MPRESENTER: Multilingual Agentic System for Paper-to-Video Generation

Wenhan Han[1], Xiao Xiao[2], Mykola Pechenizkiy[1], Meng Fang[2,1]
[1]Eindhoven University of Technology
[2]University of Liverpool
w.han@tue.nl, xiao.xiao@liverpool.ac.uk
m.pechenizkiy@tue.nl, Meng.Fang@liverpool.ac.uk

Abstract

Generating presentation videos from academic papers is challenging due to the need for long-document discourse planning and cross-lingual grounding. Existing Paper2Video systems are largely monolingual and often rely on single-pass pipelines, which can limit the coherence and informativeness of the resulting presentations. We present MPRESENTER, a multilingual agentic Paper2Video system that decomposes the task into planning, audience-oriented critique, layout-aware slide generation, and multilingual figure interpretation, enabling iterative refinement at the discourse level. To facilitate reproducible evaluation, we also introduce MPREBENCH, a multilingual benchmark that evaluates presentation videos via question answering as a proxy for effective information transfer. Experimental results indicate that MPRESENTER improves question-answering accuracy relative to prior systems, while maintaining affordable cost and latency. We release both the system and benchmark to support further research on multilingual Paper2Video [1].

1 Introduction

The volume of scientific publications continues to grow, making it increasingly difficult for researchers to efficiently disseminate and consume new findings. In this context, presentation-style videos, such as conference talks and video abstracts, have become an important medium for scientific communication, as they combine structured visual framing with spoken explanations that help audiences grasp dense technical content more effectively than text alone. Empirical studies further suggest that video abstracts can increase visibility and scientific impact, leading to higher views and citation counts (Bonnevie et al., 2023; Zong et al., 2019). Despite these benefits, producing a high-quality presentation video from a paper remains labor-intensive, requiring careful decisions about content selection, slide design, narrative structure, and alignment between visuals and narration.

Automating the generation of presentation videos from academic papers poses substantial research challenges, particularly in terms of language understanding, discourse planning, and multilingual grounding. Although recent advances in Large Language Models (LLMs) and autonomous agents have enabled progress in complex tasks such as web navigation (Zhou et al., 2024), and social simulation (Park et al., 2023), existing approaches to Paper2Video remain limited. Prior work has explored document-to-slide and document-to-poster generation (Fu et al., 2022; Pang et al., 2025; Zheng et al., 2025), as well as end-to-end agentic pipelines for paper-to-video generation (Zhu et al., 2025; Shi et al., 2025; Liu et al., 2025). However, Paper2Video is not a straightforward composition of these components: a practical system must preserve the paper's technical narrative, select and explain figures and tables, and generate slides whose layout remains readable under real presentation constraints.

In addition, existing systems are predominantly monolingual and often produce outputs of limited fidelity, restricting real-world applicability. These limitations are further amplified in multilingual settings, where long-document discourse planning, cross-lingual grounding, and figure interpretation become critical challenges. Although English dominates scholarly publishing, a substantial body of research is produced in other languages, such as Chinese, the second-largest source of academic publications. Many researchers do not speak English fluently (Kleidermacher and Zou, 2025), and scientific figures frequently contain language-specific text that cannot be handled by translation alone. Meanwhile, long-context summarization remains difficult for current models (Li et al., 2024), often leading to omissions or superficial coverage.

[1] https://github.com/aialt/mpresenter

Findings of the Association for Computational Linguistics: ACL 2026, pages 16358–16371

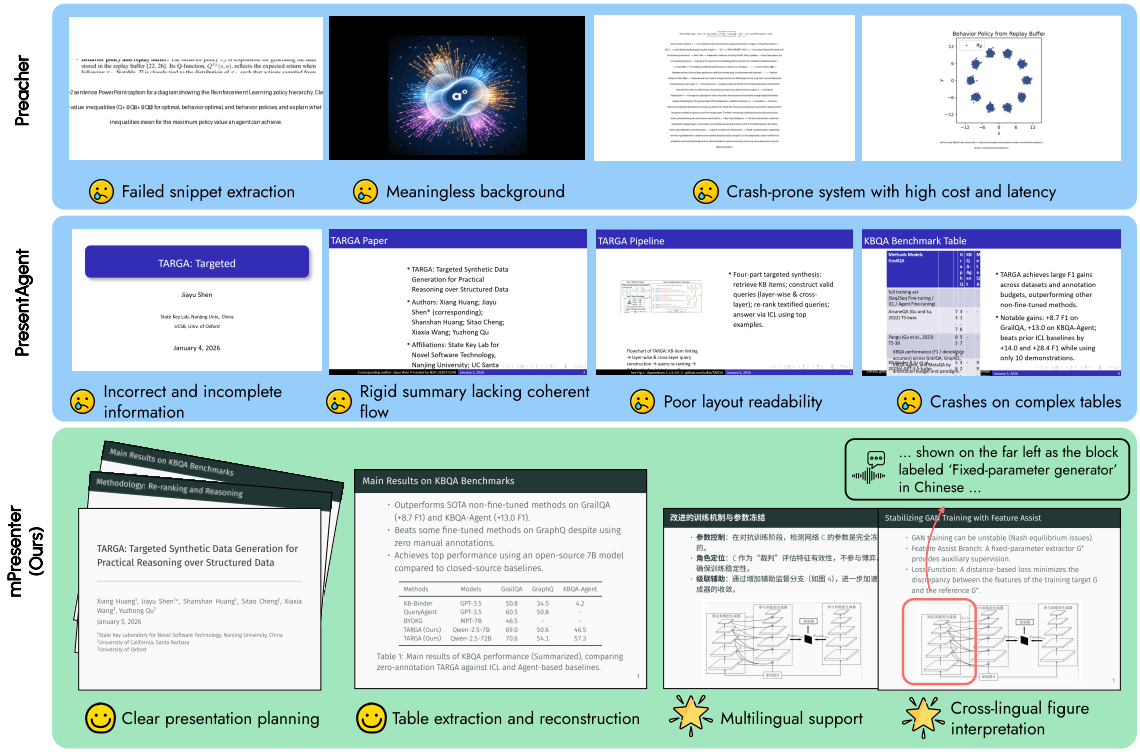


Figure 1: Paper2Video system comparison. mPresenter delivers a coherent presentation flow, strong content extraction and layout management, robust multilingual support, and an interpreter module that helps explain images in non target languages.

Together, these challenges highlight the need for methods that explicitly address discourse structure, multilingual reasoning, and iterative refinement in Paper2Video systems.

In this paper, we present MPRESENTER, a multilingual agentic system that converts an academic paper into a presentation video, focusing on effective information transfer in realistic presentations. MPRESENTER coordinates four interacting agents with complementary roles. The *planner* proposes a slide outline that balances coverage and pacing; the *reviewer* acts as an audience member who questions unclear points and requests improvements; the *coder* turns the outline into executable Beamer slides, iteratively compiling and inspecting rendered results to refine layout and visual readability; and the *interpreter* specializes in multilingual figure comprehension. This division of labor enables iterative refinement that is difficult to achieve with a single-pass generation pipeline. Figure 1 shows the comparison between MPRESENTER and previous agentic systems.

To support reproducible and scalable evaluation of Paper2Video systems, particularly in multilingual settings, we also introduce MPREBENCH, a benchmark for multilingual Paper2Video. MPREBENCH contains 40 English and Chinese papers. Each paper is paired with eight expert-written questions spanning diverse aspects of the work, and every question is provided in 5 languages. We evaluate generated presentation videos via audience QA accuracy utilizing a powerful vision-language model (VLM) to watch the video and answers the questions. This protocol provides an objective, content-grounded metric that directly reflects whether the video successfully conveys the paper's key information, while avoiding reliance on reference video.

We compare MPRESENTER against recent works. Results show that MPRESENTER improves Paper2Video quality from toy-level demos toward practical usability, achieving higher question-answering accuracy while substantially reducing cost and latency. We open-source the system and benchmark to support reproducible multilingual Paper2Video research.

To summarize, our contributions include:

- We propose MPRESENTER, the first multilingual Paper2Video agentic system with dedicated roles for planning, audience-driven cri-

tique, layout-aware slide coding, and multilingual figure interpretation.

- We release MPREBENCH, an expert-curated, high-difficulty multilingual benchmark for evaluating the quality of generated presentation videos, focusing on conveying key information and aligning evaluation with real-world use.
- We provide empirical evidence that our approach improves accuracy and usability with low cost and latency.

2 Related Work

Document-to-slides and visual summaries. A growing line of work studies converting long-form documents into structured visual artifacts, especially slide decks and posters. Early efforts such as Doc2PPT formulate slide creation as a layout-aware summarization problem grounded in scientific documents (Fu et al., 2022). More recent systems leverage stronger LLMs and iterative refinement to generate higher-quality slides beyond text-to-slides, often with planning, rendering, and evaluation components (Zheng et al., 2025). In parallel, poster automation aims to condense papers into single-page multimodal overviews with explicit layout planning and visual feedback (Pang et al., 2025). While these methods improve static visual communication, they typically stop short of modeling *spoken delivery* and the temporal structure that makes presentations effective.

From papers to narrated videos. Beyond static outputs, recent work has begun targeting narrated presentation videos. PresentAgent (Shi et al., 2025), based on PPTAgent (Zheng et al., 2025) proposes a modular pipeline that segments documents, renders slide-like frames, generates oral-style narration, and composes time-aligned videos. However, this work is limited to chaining components together and inherits PPTAgent's rigid pipeline of summarizing the paper strictly section by section, overlooking effective information delivery. PaperTalker (Zhu et al., 2025) enriches the media format by adding a talking-head presenter and cursor guidance, but it requires LaTeX source files and largely follows a fixed, stage-wise pipeline with limited content refinement. In contrast, MPRESENTER targets raw PDFs and uses interactive reviewer and render-feedback loops to improve information transfer and slide readability. Preacher (Liu et al., 2025) introduces reflection agents for content planning, but it remains largely confined to summarizing the paper section by section. Moreover, existing systems largely assume monolingual settings and do not explicitly address cross-lingual barriers that arise when papers, figures, and audiences span languages.

Agentic systems for tool-augmented multimodal generation. Our work is also connected to LLM/VLM agents that combine reasoning with action and external tools. ReAct-style frameworks emphasize interleaving deliberation and tool use for complex tasks (Yao et al., 2023), and Toolformer-like approaches highlight the value of learning to invoke tools reliably (Schick et al., 2023). For long-form video synthesis, multi-agent planning has been explored in domains such as movie generation (Wu et al., 2025), and agentic frameworks have been proposed for orchestrating generation and editing with adaptive decomposition and model selection (Yuan et al., 2024). MPRESENTER builds on this trajectory but centers on *presentation-realistic* outputs (slides + audio) and introduces an explicit interpreter role to mitigate multilingual figure-understanding gaps.

Evaluation of information delivery. Evaluating whether a generated presentation video truly delivers the paper's key information is challenging, and existing Paper2Video systems largely rely on proxy signals rather than a unified static benchmark. Preacher (Liu et al., 2025) reports rubric-based subjective scores from LLM judges and humans. PaperTalker (Zhu et al., 2025) evaluates videos with quiz-style comprehension based on LLM-generated multiple-choice questions, and also uses LLM-as-judge and human for subjective preference scoring. PresentAgent (Shi et al., 2025) combines a small set of fixed multiple-choice questions with LLM-as-judge style ratings on dimensions such as clarity and coherence. These works lack a comparable static benchmark, and do not guarantee evaluation quality that aligns with real-world use. Moreover, their evaluations are conducted only in English, limiting their applicability to multilingual presentation videos.

3 MPRESENTER

Prior Paper2Video systems generally operate within a monolingual paradigm and frequently fail to generate presentation-realistic outputs that faith-

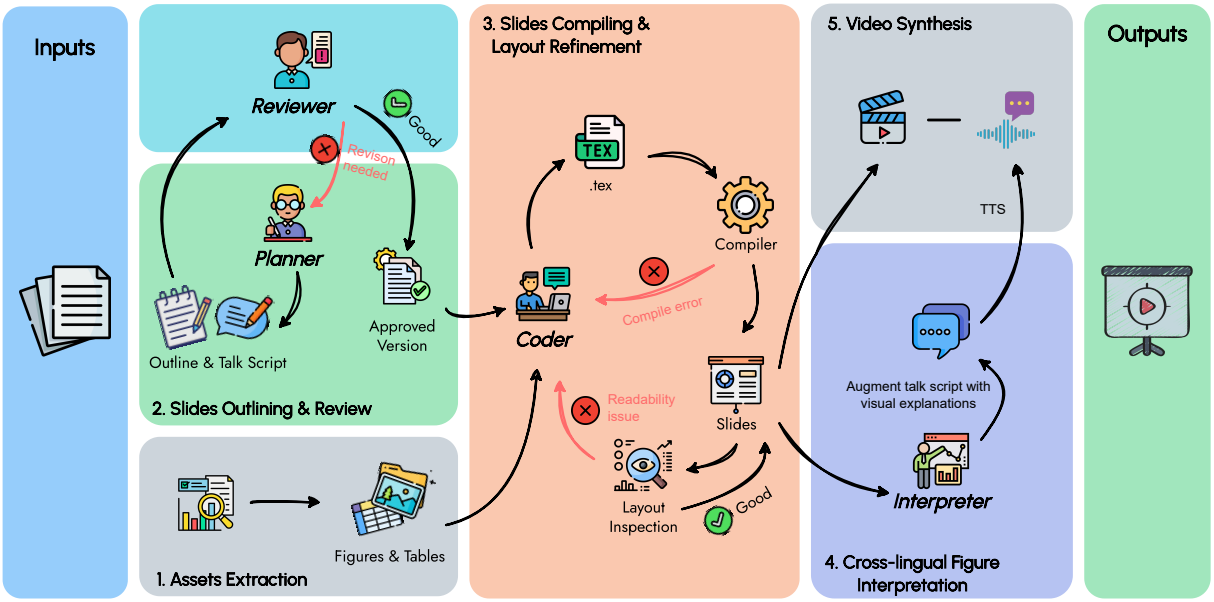


Figure 2: Overview of the MPRESENTER framework. The system orchestrates four specialized agents to transform a source paper PDF into a multilingual presentation video through iterative refinement.

fully convey technical depth, particularly in cross-lingual and long-context scenarios. To bridge this gap, we introduce MPRESENTER, a multi-agent framework designed to transform a static document into a dynamic, audio-visual presentation aimed at maximizing information transfer efficiency. Figure 2 depicts the structure of MPRESENTER system.

MPRESENTER orchestrates a set of four interacting agents: a **Planner** (A_{plan}) for structural outlining, a **Reviewer** (A_{rev}) for audience-centric critique, a **Coder** (A_{code}) for LaTeX-based slide construction and layout refinement, and an **Interpreter** (A_{int}) for resolving cross-lingual semantic barriers. This agentic workflow facilitates the iterative optimization of both content fidelity and visual presentation, aligning the output with professional presentation standards.

3.1 Task Formulation

We formalize the **Paper2Video** task as a generation mapping function. Given an academic paper PDF, denoted as $\mathcal{P}$, written in a source language ℓ_{in}, and a target presentation language ℓ_{out}, the system objective is to synthesize a presentation video $\mathcal{V}$ such that:

$$\mathcal{V} = f(\mathcal{P}, \ell_{\text{out}}) = (\mathcal{S}, \mathcal{A}) \tag{1}$$

where $\mathcal{S} = \{s_1, s_2, \ldots, s_N\}$ represents a sequence of visual slides, and $\mathcal{A} = \{a_1, a_2, \ldots, a_N\}$ represents the corresponding synchronized narration audio segments. Crucially, we address the multilingual setting where $\ell_{\text{in}} \neq \ell_{\text{out}}$, requiring the generated narration $\mathcal{A}$ to bridge the linguistic gap, particularly when visual elements in $\mathcal{S}$ contain embedded text in ℓ_{in}.

3.2 Assets Extraction

Unlike prior approaches such as PaperTalker (Zhu et al., 2025), which rely on structured LaTeX source files, MPRESENTER processes raw PDFs to maximize applicability. While modern Multimodal Large Language Models (MLLMs) possess PDF reading capabilities, the generation of evidence-based slides requires precise extraction of visual artifacts. Rather than converting PDFs to Markdown and attempting asset re-linking via retrieval like PresentAgent (Shi et al., 2025), we employ document layout analysis.

We define the extracted asset set as $\mathcal{E} = \{(v_i, c_i, id_i)\}_{i=1}^{K}$, where v_i represents the image file of a figure or table, c_i is its associated caption, and id_i is a unique stable identifier. By explicitly referencing assets via id_i, the subsequent outlining phase achieves robust grounding, significantly reducing computational overhead by eliminating redundant MLLM calls for asset matching.

3.3 Iterative Slides Outlining

Given the paper $\mathcal{P}$ and extracted assets $\mathcal{E}$, the planner agent A_{plan} synthesizes a structured outline $\mathcal{O}$. For each slide s_k, $\mathcal{O}_k$ encompasses the key technical points, referenced asset identifiers, and a draft narration script.

To mitigate the limitations of one-shot generation, which often suffers from omission of detail or poor pacing, we introduce a reviewer agent A_{rev}. The interaction is modeled as an iterative refinement loop:

$$\mathcal{O}^{(t+1)} \leftarrow A_{\text{plan}}(\mathcal{O}^{(t)}, A_{\text{rev}}(\mathcal{O}^{(t)})) \tag{2}$$

At each step t, A_{rev} audits the outline for metadata completeness, logical flow, and clarity from the perspective of a domain researcher. It poses clarification questions or requests re-ordering. A_{plan} then generates a revised outline $\mathcal{O}^{(t+1)}$ with justifications. This process terminates when A_{rev} signals approval, ensuring the narrative is both coherent and rigorous.

3.4 Slides Compiling and Layout Refinement

The coder agent A_{code} is responsible for translating the approved outline $\mathcal{O}$ into executable Beamer LaTeX code, denoted as $\mathcal{C}$. We implement a *code-compile-render-evaluate* feedback loop to ensure visual quality.

Initially, A_{code} generates code $\mathcal{C}$. Upon successful compilation, the system renders the slide images $\mathcal{I} = \text{Render}(\mathcal{C})$. These images are fed back to A_{code} to detect visual anomalies such as text overflow, misalignment, or excessive density. Unlike rule-based systems that tweak scalar parameters, A_{code} is empowered to perform structural refactoring, such as splitting a dense slide s_k into sub-slides $\{s_{k,1}, s_{k,2}\}$, while preserving the narrative arc.

To ensure scalability, this process is parallelized at the slide level.

3.5 Cross-lingual Interpretation

A distinct challenge in the setting $\ell_{\text{in}} \neq \ell_{\text{out}}$ is the presence of source-language text or conventions within figures that are opaque to the target audience. We address this via the interpreter agent A_{int}.

For every slide s_k containing visual assets, A_{int} analyzes the rendered image alongside the initial narration script T_k. The agent determines if the visual requires exegesis. If an explanatory gap is detected, A_{int} augments the script with grounded spatial and semantic cues (e.g., explicit references to color coding, legends, or specific module locations). This process maps the initial script T_k to a refined, context-aware script T'_k:

$$T'_k = \begin{cases} A_{\text{int}}(s_k, T_k) & \text{if interpretation needed} \\ T_k & \text{otherwise} \end{cases} \tag{3}$$

This ensures comprehension without requiring modification of the original visual evidence. The samples are demonstrated in Appendix A.

3.6 Video Generation

In the final synthesis phase, the system converts the finalized text scripts into speech. We utilize a Text-to-Speech (TTS) model to generate audio waveforms, taking the interpreted scripts T'_k explicitly as input:

$$a_k = \text{TTS}(T'_k) \tag{4}$$

We then align the visual sequence $\mathcal{S}$ with the generated audio sequence $\mathcal{A} = \{a_1, \ldots, a_N\}$ to produce the final video $\mathcal{V}$:

$$\mathcal{V} = \bigoplus_{k=1}^{N} \text{Align}(s_k, a_k) \tag{5}$$

where $\oplus$ denotes the temporal concatenation of aligned audiovisual segments.

3.7 Design Rationale

MPRESENTER removes fragile steps that are not directly tied to improving presentation content, which improves system stability. We also avoid additional video styles such as talking head overlays or general background video synthesis, since they provide limited benefit for communicating key technical ideas and can distract from the core message. By allocating agent interactions to content planning, critique, layout refinement, and cross lingual interpretability, MPRESENTER improves information delivery while maintaining low cost and latency.

4 MPREBENCH

A key barrier to progress in Paper2Video is the absence of a comparable static benchmark that tests whether a generated presentation includes the paper content that *should* be presented. We therefore build MPREBENCH, an expert curated multilingual benchmark designed to evaluate *content selection quality* of Paper2Video systems. Importantly, MPREBENCH targets the presenter side: questions are intended to be answerable from reading the paper without requiring complex multi-step reasoning, so that failures can be attributed primarily to missing or miscommunicated key information in the generated presentation.

4.1 Benchmark Overview

MPREBENCH is a test only benchmark built from 40 research papers, including 20 Chinese papers

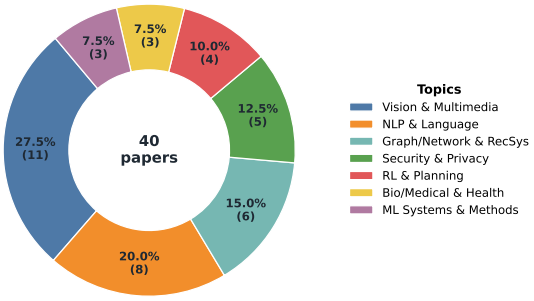


Figure 3: Topic distribution.

published in *Chinese Journal of Computers* and 20 English papers (10 from NeurIPS 2025 and 10 from ACL 2025). Figure 3 shows the topic distribution. For each paper, we provide 8 high difficulty multiple choice questions, resulting in 320 questions in total. Each question is a 4 choice single answer problem. Samples are presented in Appendix B.

4.2 Question Curating

Questions are written to reflect the core information a practical presentation video should convey, with a fixed coverage structure: Motivation, Method, Experiment, and Conclusion, with two questions per category. Annotators are instructed to focus on key details and central contributions, rather than generic background knowledge.

Concretely, Motivation questions target the research context and motivation and explicitly contrast with related work. Method questions probe the key procedure, core computation or modeling mechanism, or the correct interpretation of the paper's main figures. Experiment questions focus on experimental setup, or the correct reading of results tables and comparative conclusions. Conclusion questions test the primary takeaways supported by the experiments and analysis.

Each question contains four options with a single correct answer. Distractors must be plausible, avoid overly absolute wording, and avoid vague phrasing. To prevent shallow cues, annotators control option phrasing and length: the four options should have similar length and style.

4.3 LLM-based Validation

After writing, each question is validated by an LLM (GPT-5) that reads the full paper and answers the question. A question is accepted only if GPT-5 answers it correctly. This validation operationalizes our design goal: the questions should correspond to information that is important for a presentation and should be readily recoverable from the paper itself, without requiring complicated reasoning. The resulting benchmark therefore emphasizes whether a Paper2Video system selects and communicates the right content, instead of measuring audience-level problem solving.

4.4 Multilingual Extension

For each paper, questions are authored by an expert in the paper language. After the source questions are finalized, all questions and options are translated by human translators into five languages: English, Chinese, German, Japanese, and Arabic. Following translation, we again use GPT-5 to answer the translated questions, and we observe that GPT-5 can still answer them correctly. This provides an additional check that translation preserves semantics and the identity of the correct option.

Overall, MPREBENCH contains 1,600 multilingual question instances in total, enabling evaluation of presentation videos generated in different target languages even when the source paper is written in Chinese or English.

5 Experiments

5.1 Baselines

We compare MPRESENTER with two recent Paper2Video systems. **PresentAgent** (Shi et al., 2025) is a multimodal agent pipeline that generates presentation videos by segmenting paper content into slide-like frames and producing aligned narration. **Preacher** (Liu et al., 2025) orchestrates multiple generative components to produce narrated videos, emphasizing end-to-end automation with LLM based planning and judging.

5.2 Evaluation

Automatic evaluation. Since existing systems are not multilingual, we first compare MPRESENTER with PresentAgent and Preacher on English presentation generation. The generation is conducted under a zero-shot setting. To reduce the chance that the evaluator answers from parametric knowledge rather than the video content, we use **Gemini-2.0-Flash**, which has an early knowledge cutoff, to watch each generated presentation video and answers questions. We report three metrics. **Accuracy (ACC)** is the fraction of correctly answered questions. **Effective information density**

(EID) is defined as ACC$/T$, where T is the video duration in seconds. **Success rate (SR)** is the fraction of papers for which the system produces a valid video (i.e., no pipeline crash and the output video is successfully generated), reported over the 20 English papers.

Human evaluation. We also conduct a human evaluation. We sample five papers from the subset where both baselines successfully produce videos, and recruit three computer-science PhD students as evaluators. Each evaluator watches videos for five distinct papers and answers the corresponding questions. Evaluators see outputs from all three systems, but each paper is evaluated by at most one system for a given evaluator. Human-evaluator accuracy (**HEACC**). Neither model-based nor human-based evaluation is perfect. Human judgements can have higher variance due to attention and fatigue fluctuations when watching videos, while an LLM judge provides a more consistent and strictly controlled scoring procedure, which improves fairness across systems.

5.3 Implementation

We use **Gemini-3-Flash**[2] as the backbone of all agents in MPRESENTER. PDF assets are extracted with **PP-DocLayout-L**[3] and **PaddleOCR**[4]. Speech synthesis is performed with MegaTTS3 using `time_step=32`. We cap the number of refinement iterations: planning runs for up to 3 rounds, slide coding and layout review run for up to 6 rounds. Outline items are processed in parallel with up to 10 workers. For fairness, we also run Preacher and PresentAgent with **Gemini-3-Flash**. Although the planner and reviewer share the same backbone, they operate under different contexts and constraints: the reviewer does not see the source paper and only critiques the outline and narration plan, which encourages it to surface missing content and unclear logic rather than echo the planner. We observed that the Math Style and General Style options in Preacher often produce chaotic and uninformative visuals while incurring expensive API calls, so we restrict its video style to Slides and Static. Preacher and PresentAgent do not run stably on complex academic PDFs. We report results from versions that we repaired to the best of our ability.

[2] https://blog.google/products/gemini/gemini-3-flash

[3] https://huggingface.co/PaddlePaddle/PP-DocLayout-L

[4] https://github.com/PaddlePaddle/PaddleOCR

Method	ACC	EID	SR	HEACC
Preacher	11.1	0.07	9/20	6.9
PresentAgent	43.0	0.17	16/20	21.1
MPRESENTER	**73.8**	**0.22**	**20/20**	**53.8**

Table 1: English Paper2Video results on MPREBENCH. ACC denotes question-answering accuracy (%), EID denotes effective information density (ACC per second), SR denotes success rate (valid videos), and HEACC denotes human-evaluator accuracy (%).

5.4 Results

From the results shown in Table 1, we have findings follows.

MPRESENTER substantially outperforms prior Paper2Video agentic systems. Its generated videos enable the evaluator to reach 73.8% accuracy, a large gain over PresentAgent (41.9%). Qualitative inspection is consistent with these numbers: MPRESENTER produces clearer slide layouts and better-structured narration, making the video easier to follow and understand.

We find that Preacher nearly collapses (SR 9/20) on this setting and often fails to produce a valid presentation. Its outputs frequently contain long stretches of low-information background visuals and show limited ability to extract and organize paper content into a coherent talk. In addition, its rigid pipeline leads to poor robustness on complex academic PDFs, with frequent failures across intermediate agent steps. PresentAgent is more stable and can generate videos, but its design largely resembles a linear composition of models with limited adaptive decision making; it often reduces the paper to coarse slicing and summarization, which limits practical usefulness.

Model-based ACC is higher than human-evaluator accuracy. This gap has several plausible causes. First, humans may miss fine-grained details due to limited attention and memory, whereas LLM judges can systematically search for evidence across the video. Second, human evaluators have narrower and more variable domain knowledge; for unfamiliar topics, plausible distractors can be more misleading to humans than to LLMs. Finally, although we use a judge model with an earlier knowledge cutoff, we cannot fully rule out residual parametric knowledge about some papers, which may help the model infer the intended answer beyond what is explicitly conveyed in the presentation.

	EN Pre.		ZH Pre.	
	En Q.	ML Q.	Zh Q.	ML Q.
EN Paper	**73.8**	68.1	71.9	65.6
ZH Paper	66.9	67.1	**75.6**	66.8

Table 2: Multilingual results. ML Q. denotes the average of ACC in all languages excluding the presentation language.

5.5 Multilingual Performance of mPresenter

Evaluation across all languages is costly. We therefore evaluate mPresenter in Chinese and English by generating two videos per paper for all 40 papers in mPreBench, one video in Chinese and one in English. Each video is evaluated using questions in all languages.

Table 2 shows the multilingual result. It can be seen that generating presentations in the source language yields the highest quality for both Chinese and English papers. Cross lingual generation leads to a drop, and the degradation from Chinese papers to English videos is larger than from English papers to Chinese videos. Overall, mPresenter maintains strong performance under cross lingual settings.

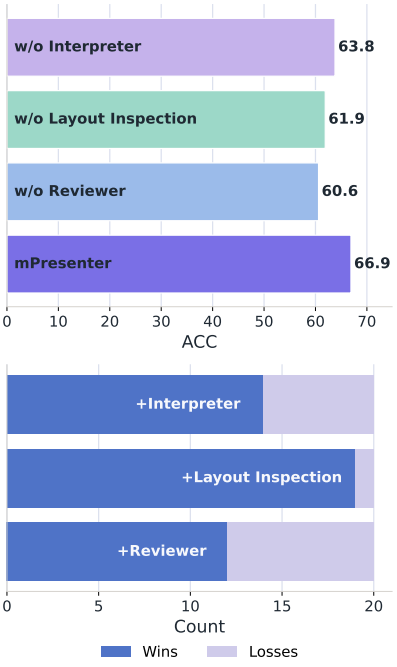


Figure 4: Ablation results. Each component improves presentation quality, with the reviewer contributing the largest gain and the full system being consistently preferred by human evaluators.

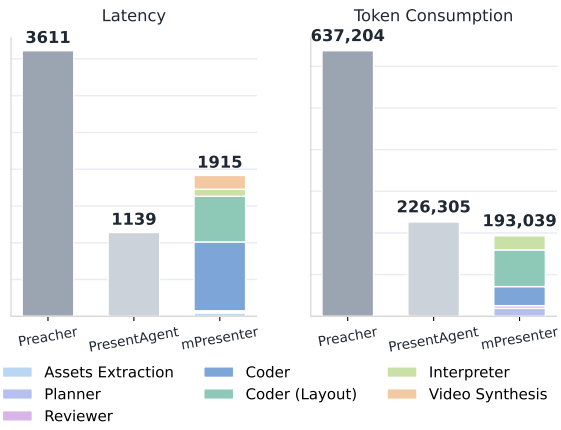


Figure 5: Latency and token consumption per video. mPresenter achieves the strongest quality efficiency tradeoff, with lower token consumption than prior systems and substantially lower latency than Preacher.

5.6 Ablation Study

To quantify the contribution of key components, we conduct ablations on the 20 Chinese papers by generating English presentation videos while removing one component at a time: (1) **w/o reviewer** (2) **w/o layout inspection** and (3) **w/o interpreter**. We evaluate each variant with the same question-answering protocol and report the change in ACC.

We also run a human preference study. We recruit three annotators and conduct pairwise comparisons between the full system and each ablation variant. Annotators select the better video based on overall viewing quality without a fixed rubric. We count a win when at least two of the three annotators prefer the same video.

In Figure 4 (top), removing individual components reduces ACC, with the largest drop observed when removing the reviewer. In contrast, removing the interpreter has a smaller effect on ACC, likely because the evaluator model is multilingual and therefore less sensitive to missing cross-lingual figure explanations. By comparison, the human preference study in Figure 4 (bottom) suggests a different emphasis: humans are more sensitive to slide readability and narration quality when judging overall presentation quality.

5.7 Efficiency Analysis

Figure 5 shows the average end-to-end time and token consumption per video. Preacher relies heavily on multimodal generation, resulting in much higher cost and latency. mPresenter achieves the best video quality with the lowest token con-

sumption. Within MPRESENTER, coding and layout inspection account for most of the runtime and token usage. The token counts in Figure 5 include only backbone LLM usage. Preprocessing such as PDF layout analysis and OCR runs locally and contributes only a small fraction of the end-to-end latency, corresponding to the Assets Extraction stage.

6 Conclusion

We presented MPRESENTER, a multilingual agentic system that converts academic paper PDFs into presentation videos, with an explicit focus on effective information transfer. We also introduced MPREBENCH, an manually curated multilingual benchmark that targets content presentation quality multiple-choice questions. Experiments show that MPRESENTER substantially improves video quality over prior Paper2Video systems while maintaining lowest cost.

Limitations

Language coverage. Although MPREBENCH provides questions in five languages and MPRESENTER supports multilingual generation, our system development and most experiments focus on a limited set of source and target languages, and results may not fully transfer to languages with different scripts or presentation conventions.

Evaluation scope. Our benchmark emphasizes content delivery and omits other aspects of presentation quality such as speaking style, prosody, engagement, and aesthetics. Question-answering accuracy is an application-aligned proxy, but it does not capture every dimension of a good talk.

Judge bias. Model-based evaluation can be affected by the judge model's capabilities and possible residual parametric knowledge. While we mitigate this by using an earlier-cutoff model, we cannot completely eliminate such effects. Human evaluation is included but remains limited in scale.

Reliance on external tools. MPRESENTER depends on PDF layout detection, OCR, LaTeX compilation, and TTS components. Failures or domain shifts in any component can degrade end-to-end quality, and computational requirements may vary across hardware and deployment settings.

References

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

Tsu-Jui Fu, William Yang Wang, Daniel McDuff, and Yale Song. 2022. DOC2PPT: Automatic Presentation Slides Generation from Scientific Documents. *Preprint*, arXiv:2101.11796.

Hannah Calzi Kleidermacher and James Zou. 2025. Science Across Languages: Assessing LLM Multilingual Translation of Scientific Papers. *Preprint*, arXiv:2502.17882.

Jiaqi Li, Mengmeng Wang, Zilong Zheng, and Muhan Zhang. 2024. LooGLE: Can Long-Context Language Models Understand Long Contexts? *Preprint*, arXiv:2311.04939.

Jingwei Liu, Ling Yang, Hao Luo, Fan Wang, Hongyan Li, and Mengdi Wang. 2025. Preacher: Paper-to-Video Agentic System. *Preprint*, arXiv:2508.09632.

Wei Pang, Kevin Qinghong Lin, Xiangru Jian, Xi He, and Philip Torr. 2025. Paper2Poster: Towards Multimodal Poster Automation from Scientific Papers. *Preprint*, arXiv:2505.21497.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. *Preprint*, arXiv:2304.03442.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language Models Can Teach Themselves to Use Tools. *Preprint*, arXiv:2302.04761.

Jingwei Shi, Zeyu Zhang, Biao Wu, Yanjie Liang, Meng Fang, Ling Chen, and Yang Zhao. 2025. PresentAgent: Multimodal Agent for Presentation Video Generation. *Preprint*, arXiv:2507.04036.

Weijia Wu, Zeyu Zhu, and Mike Zheng Shou. 2025. Automated Movie Generation via Multi-Agent CoT Planning. *Preprint*, arXiv:2503.07314.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing Reasoning and Acting in Language Models. *Preprint*, arXiv:2210.03629.

Zhengqing Yuan, Yixin Liu, Yihan Cao, Weixiang Sun, Haolong Jia, Ruoxi Chen, Zhaoxu Li, Bin Lin, Li Yuan, Lifang He, Chi Wang, Yanfang Ye, and Lichao Sun. 2024. Mora: Enabling Generalist Video Generation via A Multi-Agent Framework. *Preprint*, arXiv:2403.13248.

Hao Zheng, Xinyan Guan, Hao Kong, Jia Zheng, Weixiang Zhou, Hongyu Lin, Yaojie Lu, Ben He, Xianpei Han, and Le Sun. 2025. PPTAgent: Generating and Evaluating Presentations Beyond Text-to-Slides. *Preprint*, arXiv:2501.03936.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. 2024. WebArena: A Realistic Web Environment for Building Autonomous Agents. *Preprint*, arXiv:2307.13854.

Zeyu Zhu, Kevin Qinghong Lin, and Mike Zheng Shou. 2025. Paper2Video: Automatic Video Generation from Scientific Papers. *Preprint*, arXiv:2510.05096.

Qianjin Zong, Yafen Xie, Rongchan Tuo, Jingshi Huang, and Yang Yang. 2019. The impact of video abstract on citation counts: Evidence from a retrospective cohort study of New Journal of Physics. *Scientometrics*, 119(3):1715–1727.

A Cross-lingual Figure Interpretation

Table 3 presents how Interpreter modifies the speech script. By interpreting the figures, it adds details such as spatial cues or color references to help the audience understand visual elements that are not in the target language.

B MPREBENCH samples

Table 4 and Table 5 demonstrates the samples from MPREBENCH. Each question follows the minimal-context principle. The prompt is phrased in the most generalized way possible, avoiding any specific keywords that could hint at the correct answer.

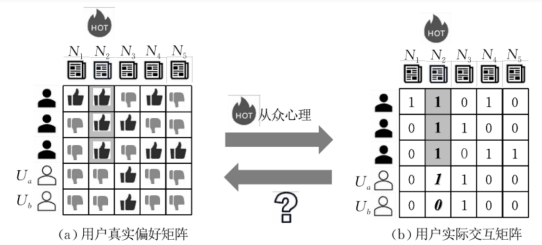
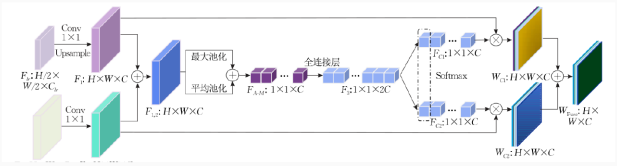
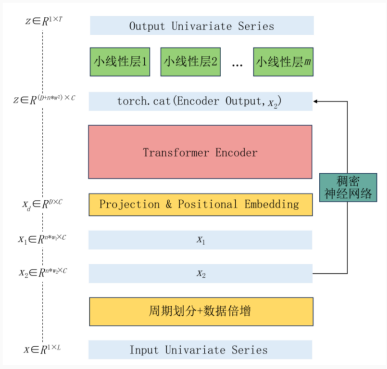
Slide	Speech
Motivation: Conformity Bias in Recommendations • Click behavior does not always equal genuine preference. • Users often click popular news due to the "Bandwagon Effect" or conformity. • Traditional debiasing oversimplifies the complex cognitive factors behind clicks.  Figure 1: Conformity bias in news recommendation. 1	**Original.** In the world of news recommendation, a click doesn't always mean a user truly likes the content. As you can see in Figure 1, news N2 is highly popular. User A might click it simply because of its popularity—a phenomenon known as the bandwagon effect or conformity—whereas User B, who has lower conformity, might ignore it. **Modified.** This distinction is illustrated in Figure 1 by comparing the true preference matrix on the left, labeled (a) with the actual interaction matrix on the right, labeled (b). As you can see, news N2 is highly popular, indicated by the 'HOT' icon. User A might click it simply because of its popularity—a phenomenon known as the bandwagon effect or conformity.
SE-MSFF: Multi-Scale Feature Fusion • **Multi-Scale Extraction:** Concatenates features from three different layers to capture local and global context. • **Global Recalibration:** SE weights are applied globally to the fused feature map, not individual scales. • **Effective Representation:** Helps in identifying targets that vary from single pixels to small streaks.  Figure 8: Squeeze-Excitation Multi-Scale Feature Fusion module. 6	**Original.** To address scale variance, we use the SE-MSFF module. We first concatenate feature maps from multiple scales. Then, we apply a global squeeze-excitation mechanism to the entire fused map. This recalibrates the channel-wise importance across all scales simultaneously. **Modified.** We first concatenate feature maps from multiple scales into the large blue block on the left. Then, we apply a global squeeze-excitation mechanism to the entire fused map. You can see this process in the center of the diagram, where it passes through max and average pooling.
Proposed Method: CycleLLH Architecture  Figure 3: Structure of the CycleLLH model 3	**Original.** This architecture, shown in Figure 3, allows us to use a standard Transformer encoder while drastically reducing the total parameter count. By structuring the model this way, we can focus on both inter-cycle and intra-cycle temporal dependencies efficiently. **Modified.** This architecture, shown in Figure 3, allows us to use a standard Transformer encoder—the central pink block—while drastically reducing the total parameter count. By structuring the model with the cycle partitioning shown in the lower yellow block.

Table 3: The Interpreter edits speech samples.

Type	Question (EN)	Options (EN)
Motivation	Which limitation is identified in the paper?	A) Temporal instability B) Absence of metric scaling C) Emphasis on semantic categorization D) Sensitivity to illumination variations
Method	Which combination of data sources is reported in the paper?	A) Virtual environments, physical models, and clinical records B) Tissue samples, volumetric scans, and procedural videos C) Stereoscopic systems, surface scans, and alternative modalities D) Generative models, simulation, and expert reports
Experiment	Which metric is used to report the final estimation error?	A) Mean absolute error B) Mean intersection over union C) Root mean square error D) Average precision
Conclusion	What is reported about the method's performance relative to human assessment?	A) The method outperformed practitioners B) The method matched practitioner accuracy C) The method was outperformed by practitioners D) Both the method and practitioners showed high error rates

Table 4: Question type samples from MPREBENCH.

Q	Language	Question	Options
Q1	English	What element is identified as missing from existing datasets?	A) Endoscope trajectory information B) Scalable image sequences C) Real-world volumetric models D) Target instances with verified physical dimensions
Q1	Chinese	现有数据集中确定缺失了什么元素?	A) 内窥镜轨迹信息 B) 可扩展的图像序列 C) 真实世界的体积模型 D) 具有经过验证的物理尺寸的目标实例
Q1	Japanese	既存のデータセットに欠けていると特定された要素は何か？	A) 内視鏡の軌道情報 B) スケーラブルな画像シーケンス C) 現実世界のボリュメトリックモデル D) 検証済みの物理的寸法を持つターゲットインスタンス
Q1	German	Welches Element wird als fehlend in bestehenden Datensätzen identifiziert?	A) Informationen zur Endoskop-Trajektorie B) Skalierbare Bildsequenzen C) Reale volumetrische Modelle D) Zielinstanzen mit verifizierten physischen Dimensionen
Q2	English	Which approach is used for establishing ground truth in the paper?	A) Projection of 3D reconstructions B) Calculation from estimated variables C) Application of fixed scale factors D) Calibration with a known physical reference
Q2	Chinese	文中使用哪种方法来确定真值（ground truth）？	A) 3D 重建投影 B) 通过估计变量进行计算 C) 应用固定比例因子 D) 使用已知物理参照进行标定
Q2	Japanese	この論文でグラウンドトゥルースを確立するために使用されているアプローチはどれか？	A) 3D再構成の投影 B) 推定変数からの計算 C) 固定スケール因子の適用 D) 既知の物理的参照によるキャリブレーション
Q2	German	Welcher Ansatz wird in der Arbeit zur Bestimmung der Ground Truth verwendet?	A) Projektion von 3D-Rekonstruktionen B) Berechnung aus geschätzten Variablen C) Anwendung fester Skalierungsfaktoren D) Kalibrierung mit einer bekannten physischen Referenz

Table 5: Multilingual Sample (Arabic not displayed).