scalability: generating only a few minutes of talking-head video typically takes several hours, and some models(*e.g.*, FantasyTalking), do not yet natively support long-video generation. Inspired by the common practice of slide-by-slide recording and the independence between each slides, we synthesize the presenter video on a per-slide basis. Specifically, for each slide s_i, given the audio condition a_i and portrait v_o, we generate an independent clip v_i and execute these jobs in parallel, markedly reducing generation time,

$$v_i = \mathcal{G}(a_i, v_o)\,, \quad i = 1, \ldots, n, \tag{3}$$

where $\mathcal{G}$ represents the talking-head generation model. This design is justified because slide transitions are hard scene changes, temporal continuity of the presenter across adjacent slides is unnecessary.

4.4 Cursor Highlighting

Spatial-Temporal Grounding. In practice, presenters leverage the cursor as an attentional guide: a well-designed trajectory minimizes extraneous cognitive load, helps the audience track the presentation, and keeps focus on the key content. However, automatic cursor-trajectory grounding is nontrivial, requiring simultaneous alignment to the timing of speech and the visual semantics of slide. To simplify the task, we assume that the cursor will stay still within a sentence and only move between the sentences. Thus, we estimate a per-sentence cursor location and time span. For spatial alignment, motivated by strong computer-use models [16, 21] which simulate user interaction with the screenshot, we propose to ground the cursor location for each sentence with the visual focus prompt $d_{i,j}$ by UI-TARS [21]. To achieve precise temporal alignment, we then use WhisperX [1] to extract word-level timestamps and align them with the subtitles to derive the start and end times of each cursor segment.

5 Experiments

5.1 Baseline and Settings

We evaluate three categories of baselines: **(i) End-to-end Methods**, where a natural video generation model produces the presentation video directly from the paper; **(ii) Multi-Agent Frameworks**, which

Table 2: **Detailed evaluation result of Paper2Video across three baselines.** PaperTalk* represents a simple version without presenter and cursor. **Bold** and Underline indicates the best and the second-best. Detail-P and Under.-P represent the corresponding QA scores after video length penalty.

Method	Similarity↑		Arena↑	PresentQuiz Accuracy↑			
	Speech	Content		Detail	Detail-P	Under.	Under.-P
Veo3 [9]	0.133	NA	1.20%	0.367	0.367	0.585	0.585
PresentAgent [23]	0.045	2.23	2.00%	**0.878**	0.515	**0.961**	0.558
PaperTalk*	**0.646**	**2.38**	7.81%	0.839	0.795	0.949	0.911
PaperTalk (Ours)	**0.646**	**2.38**	**21.0%**	0.842	**0.824**	0.952	**0.940**

combine slide generation with text-to-speech generation and then compose them into a presentation video; and **(iii) PaperTalker**, our method and its variants. For the VLM and VideoLLM, we choose *GPT-4.1-mini* and *Gemini-2.5-Flash* respectively for a favorable efficiency and performance trade-off. We performed inference using eight NVIDIA RTX A6000 GPUs for talking-head video generation.

5.2 Main Results

Meta Similarity. We evaluate the alignment of the generated slides, subtitles, and speech with their human-authored counterparts. For speech, we randomly sample a 10-second audio segment from each method and compute the cosine similarity between the its embeddings [22] and those of the author's speech. As reported in Table 2, PaperTalker achieves the highest similarity among all baselines in terms of both speech and content similarity. We attribute this performance to personalized TTS and our slide-generation design: **(i)** adopting *Beamer*, which provides formal, academically styled templates while LaTeX automatically arranges content within each slide; and **(ii)** a tree search visual choice layout refinement that further enforces fine-gained slide layouts as commonly observed in human-authored slides.

PresnetArena. We compare the presentation videos generated by each method against human-made videos. As an automatic evaluator, we prompt the VideoLLMs as judge to determine which presentation is better with respect to clarity, delivery, and engagement. As shown in Table 2, PaperTalker attains the highest pairwise win rate among all baselines, indicating that our method produces presentation videos with superior overall perceived quality.

PresentQuiz. To assess information coverage, we conduct a VideoQA evaluation. We construct QA sets by prompting an LLM to generate questions targeting (i) fine-grained details and (ii) higher-level understanding. As shown in Table 2, PaperTalker attains the best performance in terms of penalized scores. PaperTalker's gains stem from our slide-generation design and our principled cursor-trajectory synthesis, which together guide attention and enable accurate grounding of key content for the VideoLLMs. Although PresentAgent attains slightly higher unpenalized accuracy, its length-penalized score is markedly lower owing to the excessive duration of its generated videos compared with human-made videos.

Table 3: **Generation cost for each method.** PaperTalker^s denotes PaperTalker with serial talking-head generation.

Method	Token (K) ↓	Time (min.)↓	Cost ($)↓
Veo3 [9]	NA	**0.4**	1.667
PresentAgent [23]	241	39.5	0.003
PaperTalker^s	**62**	287.2	**0.001**
PaperTalker	**62**	64.8 (**4**×)	**0.001**

Efficiency Analysis. As shown in Table 3, PaperTalker achieves the lowest cost. This efficiency stems from our slide-generation design: adopting *Beamer* reduces token usage for slide creation, and our visual-selective layout refinement is a lightweight post-processing step. Runtime is further reduced by our parallel talking-head generation mechanism. By contrast, PresentAgent [23] incurs higher token costs due to frequent refinement queries during slide editing.

5.3 Qualitative Analysis.

As shown in Figure 6, PaperTalker produces presentation videos that most closely align with the human-made videos. While Veo3 [9] renders a high-quality speaker in front of the screen, it is constrained by short duration (*e.g.*, 8s) and blurred text. Besides, PresentAgent[23] typically suffers from missing of the presenter and slide-design errors (*e.g.*, incorrect titles, incomplete author lists).

5.4 Key Ablations

Is Talking Head Necessary? In presentation videos, the presence of a presenter substantially affects audience engagement. Empirically, the VideoArena results in Table 2 show

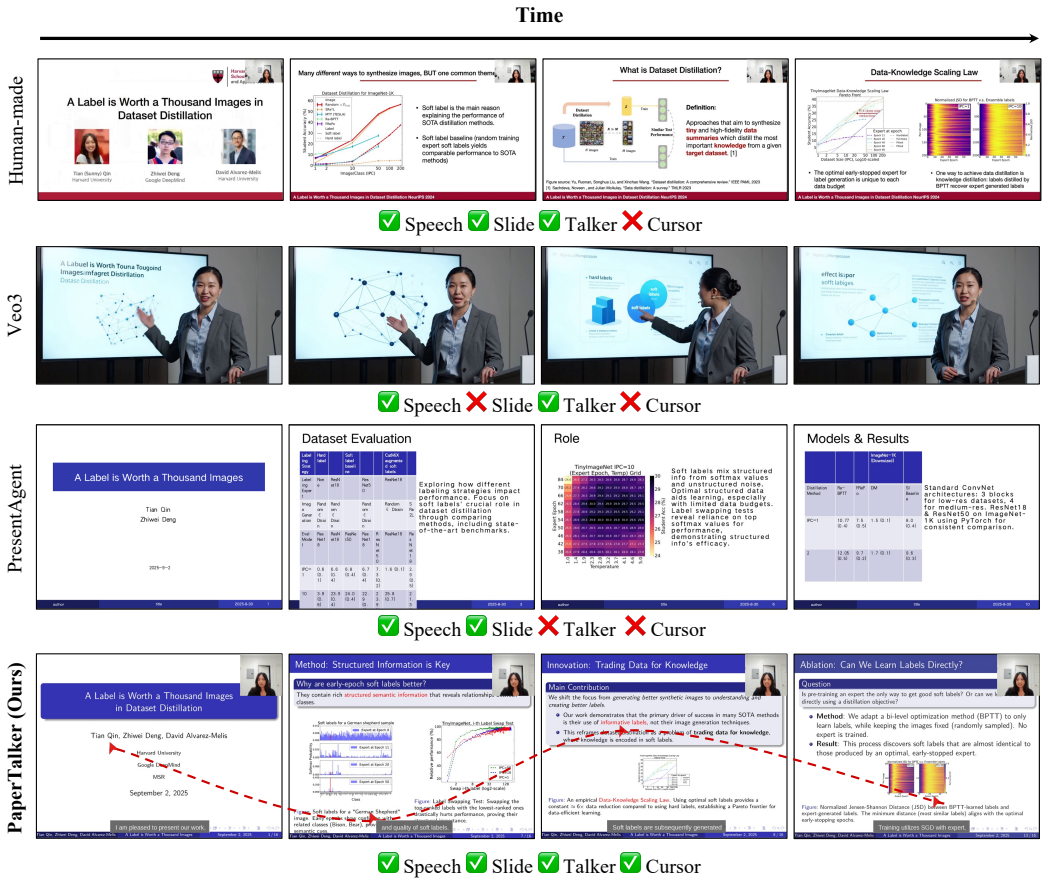


Figure 6: **Visualization of generated results.** PaperTalker produces presentation videos with rich, fine-grained slide content, accurate cursor grounding, and an engaging presenter; in contrast, Veo3 [9] yields blurred text and incomplete information coverage, while PresentAgent [23] produces text-heavy slides and suffers from overfull layout issues and inaccurate information (*e.g.*, title and institutions).

PaperTalker outperforming its no-presenter, no-cursor variant PaperTalker* by more than 10%, implying that the VideoLLM prefers presentation videos which includes a presenter.

Table 4: **Ablation study on cursor.** PaperTalker^o denotes PaperTalker without cursor generation.

Method	Accuracy↑
PaperTalker^o	0.371
PaperTalker	**0.383**

What benefits are brought by Cursor Highlight? Motivated by the observation that a cursor typically helps audiences locate the relevant region, we hypothesize that a visible cursor, by providing an explicit spatial cue, facilitates content grounding for VLMs. To test this, we construct a localization QA task: for each subtitle sentence and its corresponding slide, a VLM is prompted to generate a four-option multiple-choice question targeting the sentence's position on the slide. We then measure answer accuracy on paired screenshots with and without the cursor overlay. As shown in Table 4, the accuracy is higher with the cursor, corroborating its importance for the audience visual grounding accessibility of presentation videos.

6 Conclusions

This work tackles the long-standing bottleneck of presentation video generation by agent automation. With Paper2Video, we provide the first comprehensive benchmark and well-designed metrics to rigorously evaluate presentation videos in terms of quality, knowledge coverage, and academic memorability. Our proposed PaperTalker framework demonstrates that automated generation of ready-to-use academic presentation videos is both feasible and effective, producing outputs that closely approximate author-recorded presentations while significantly reducing production time by 4 times. We hope our work advances AI for Research and supports scalable scholarly communication.

References

[1] Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. Whisperx: Time-accurate speech transcription of long-form audio. *arXiv preprint arXiv:2303.00747*, 2023.

[2] Eric Chamoun, Michael Schlichtkrull, and Andreas Vlachos. Automated focused feedback generation for scientific writing assistance. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9742–9763, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.580. URL https://aclanthology.org/2024.findings-acl.580/.

[3] Qiguang Chen, Mingda Yang, Libo Qin, Jinhao Liu, Zheng Yan, Jiannan Guan, Dengyun Peng, Yiyan Ji, Hanjing Li, Mengkang Hu, et al. Ai4research: A survey of artificial intelligence for scientific research. *arXiv preprint arXiv:2507.01903*, 2025.

[4] Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Jian Zhao, Kai Yu, and Xie Chen. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. *arXiv preprint arXiv:2410.06885*, 2024.

[5] Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Jian Zhao, Kai Yu, and Xie Chen. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. *arXiv preprint arXiv:2410.06885*, 2024.

[6] Ziru Chen, Shijie Chen, Yuting Ning, Qianheng Zhang, Boshi Wang, Botao Yu, Yifei Li, Zeyi Liao, Chen Wei, Zitong Lu, et al. Scienceagentbench: Toward rigorous assessment of language agents for data-driven scientific discovery. *arXiv preprint arXiv:2410.05080*, 2024.

[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.

[8] Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. A dataset of information-seeking questions and answers anchored in research papers. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4599–4610, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.365. URL https://aclanthology.org/2021.naacl-main.365/.

[9] DeepMind. Veo 3 technical report. Technical report, DeepMind, May 2025. URL https://storage.googleapis.com/deepmind-media/veo/Veo-3-Tech-Report.pdf. Technical Report.

[10] Jay DeYoung, Iz Beltagy, Madeleine van Zuylen, Bailey Kuehl, and Lucy Lu Wang. Ms2: Multi-document summarization of medical studies. *arXiv preprint arXiv:2104.06486*, 2021.

[11] Tomas Goldsack, Zhihao Zhang, Chenghua Lin, and Carolina Scarton. Making science simple: Corpora for the lay summarisation of scientific literature. *arXiv preprint arXiv:2210.09932*, 2022.

[12] Linmei Hu, Duokang Wang, Yiming Pan, Jifan Yu, Yingxia Shao, Chong Feng, and Liqiang Nie. Novachart: A large-scale dataset towards chart understanding and generation of multimodal large language models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 3917–3925, 2024.

[13] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024.

[14] Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, et al. Vbench++: Comprehensive and versatile benchmark suite for video generative models. *arXiv preprint arXiv:2411.13503*, 2024.

[15] Uri Katz, Mosh Levy, and Yoav Goldberg. Knowledge navigator: Llm-guided browsing framework for exploratory search in scientific literature. *arXiv preprint arXiv:2408.15836*, 2024.

[16] Kevin Qinghong Lin, Linjie Li, Difei Gao, Zhengyuan Yang, Shiwei Wu, Zechen Bai, Stan Weixian Lei, Lijuan Wang, and Mike Zheng Shou. Showui: One vision-language-action model for gui visual agent. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19498–19508, 2025.

[17] Yao Lu, Yue Dong, and Laurent Charlin. Multi-XScience: A large-scale dataset for extreme multi-document summarization of scientific articles. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8068–8074, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.648. URL https://aclanthology.org/2020.emnlp-main.648/.

[18] Yue Ma, Kunyu Feng, Zhongyuan Hu, Xinyu Wang, Yucheng Wang, Mingzhe Zheng, Xuanhua He, Chenyang Zhu, Hongyu Liu, Yingqing He, et al. Controllable video generation: A survey. *arXiv preprint arXiv:2507.16869*, 2025.

[19] Ludovico Mitchener, Jon M Laurent, Benjamin Tenmann, Siddharth Narayanan, Geemi P Wellawatte, Andrew White, Lorenzo Sani, and Samuel G Rodriques. Bixbench: a comprehensive benchmark for llm-based agents in computational biology. *arXiv preprint arXiv:2503.00096*, 2025.

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

[21] Yujia Qin, Yining Ye, Junjie Fang, Haoming Wang, Shihao Liang, Shizuo Tian, Junda Zhang, Jiahao Li, Yunxin Li, Shijue Huang, et al. Ui-tars: Pioneering automated gui interaction with native agents. *arXiv preprint arXiv:2501.12326*, 2025.

[22] Mirco Ravanelli, Titouan Parcollet, Adel Moumen, Sylvain de Langen, Cem Subakan, Peter Plantinga, Yingzhi Wang, Pooneh Mousavi, Luca Della Libera, Artem Ploujnikov, Francesco Paissan, Davide Borra, Salah Zaiem, Zeyu Zhao, Shucong Zhang, Georgios Karakasidis, Sung-Lin Yeh, Pierre Champion, Aku Rouhe, Rudolf Braun, Florian Mai, Juan Zuluaga-Gomez, Seyed Mahed Mousavi, Andreas Nautsch, Ha Nguyen, Xuechen Liu, Sangeet Sagar, Jarod Duret, Salima Mdhaffar, Gaëlle Laperrière, Mickael Rouvier, Renato De Mori, and Yannick Estève. Open-source conversational ai with speechbrain 1.0. *Journal of Machine Learning Research*, 25(333), 2024. URL http://jmlr.org/papers/v25/24-0991.html.

[23] Jingwei Shi, Zeyu Zhang, Biao Wu, Yanjie Liang, Meng Fang, Ling Chen, and Yang Zhao. Presentagent: Multimodal agent for presentation video generation. *arXiv preprint arXiv:2507.04036*, 2025.

[24] Parshin Shojaee, Ngoc-Hieu Nguyen, Kazem Meidani, Amir Barati Farimani, Khoa D Doan, and Chandan K Reddy. Llm-srbench: A new benchmark for scientific equation discovery with large language models. *arXiv preprint arXiv:2504.10415*, 2025.

[25] Giulio Starace, Oliver Jaffe, Dane Sherburn, James Aung, Jun Shern Chan, Leon Maksin, Rachel Dias, Evan Mays, Benjamin Kinsella, Wyatt Thompson, et al. Paperbench: Evaluating ai's ability to replicate ai research. *arXiv preprint arXiv:2504.01848*, 2025.

[26] Edward Sun, Yufang Hou, Dakuo Wang, Yunfeng Zhang, and Nancy XR Wang. D2s: Document-to-slide generation via query-based text summarization. *arXiv preprint arXiv:2105.03664*, 2021.

[27] Shuai Tan, Bin Ji, and Ye Pan. Flowvqtalker: High-quality emotional talking face generation through normalizing flow and quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26317–26327, 2024.

[28] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

[29] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

[30] Mengchao Wang, Qiang Wang, Fan Jiang, Yaqi Fan, Yunpeng Zhang, Yonggang Qi, Kun Zhao, and Mu Xu. Fantasytalking: Realistic talking portrait generation via coherent motion synthesis. *arXiv preprint arXiv:2504.04842*, 2025.

[31] Ronghuan Wu, Wanchao Su, Kede Ma, and Jing Liao. Iconshop: Text-guided vector icon synthesis with autoregressive transformers. *ACM Trans. Graph.*, 42(6), December 2023. ISSN 0730-0301. doi: 10.1145/3618364. URL https://doi.org/10.1145/3618364.

[32] Weijia Wu, Mingyu Liu, Zeyu Zhu, Xi Xia, Haoen Feng, Wen Wang, Kevin Qinghong Lin, Chunhua Shen, and Mike Zheng Shou. Moviebench: A hierarchical movie level dataset for long video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 28984–28994, 2025.

[33] Weijia Wu, Zeyu Zhu, and Mike Zheng Shou. Automated movie generation via multi-agent cot planning. *arXiv preprint arXiv:2503.07314*, 2025.

[34] Yanzheng Xiang, Hanqi Yan, Shuyin Ouyang, Lin Gui, and Yulan He. Scireplicate-bench: Benchmarking llms in agent-driven algorithmic reproduction from research papers. *arXiv preprint arXiv:2504.00255*, 2025.

[35] David Junhao Zhang, Jay Zhangjie Wu, Jia-Wei Liu, Rui Zhao, Lingmin Ran, Yuchao Gu, Difei Gao, and Mike Zheng Shou. Show-1: Marrying pixel and latent diffusion models for text-to-video generation. *International Journal of Computer Vision*, pages 1–15, 2024.

[36] Hao Zheng, Xinyan Guan, Hao Kong, Jia Zheng, Weixiang Zhou, 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.

Appendix

Contents

A PaperTalker

The details of repair and visual-selective MCTS layout refinement are shown in Algorithm 1 and Algorithm 2 respectively.

Algorithm 1 Slide Generation via Compile–Repair

Require: Text source t; figure set $\mathcal{F} = \{f_i\}_{i=1}^{n}$; max iterations K
Ensure: Compilable slide code c^*

$c \leftarrow \mathrm{LLM}(t, \mathcal{F})$ ▷ initial draft $c^{(0)}$
for $k = 1$ **to** K **do**
  $(e, w) \leftarrow \mathrm{Compile}(c)$
  if $e = \varnothing$ **then**
    $c^* \leftarrow c$; **return** c^* ▷ no errors; stop
  end if
  $c \leftarrow \mathrm{LLM}(c, e)$ ▷ repair using compiler errors only
end for
$c^* \leftarrow c$; **return** c^* ▷ best-effort if max iterations reached

Algorithm 2 Visual-Selective Layout Refinement

Require: Compilable slide code c^*; text font scales $\mathcal{A}_{\text{text}}$; figure font scale α_{fig}; figure scales $\mathcal{B}$
Ensure: Refined slide code $c^\dagger$ and slides s

$(\_, w) \leftarrow \mathrm{Compile}(c^*)$; $\mathcal{I} \leftarrow \mathrm{OverfullFrames}(w)$
if $\mathcal{I} = \varnothing$ **then**
  return $c^\dagger \leftarrow c^*$
end if
for each $i \in \mathcal{I}$ **do**
  $s \leftarrow \mathrm{ExtractFrame}(c^*, i)$; $\mathcal{C} \leftarrow \varnothing$
  if $\mathrm{IsTextOnly}(s)$ **then**
    $\mathcal{C} \leftarrow \{\mathrm{ApplyFontScale}(s, \alpha) : \alpha \in \mathcal{A}_{\text{text}}\}$
  else
    $s_{\text{font}} \leftarrow \mathrm{ApplyFontScale}(s, \alpha_{\text{fig}})$
    $\mathcal{C} \leftarrow \{s_{\text{font}}\} \cup \{\mathrm{ApplyFigureScale}(s_{\text{font}}, \beta) : \beta \in \mathcal{B}\}$
  end if
  $\mathrm{img} \leftarrow \{\mathrm{RenderAsImage}(c \text{ with } s' \text{ at } i) : s' \in \mathcal{C}\}$
  $s^\star \leftarrow \arg\max_{s' \in \mathcal{C}} \mathrm{VLM\_Judge}(\mathrm{img}[s'], \mathrm{prompt}_{\text{layout}})$
  $c \leftarrow \mathrm{ReplaceFrame}(c, i, s^\star)$
end for
$c^\dagger \leftarrow c$
return $s \leftarrow \mathrm{CompilePDF}(c^\dagger)$

B Prompts

Prompt: Slide Generation

System Prompt: Please generate a complete English PPT introduction based on the following TeX source text content, using LaTeX Beamer. The specific requirements are as follows.

Content structure:

- The PPT should contain the following chapters (arranged in order), and each chapter must have a clear title and content:
- Motivation (research background and problem statement)
- Related work (current status and challenges in the field)
- Method (core technical framework) [The content of the method needs to be introduced in detail, and each part of the method should be introduced on a separate page]
- Innovation (differentiation from existing work)
- Experimental method (experimental design and process)
- Experimental setting (dataset, parameters, environment, etc.)

- Experimental results (main experimental results and comparative analysis)
- Ablation experiment (validation of the role of key modules)
- Deficiencies (limitations of current methods)
- Future research (improvement direction or potential application)
- End slide (Thank you)

Format requirements:

- Use Beamer's theme suitable for academic presentations, with simple color matching.
- The content of each page should be concise, avoid long paragraphs, and use itemize or block environment to present points. The title page contains the paper title, author, institution, and date.
- Key terms or mathematical symbols are highlighted with alert{}.

Image and table processing:

- All image paths are given, and relative paths are used when citing, the picture names must "be consistent with the name in tex file".
- Images should automatically adapt to width, and add titles and labels
- Experimental result tables should be extracted from the source text, formatted using tabular or booktabs environments, and marked with reference sources ("as shown in table").

Code generation requirements:

- The generated LaTeX code must be complete and can be compiled directly (including necessary structures).
- Mark the source text location corresponding to each section in the code comments (for example,
- If there are mathematical formulas in the source text, they must be retained and correctly converted to LaTeX syntax (such as $y = f(x)$).

Other instructions:

- Image content should be read from the tex file, and the source name should be used directly without arbitrary modification. Image references should use real image names and should not be forged;
- Table content should first extract real data from the source document.
- All content should be in English.
- If the source text is long, it is allowed to summarize the content, but the core methods, experimental data and conclusions must be retained.
- To enhance readability, a transition page can be added (for example, "This section will introduce the experimental part").
- Perfer more images than heavy text. **The number of slides should be around 10.**
- **& in title is not allowed which will cause error "Misplaced alignment tab character &̈** **Pay attention to this "error: !File ended while scanning use of framë**
- Only output latex code which should be ready to compile using tectonic(simple verson of TeX Live). Before output check if the code is grammatically correct.

Prompt: Error Correction

System Prompt: You are given a LaTeX Beamer code for the slides of a research paper and its error information. Correct these errors *without changing* the slide content (text, figures, layout).

Instructions:

- Apply the minimal edits required to make the file compile: add missing packages, close/open environments, balance braces, escape special characters, fix math delimiters, resolve duplicate labels, and correct obvious path or option typos.
- Do *not* paraphrase or delete text; do *not* change figure/table content, captions, labels, or layout semantics.
- Keep all image/table file names and relative paths as given; do not invent or rename assets.
- Preserve the original Beamer theme, colors, and structure.
- Ensure the final output compiles with **Tectonic**; close all environments and avoid undefined commands.

Output (strict): Output *only* the corrected LaTeX source, beginning with `beamer` and ending with `document`; no extra commentary.

Prompt: MSTS Judge

System Prompt: You are a slide layout judge. You see four slides A–D in a 2×2 grid: A (top-left), B (top-right), C (bottom-left), D (bottom-right).

Definitions

- **Overfull:** any part of the figure or its caption is clipped, outside the frame, or overlapped/hidden.
- **Coverage:** among non-overfull options, larger visible content with less empty background is better.
- **Risk:** risk of overfull decreases from A → D (A largest, D smallest).
- **Coverage trend:** coverage decreases from A → D.

Rules (judge only the given images)

1. Disqualify any option with overfull (caption must be fully visible).
2. From the remaining, pick the one with the greatest coverage.
3. Practical method: scan **A** → **B** → **C** → **D**; choose the *first* slide in that order that is not overfull.

Output only (strict; do *not* output ‘‘‘`json`):

```
{
"reason": "concise comparison",
"choice": "A" | "B" | "C" | "D"
}
```

Prompt: Slide Script with Cursor Positions

System Prompt: You are an academic researcher presenting your own work at a research conference. You are provided with a sequence of adjacent slides.

Instructions:

- For each slide, write a smooth, engaging, and coherent first-person presentation script.
- Clearly explain the *current* slide with academic clarity, brevity, and completeness; use a professional, formal tone and avoid content unrelated to the paper.

- Each sentence must include *exactly one* cursor position description drawn from the *current slide* and listed in order, using the format `script | cursor description`. If no cursor is needed for a sentence, write `no`.
- Limit the total script for each slide to **50 words** or fewer.
- Separate slides using the delimiter `###`.

Output Format (strict):

```
sentence 1 | cursor description
sentence 2 | cursor description
...
###
sentence 1 | cursor description
...
```

Prompt: Meta Similarity

System Prompt: You are an evaluator. You will be given two presentation videos of the same talk: (1) a human-presented version and (2) an AI-generated version. Evaluate *only* the slides and subtitles; ignore the presenter's face, voice quality, background music, camera motion, and any non-slide visuals.

Inputs You May Receive

- Human video (and optionally its slide images and subtitles/transcript)
- AI video (and optionally its slide images and subtitles/transcript)

Evaluation Scope (focus strictly on slides + subtitles)

1. **Slide Content Matching:** Do AI slides convey the same key points and comparable layout/visual elements (titles, bullets, diagrams, tables, axes annotations) as the human version?
2. **Slide Sequence Alignment:** Is slide order consistent? Any sections missing, added, or rearranged?
3. **Subtitle Wording Similarity:** Do AI subtitles reflect similar phrasing/terminology and information as the human speech/subtitles? Focus on semantic equivalence; minor style/spelling differences do not matter.
4. **Slide–Subtitle Synchronization:** Within the AI video, does narration/subtitle content match the on-screen slide at the same time? Does this broadly align with the human presenter's per-slide content?

Evidence-Only Rules

- Base the judgment solely on the provided materials (videos, slides, subtitles). Do *not* use outside knowledge.
- If some inputs are missing (*e.g.*, no subtitles), judge from what is available and briefly note the missing piece in the Reasons.

Relaxed Scoring Rubric (0–5)

- **5** — Nearly identical: slides and subtitles closely match the human version in content, layout, sequence, and timing; wording is near-paraphrase.
- **4** — Highly similar: only minor layout/phrasing differences; content, order, and alignment clearly match.
- **3** — Moderate differences yet same core content: several layout/wording/sequence deviations but main sections and key points are preserved. (Leniency: borderline cases between 2 and 3 *round up* to 3.)
- **2** — Partial overlap: substantial omissions/rearrangements or subtitle drift; multiple slide mismatches or sync issues.

- **1** — Minimal overlap: only a few matching fragments; most slides/subtitles diverge.
- **0** — No meaningful match: AI slides/subtitles do not correspond to the human version.

*Lenient mapping: if borderline between adjacent levels, choose the higher score. If computing subscores, average and **round up** to the nearest integer in [0,5].*

Output Format (STRICT; exactly one line)

`Content Similarity: X/5; Reasons`

Where `X` is an integer 0–5 from the rubric, and `Reasons` is 1–3 short sentences referencing content, sequence, wording, and synchronization as relevant.

Prompt: PresentArena

System Prompt: You are an expert in evaluating academic presentation videos. You are given two videos (Video A and Video B) on the same research topic. Evaluate each video independently and then decide which is better, or if they are basically the same (preferred when not confident).

Evaluation Criteria

- **Content Clarity**: Are key ideas and findings clearly explained?
- **Speaker Delivery**: Is the speaker confident, fluent, and engaging?
- **Visual Aids**: Are slides/visuals clear, helpful, and well-integrated?
- **Structure & Pacing**: Is the talk logically organized and appropriately paced?
- **Audience Engagement**: Does the speaker maintain interest and attention?

Steps

1. **Step 1:** Write a short (1–2 sentence) evaluation of **Video A** based on the criteria.
2. **Step 2:** Write a short (1–2 sentence) evaluation of **Video B** based on the criteria.
3. **Step 3:** Decide which video is better, or if they are basically the same (prefer "Same" if not confident).

Output Format (Strict; only these three blocks):

```
Step 1:
[1-2 sentences evaluating Video A]

Step 2:
[1-2 sentences evaluating Video B]

Step 3:
Final Judgment:
[A] | [B] | [Same]

Reason: [One concise sentence justifying the judgment based on Steps1-2.]
```

Prompt: PresentationQuiz

System Prompt: You are an answering agent. You will be provided with: 1) a presentation video of a paper, and 2) a JSON object called `"questions"` containing multiple questions, each with four options (A–D). Analyze the video thoroughly and answer each question *solely* based on the video content (no external knowledge). Do not reference timesteps that exceed the video length.

Instructions:

- For each question, if the video provides sufficient evidence for a specific option (A, B, C, or D), choose that option.
- Include a brief reference to where in the video the evidence appears (*e.g.*, "Top-left text", "Event date section").
- Rely only on the video; do not use outside context.
- Provide an answer entry for *all* questions present in `"questions"`.

Template (steps to follow):

1. Study the presentation video together with `"questions"`.
2. For each question, determine whether the video clearly supports one of the four options; if so, pick that answer.
3. Provide a brief reference indicating where in the video you found the evidence.
4. Format the final output strictly as a JSON object with the following pattern (and no extra keys or explanations).

Output Format (strict):

```
{
  "Question 1": { "answer": "X", "reference": "some reference" },
  "Question 2": { "answer": "X", "reference": "some reference" },
  ...
}
```

questions payload:

```
{{questions}}
```