SlideTailor: Personalized Presentation Slide Generation for Scientific Papers

Wenzheng Zeng*, Mingyu Ouyang*, Langyuan Cui*, Hwee Tou Ng†

Department of Computer Science, National University of Singapore
{wenzhengzeng, ouyangmingyu04, langyuan.c}@u.nus.edu, dcsnght@nus.edu.sg

Abstract

Automatic presentation slide generation can greatly streamline content creation. However, since preferences of each user may vary, existing under-specified formulations often lead to suboptimal results that fail to align with individual user needs. We introduce a novel task that conditions paper-to-slides generation on user-specified preferences. We propose a human behavior-inspired agentic framework, SlideTailor, that progressively generates editable slides in a user-aligned manner. Instead of requiring users to write their preferences in detailed textual form, our system only asks for a paper-slides example pair and a visual template—natural and easy-to-provide artifacts that implicitly encode rich user preferences across content and visual style. Despite the implicit and unlabeled nature of these inputs, our framework effectively distills and generalizes the preferences to guide customized slide generation. We also introduce a novel chain-of-speech mechanism to align slide content with planned oral narration. Such a design significantly enhances the quality of generated slides and enables downstream applications like video presentations. To support this new task, we construct a benchmark dataset that captures diverse user preferences, with carefully designed interpretable metrics for robust evaluation. Extensive experiments demonstrate the effectiveness of our framework.

Project website — https://github.com/nusnlp/SlideTailor

1 Introduction

Presentations, usually delivered through slides, are a widely used medium for effectively communicating information in a visually engaging and accessible way (Bartsch and Cobern 2003). Crafting high-quality presentations, however, demands considerable effort, requiring the author to put in informative and focused content, craft a coherent and compelling narrative, and create an appealing visual design. Given the time and expertise required, there is a growing interest in developing automated systems that can generate presentation slides to reduce the manual workload involved.

Recent works (Fu et al. 2022; Zheng et al. 2025; Xu et al. 2025) have exploited the inherent multimodal nature of the academic document-to-slides generation task, showing promising results in both content quality and visual layout. Despite the effectiveness, they typically treat slide generation as a straightforward document-to-slides conversion task. This overlooks a crucial aspect: the user. We argue that presentation design is inherently subjective. Users have different preferences in terms of narrative structure, emphasis, conciseness, and aesthetic choices. Consequently, under-specified or one-size-fits-all generation frameworks often yield outputs that do not align well with individual needs. To enable more personalized, user-aligned presentations, it is essential to incorporate individual preferences.

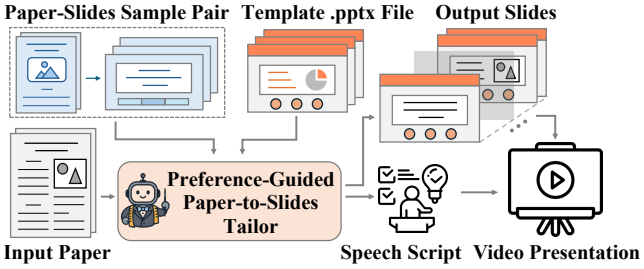

Figure 1: Preference-guided paper-to-slides generation. Based on user preferences inferred from a paper-slides sample pair and a visual template, the system produces personalized slides accompanied by a speech script, supporting downstream applications such as video presentations.

Motivated by this, we shed light on this under-explored research problem: preference-guided paper-to-slides generation, where the generation process is explicitly guided by user-specified preferences. We categorize such preferences into two main aspects: (1) Content preferences, which affect narrative flow and the level of emphasis or conciseness applied to specific topics or sections; and (2) Aesthetic preferences, which govern layout structure, background design, decorative elements, and overall stylistic choices.

Instead of asking users to articulate their preferences in detailed textual instructions, we propose a more user-friendly way. The system takes (1) a paper-slides sample pair, which implicitly encodes the user's content structuring-related preferences, and (2) a `.pptx` slide template, which reflects aesthetic choices. These inputs are natural to pro-

*These authors contributed equally.
†Corresponding author.

vide and align with how users often prepare slides—by referencing prior slide decks and reusing templates. Besides, these two types of inputs are relatively orthogonal, offering greater clarity and flexibility by enabling customized slide generation along separate content and aesthetic dimensions. However, while these inputs are convenient for users to provide, they pose nontrivial challenges for the slide generation model: the embedded preferences are implicit, entangled, and unlabeled, making them difficult for the model to extract and apply effectively.

To tackle this challenge, we propose a human behavior-inspired agentic framework termed SlideTailor. It progressively constructs editable slides aligned with user preferences. The process begins with preference distillation, similar to a human that summarizes and learns multi-aspect user preferences from both the given sample pair and `.pptx` template file, forming an internalized preference profile. Guided by this profile, the system performs a preference-guided and presentation-oriented summarization process. It extracts and reorganizes salient content from the input paper, while adjusting the level of detail, emphasis, and narrative flow to align with the user's preferred presentation style. The resulting content is then structured into a coherent outline across slides, specifying the intended message and supporting points for each slide.

As a novel component of our framework, we introduce a chain-of-speech mechanism during the outline construction. Inspired by how human presenters plan their speech alongside slide design, this mechanism prompts the system to simulate narrative when outlining each slide. As a result, slide content can better align with the anticipated speech, improving coherence and clarity. It also enables downstream applications such as full video presentations (Fig. 1).

Based on the constructed outline, the system proceeds to template planning, selecting the most appropriate layout for each slide based on its semantic content and intended visual emphasis. This step ensures that the slide structure and aesthetics are jointly optimized in line with user preferences. Finally, the system generates slides by editing the selected templates and exporting them in standard `.pptx` format, which enables flexible user refinement and downstream use.

To facilitate research on this new task, we construct a benchmark dataset that captures and simulates diverse user preferences to comprehensively evaluate customized paper-to-slides generation methods. We also carefully craft interpretable metrics for robust evaluation of preference alignment and presentation quality. Experimental results demonstrate that our method not only better aligns with user intent, but also produces slides with higher overall quality compared to existing approaches.

2 Related Work

2.1 Document-to-Slides Generation

Prior works primarily considered slide generation as a text summarization problem (Li et al. 2021; Sun et al. 2021; Costa, Amaro, and Gonçalo Oliveira 2023; Maheshwari et al. 2024; Cachola et al. 2024), overlooking layout design and the inherent multimodal nature of engaging presentations. Some works (Xu and Wan 2021; Fu et al. 2022; Bandyopadhyay et al. 2024), especially recent and contemporaneous studies (Zheng et al. 2025; Pang et al. 2025; Shi et al. 2025; Zhu, Lin, and Shou 2025), began integrating visual and layout elements for multimodal presentation. Despite their effectiveness, most methods treat slide generation as a direct document-to-slides process, lacking constraints that capture diverse user preferences. Persona-Aware-D2S (Mondal et al. 2024b) considers customization, but restricts preferences to four fixed categories. PPTAgent (Zheng et al. 2025) enables flexible template input, but focuses solely on layout aspects, overlooking content-related preferences. In contrast, we explore the subjective nature of slide generation in a more realistic formulation, with essential contributions on definition, methodology, and dataset, aiming to facilitate future research on personalized slide generation.

2.2 Conditional Summarization

This task generates summaries conditioned on auxiliary inputs beyond the source itself, such as queries (Yu and Han 2022; Xu et al. 2023b; Cao et al. 2024), topic cues (Li et al. 2021; Mukherjee et al. 2022), timeline (Hu, Moon, and Ng 2024; Qorib, Hu, and Ng 2025), diagram (Mondal et al. 2024a), and user preferences (Xu et al. 2023a). The subarea most relevant to our work concerns user preferences. Within this subarea, different studies focus on different facets—for example, review summarization with personalized recommendations (Li, Li, and Zong 2019; Cheng et al. 2023; Xu et al. 2023a; Ghodratnama and Zakershahrak 2024) and controllable abstractive summarization where users specify attributes like style or length (Fan, Grangier, and Auli 2018). In contrast to these tasks, we focus on the conditional summarization of scientific papers for presentation slide generation. The work most similar to ours is (Mondal et al. 2024b). However, it is limited to four predefined preferences (i.e., expert/non-expert, long/short), which fail to capture the diverse and fine-grained nature of real user needs. Our formulation instead models diverse and fine-grained user preferences in a more realistic, open-ended setting.

3 Task Formulation

We now formulate the preference-guided paper-to-slides generation task. We begin by defining the key components that constitute user preferences in the context of presentation slide generation. We distinguish two preference dimensions: (1) Content preferences, which shape how a paper is mapped into presentation—affecting the narrative flow, level of detail, emphasis placed on certain sections (e.g., highlighting results over methodology), and decisions about what content to omit or elaborate; (2) Aesthetic preferences, which reflect the user's visual and stylistic inclinations—governing choices such as slide layout (e.g., text-only vs. text-image combination), background themes, color palettes, typography, and the use of decorative elements like icons or logos.

Next, we explore what forms of user input can both effectively capture expressive user preferences and align well with real-world presentation creation workflows. Rather

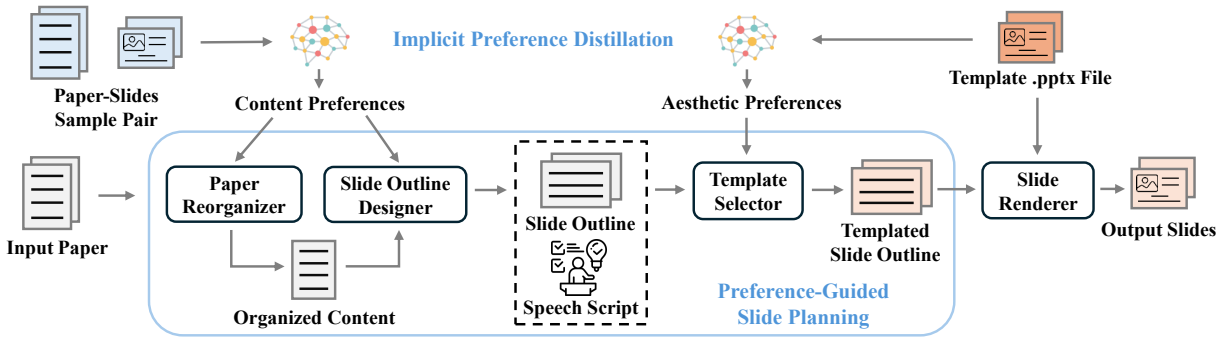

Figure 2: The conceptual pipeline of our proposed preference-guided paper-to-sildes generation framework.

than requiring users to explicitly articulate their preferences through detailed textual instructions—which can be unnatural and burdensome—we adopt a more user-friendly and example-driven setting. Specifically, the system takes two types of input: (1) a *paper-slides sample pair*, which captures the content preferences; and (2) a `.pptx` *slide template*, which conveys aesthetic preferences such as layout style and visual theme. These inputs are natural for users to provide and mirror common practices in presentation authoring, where individuals often reference previous slide decks and reuse institutional or personal templates. Moreover, the two preference types are relatively orthogonal, enabling modular flexibility—any aesthetic template can, in principle, pair with any content preference profile. In practice, we assume users have a general sense of their desired outcomes, yielding self-consistent preferences, while the designed model is expected to exhibit adaptive capability to achieve preference harmony across different preference sources during slide generation.

Formally, given a paper D in PDF format, the goal is to synthesize the corresponding slides S that (i) satisfy the general quality requirements of slide production from D, and (ii) follow the subjective preference embodied by the aforementioned two types of conditional user inputs: a paper-slides sample pair $(D_{\text{ref}}, S_{\text{ref}})$ for content preferences, and a template `.pptx` file S_{tmpl} for aesthetic preferences. This can be formally defined as:

$$S = F\left(D, (D_{\text{ref}}, S_{\text{ref}}), S_{\text{tmpl}}\right), \tag{1}$$

where F is the function that jointly performs conditional paper summarization and slide generation.

4 Method

In this section, we introduce our proposed framework SlideTailor for personalized paper-to-slides generation. We begin by analyzing the core challenges of this newly introduced task, which in turn motivate our design. We then present the framework's components.

4.1 Key Challenges

While the aforementioned setup in Sec. 3 eases the burden on users and better reproduces real-world authoring practice, it poses two key challenges:

C_1: Learning from implicit, unlabelled preference signals. Although the user-supplied sample pair $(D_{\text{ref}}, S_{\text{ref}})$ and template S_{tmpl} provide rich preference signals, they are implicit and entangled without structured labels. Thus, the system must distill diverse content and aesthetic preferences without fine-grained supervision, making preference extraction inherently ambiguous and under-constrained.

C_2: Multi-aspect alignment. Guided by the distilled preferences, slide generation should achieve harmonious alignment across content and aesthetic dimensions.

To address C_1 and C_2, we introduce a human behavior-inspired agentic framework SlideTailor that progressively models and applies user preferences in slide generation. As illustrated in Fig. 2, the overall process involves three stages: (1) implicit preference distillation from unlabeled user inputs, (2) preference-guided slide planning from the distilled profile, and (3) slide realization via appropriate template editing. This modular design mirrors slide preparation by humans: starting from internalizing personal style, then organizing content, and finally creating slides for presentation.

4.2 Implicit Preference Distillation

The first stage extracts user preferences from two types of example-driven inputs: a paper-slides sample pair $(D_{\text{ref}}, S_{\text{ref}})$, and a template `.pptx` file S_{tmpl}. These inputs are unlabeled, with no explicit annotation of what the user prefers, but still carry rich preference cues across content structure and visual design. We approach this preference inference by designing a dual-branch distillation process that aims at converting them into explicit, structured, and interpretable preferences for subsequent slide generation. Our key insight is to treat the sample pair $(D_{\text{ref}}, S_{\text{ref}})$ as an implicit demonstration of how the user transforms source content into a presentation, while treating the template S_{tmpl} as a reflection of their aesthetic inclinations. These two sources are complementary: the former governs *what* and *how* to present; the latter specifies *how it should look*.

Content preference as a latent mapping. We model the transformation $f_{\text{content}} : D_{\text{ref}} \rightarrow S_{\text{ref}}$ as a latent function that captures the user's personal preferences for abstraction and organization. Instead of aligning exact sentences or

Narrative Flow:
Title & Author Info. → Background & Motivation → …
Section-Level Preference:
1. Title & Author Info
Content Handling: Condensed
Formatting Preferences: Minimal text / full sentences
Additional Comments: Only essentials: author names, institution ...
2. Background & Motivation
Content Handling: Condensed
Formatting Preferences: Bullet Points / Short Phrases
Additional Comments: Presents general context and problem statement through … avoids technical depth …
…
Omitted Sections:
Detailed Related Work, Implementation Details, …

Figure 3: An example of distilled content preference.

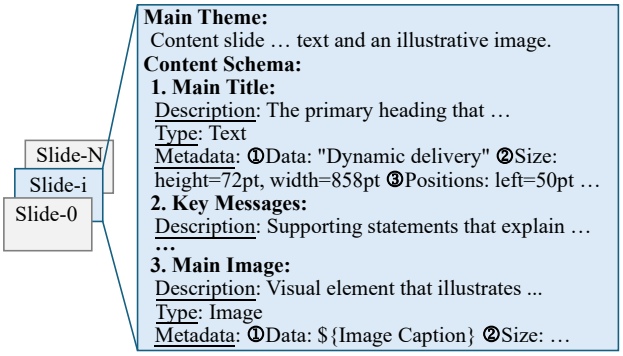

Figure 4: An example of distilled aesthetic preference.

keywords, we leverage a large language model (e.g., GPT-4.1 (OpenAI 2025b)) to infer how the content is selected, emphasized, omitted, or reordered, beyond surface form. This yields a structured content preference profile P_C comprising: (1) narrative flow (e.g., Title → Background & Motivation → ... → Future Work), and (2) section-level emphasis or omission, as well as formatting preferences (e.g., use of bullet points). The LLM is also encouraged to provide additional comments when deemed helpful, allowing the extracted structure to remain flexible and adaptive to context. An illustration example is shown in Fig. 3.

Aesthetic preference from the template. In parallel, we distill layout-related aesthetic preferences from the provided template S_{tmpl}. We employ a vision-language model (VLM) to infer the functional roles of both slide-level components (e.g., title, main content, conclusion) and element-level components (e.g., text boxes, image regions) within each slide. In addition, we incorporate fine-grained metadata directly parsed from the raw `.pptx` file, such as precise bounding box/image positions and sizes. The resulting structured schema P_A, as shown in Fig. 4, serves as a layout grounding mechanism that facilitates subsequent template understanding and selection.

Finally, the union $P = P_C \cup P_A$ constitutes a modular, symbolic representation of user preferences. Subsequent stages use P as the conditioning context, ensuring that the generated slides simultaneously reflect the sample's structural style and the template's aesthetics.

4.3 Preference-Guided Slide Planning

Guided by distilled preferences P, the system progressively plans the presentation by deciding what to say, structuring it into slides, and assigning visual formats. This process is handled by three LLM-powered agents: a paper reorganizer that restructures an input paper based on preferences, a slide outline designer that segments content into slides and drafts narration, and a template selector that assigns a suitable template for each planned slide.

Conditional paper reorganizer. Unlike generic summarization, our LLM agent reorganizes the input paper to reflect user-specific priorities encoded in P_C—adjusting the narrative flow and the level of detail based on presentation preferences. The result is a presentation-oriented document that forms the content for subsequent slide generation, as shown in the left part of Fig. 5.

Slide-wise outline generation with chain-of-speech. Next, the reorganized content is segmented into a coherent slide-wise outline, where each slide plan specifies the intended message and visual cues (e.g., images, tables from the paper). Inspired by how presenters rehearse during slide creation, we introduce a *chain-of-speech* mechanism that simultaneously drafts speech for each slide. This encourages alignment between visual content and anticipated oral delivery, improving the coherence and usability of the final output. The resulting slide outline and speech script (as shown in the middle part of Fig. 5) will jointly guide both content realization and layout selection. This design also facilitates downstream applications, as illustrated in Sec. 4.5.

Template-aware layout selection. Once the content outline is finalized, the system selects an appropriate layout for each slide by matching it with one slide from the user-provided template, based on the slide-level aesthetic schema P_A, as shown in the right part of Fig. 5. This per-slide matching aligns with real-world authoring practices and enhances coherence between slide content and visual presentation.

4.4 Slide Realization

Finally, the system realizes each slide by editing the selected template layout using the outline and corresponding speech draft. A layout-aware agent maps planned content (e.g., titles, text, visuals) to specific elements (e.g., text boxes, image placeholders) in the assigned template. This structured mapping may involve modifying, replacing, or inserting elements for coherence. A code agent then generates executable code to apply these edits directly to the `.pptx` file. This editing-based strategy preserves the original layout and theme while producing fully editable slides suitable for further user refinement.

4.5 Downstream: Video Presentations

Beyond slide generation, our framework could potentially support downstream applications such as automated, speaker-aware video presentations. Thanks to the proposed chain-of-speech mechanism, each slide is paired with a generated speech script T, which can be directly transformed into personalized narration using existing zero-shot text-to-speech systems (Qin et al. 2023; Jiang et al. 2025). With just

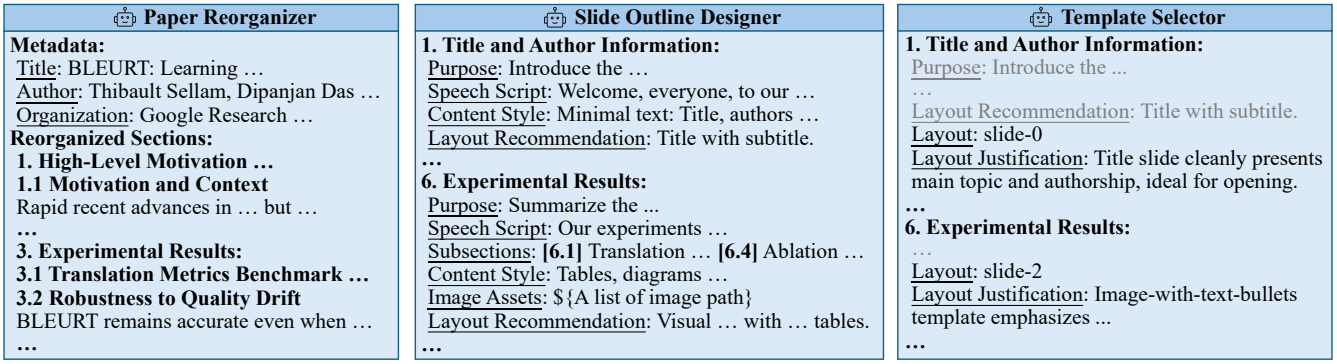

Figure 5: Example output from paper reorganizer (left), slide outline designer (middle), and template selector (right).

a short voice sample, the synthesized speech could preserve the user's vocal identity (Qin et al. 2023). Combined with the visual slides S, this could enable the automatic creation of synchronized, personalized presentation videos, offering a scalable solution for remote teaching or pre-recorded conference talks. In addition, other compelling extensions can be integrated. For example, an identity-preserving talking head can be synthesized using existing audio-driven generation methods (Zheng et al. 2024; Hong et al. 2025) and incorporated into the video to further enhance realism and audience engagement. In this work, we also take a step toward realizing this downstream extension. The implementation details can be found in our project website.

5 The PSP Dataset

To facilitate research on the proposed new task, we construct a dedicated benchmark dataset, PSP (Paper-to-Slides with Preferences), with effort on both data and evaluation protocol. Unlike prior datasets that focus on direct paper-to-slides conversion with limited consideration of user preferences, PSP explicitly incorporates diverse user preferences, covering both content structuring and aesthetic presentation, thereby paving the way for comprehensive evaluation of customized paper-to-slides generation methods.

We manually curated data from the public proceedings of leading AI and scientific venues, including top conferences such as AAAI, ACL, CVPR, NeurIPS, as well as high-impact journals like Nature, Science, The Lancet, and Chemical Reviews Letters. The source corpus encompasses papers spanning a broad spectrum of research fields, including general AI, machine learning, natural language processing, computer vision, chemistry, and medicine, ensuring both topical and stylistic diversity. To capture variation in user presentation preferences, we collected 50 distinct and high-quality paper-slides pairs that reflect diverse structuring and stylistic preferences across presenters and disciplines. Additionally, we curated a set of 10 academic slide templates representative of common research-oriented layout and aesthetic conventions. Finally, we gathered 200 scientific papers to serve as target input papers for slide generation. Altogether, this yields a pool of 200 target papers, 50 sample paper-slides pairs, and 10 layout templates, resulting in up to $200 \times 50 \times 10 = 100{,}000$ unique input combinations for conditional slide generation. As shown in Table 1, our dataset is the largest among existing paper-to-slides generation benchmarks, offering significantly more input combination possibilities and uniquely supporting open-ended preference modeling.

6 Evaluation Metrics

To support proper evaluation, we introduce two complementary sets of metrics: preference-based metrics, which assess how well the generated slides can follow user preferences, and preference-independent metrics, which evaluate the overall slide quality independent of those preferences.

Preference-Based Metrics We propose four metrics to evaluate a system's ability to follow user preferences across different aspects.

1. Coverage. It evaluates whether the generated slides cover a similar set of high-level subtopics (e.g., introduction, motivation, method) as the sample slides. We use an LLM to extract these structural topics from both S and S_{ref} and compute the intersection-over-union (IoU) between them.

2. Flow. It assesses whether the subtopics are presented in a similar order. Using the same LLM-based extraction method, we obtain the subtopic sequences from both the generated and sample slides, and compute the Normalized General Levenshtein Distance (NGLD) (Yujian and Bo 2007) between them. The similarity score is defined as $1 - \text{NGLD}(S, S_{\text{ref}})$, where NGLD lies in $[0, 1]$.

3. Content Structure. We assess how well the generated slides align with the structural presentation style of the input paper–slides sample pair. Using an LLM-as-a-judge framework, the model is instructed to focus on content organization such as pace, level of detail, visual formatting, and slide transitions—while ignoring the actual subject matter. A score from 1 to 5 is assigned based on the degree of structural and stylistic alignment.

4. Aesthetic. It evaluates how well the generated slides visually align with the given template, focusing on layout, background, color scheme, fonts, and recurring elements (e.g., headers or logos). The assessment targets visual design only, ignoring textual or semantic content. We feed slide screenshots into a VLM to produce a 1–5 score.

Benchmark	# Test Papers	# Preference Types	# Combinations	Open-Ended Preference?	Source Domain
SciDuet (Sun et al. 2021)	81	-	81	-	3 AI Conferences
DOC2PPT (Fu et al. 2022)	595	-	595	-	9 AI Conferences
Persona-Aware-D2S (Mondal et al. 2024b)	50	4	200	No	Subset of SciDuet
PSP (Ours)	200	500	100,000	Yes	7 AI Conferences, 3 Biomedical Journals, 1 Chemistry Journal, 1 General Journal

Table 1: Comparison among existing paper-to-slides generation benchmarks.

Method	Preference-based				Preference-independent		Overall
	Coverage	Flow	Content Structure	Aesthetic	Content	Aesthetic	
ChatGPT	62.62	56.84	61.60	80.80	47.00	68.32	62.86
AutoPresent (GPT-4.1)	72.84	59.58	49.60	22.40	28.05	60.20	48.78
PPTAgent (GPT-4.1)	64.41	54.24	57.60	97.20	58.36	71.96	67.30
Ours (Qwen2.5+Qwen2.5VL)	70.19	62.16	68.41	92.80	48.84	72.84	69.21
Ours (GPT-4.1)	**74.47**	**66.65**	**72.80**	**98.00**	**67.64**	**75.24**	**75.80**

Table 2: Performance on the PSP dataset. Comparison of our framework (two backbone variants) against three state-of-the-art baselines. Scores are averaged over 50 target papers.

Preference-Independent Metrics While capturing user preferences is central to our system, the generated slides should also be of high quality regardless of those preferences. To support a more holistic evaluation, we also introduce a set of metrics that assesses the presentation quality independent of user-specific preferences. Each metric is scored using an MLLM-as-a-judge framework, where the model rates specific aspects of the slides from 1 to 5 following a defined rubric.

1. Content. It evaluates how clearly and accurately the slides convey the key information of the target paper. The model considers the relevance and depth of the content and clarity. The goal is to assess whether the slides provide a coherent, focused, and informative summary of the original work.

2. Aesthetic. It assesses the overall visual appeal of the generated slides. An MLLM is instructed to focus on the presentation's design elements—such as layout, color choices, font consistency, visual balance, and spacing—without considering content semantics. The model reviews the slides holistically to determine their professional and aesthetic quality.

Note that all metrics are normalized to a 0–100 scale for consistent comparison across evaluation dimensions. We use GPT-4.1 as the MLLM for all evaluations. Detailed grading rubrics are provided at our project website. We also report the average of these metrics, denoted as "Overall".

7 Experiments

7.1 Experimental Setup

For the primary evaluation, we randomly sampled 50 papers from the PSP dataset as target input papers. Each paper was independently paired with one paper–slides sample pair and one `.pptx` template file as preferences, both randomly selected from the benchmark dataset.

We benchmarked our framework against three state-of-the-art slide generation baselines: (1) **ChatGPT** (OpenAI 2025a): We manually upload all input components (target paper, paper-slides sample pair, and template) via its web interface and prompt it to generate slides of the target paper based on the supplied preferences. (2) **AutoPresent** (Ge et al. 2025): As a text-to-slides generation method, AutoPresent takes only raw text as input. To simulate preference conditioning, we adapted it to our setting by concatenating the plain-text versions of the paper-slides pair with the target paper. (3) **PPTAgent** (Zheng et al. 2025): Since PPTAgent does not accept a paper-slides pair as preferences, we only provide the target paper and template as input. For all methods (including ours), we constrain the generation to 10 slides by embedding the instruction into the prompt for a fair comparison.

All compared systems were evaluated in a zero-shot setting powered by MLLMs. Unless otherwise noted, we employed and evaluated each system with the proprietary GPT-4.1 (checkpoint `gpt-4.1-2025-04-14`) serving as both vision and language model. We also instantiated and evaluated our system with the open-source Qwen2.5-72B-Instruct (Yang et al. 2024) and Qwen2.5-VL-72B-Instruct models (Bai et al. 2025) to demonstrate its adaptability and robustness across base LMs. These open-source models were served through the LMDeploy (LMDeploy Contributors 2023) framework and ran on NVIDIA H200 GPUs.

7.2 Experimental Results

Quantitative results. From Tab. 2, it can be observed that: (1) No method achieves an overall average score above 80%, highlighting the inherent difficulty of the preference-guided paper-to-slides generation task. (2) Our method (GPT-4.1) achieves the highest overall score (75.8%) and consistently outperforms all baselines across both preference-based and preference-independent metrics, suggesting that our framework produces slides that are not only well-aligned with user intent but also more informative and coherent from a general perspective. (3) Our approach also performs competitively when using the open-source Qwen2.5 + Qwen2.5VL models, demonstrating strong adaptability across different MLLM backbones without requiring model-specific tuning.

Setting	Preference-based				Preference-independent		Overall
	Coverage	Flow	Content Structure	Aesthetic	Content	Aesthetic	
Without content preference	65.80	56.83	54.67	94.67	65.73	73.93	68.61
Without chain-of-speech	73.60	63.99	**66.00**	94.00	47.33	**74.53**	69.91
Full system	**74.82**	**68.38**	**66.00**	**96.67**	**66.40**	73.60	**74.31**

Table 3: Ablation results on key model components. Results on a 30-sample subset of the PSP dataset.

Cost analysis. We sample five instances, each generating a 10-slide deck, with an average cost of $0.665 (GPT version) or $0.016 (Qwen version), based on OpenRouter (OpenRouter Team 2025) API pricing as of October 13, 2025.

Qualitative analysis. Due to space limitations, detailed visualizations are included at our project website. Here we summarize key observations: AutoPresent (Ge et al. 2025) fails to reflect aesthetic preferences due to its text-only input format. Although it can generate interleaved image-text output, the generated images are not faithfully derived from the paper, leading to weaker informativeness and potentially misleading content. ChatGPT supports multi-modal inputs but still struggles to consistently capture the desired visual style, and often omits figures and tables from the original paper, likely due to long context and the difficulty of extracting meaningful visuals from the paper. PPTAgent better preserves layout templates but still lacks alignment with the content structure. It also has a higher failure rate for image-related content extraction compared to our method. It frequently produces slides with large areas of unreasonable blank space and sometimes retains placeholder elements from the template that should have been removed. These observations highlight the challenge of slide generation and the effectiveness of our method.

Human evaluation. We recruited four graduate students with over two years of AI research experience to compare our method with PPTAgent, the strongest existing approach. Each participant completed 15 case studies. For each study, they were given the full input context (i.e., the target paper, the paper-slides sample pair, and the `.pptx` template), as well as anonymized outputs from both systems. They were asked to score each case on metrics mirroring those in our MLLM-based evaluation, covering two preference-based and two preference-independent metrics, as well as selecting an overall preferred output. The scoring rubric and instructions were identical to those used in the automatic evaluation. In total, we collected 60 independent human ratings, with each of the 30 unique cases evaluated by two evaluators. Our method achieved an 81.63% win rate when compared to PPTAgent, demonstrating its superiority and consistent with evaluation by MLLMs. We also examined the agreement between human ratings and MLLM-based evaluations, observing an average Pearson correlation of 0.64 (with 0.5 generally considered a strong correlation). Further details are provided at our project website.

7.3 Ablation Studies

To assess the effectiveness of different components, we randomly sampled 30 cases and conducted ablations on two variants: (1) removing content preference guidance, and (2) disabling the chain-of-speech mechanism.

From Tab. 3, we highlight two key observations: (1) Removing content preference distillation notably degrades performance across all preference-based metrics, especially coverage, flow, and content structure (by around 10%). This not only validates the core hypothesis that modeling user-specific content preferences, even from implicit and unlabeled examples, is essential for generating slides aligned with communicative intent, but also demonstrates the effectiveness of our preference distillation module in capturing and leveraging such nuanced user signals. (2) Disabling the chain-of-speech module results in a clear drop in overall performance, especially in general content quality (66.4% $\rightarrow$ 47.3%). This underscores the importance of aligning slide planning with anticipated narration to improve clarity and informativeness.

8 Conclusion and Limitations

In this paper, we explore the subjective nature of paper-to-slides generation. We propose a practical yet challenging task conditioned on user preferences captured through natural, real-world inputs. We introduce a human-like agentic framework that distills implicit preferences and progressively generates editable slides. A novel chain-of-speech mechanism bridges slide planning with oral narration, enhancing coherence and enabling downstream applications like video presentation. We also construct a benchmark dataset that simulates diverse user preferences and design interpretable metrics for robust evaluation. Experiments show the superiority of our method in both preference alignment and overall generation quality, paving the way for more personalized and flexible slide generation.

However, several limitations remain. First, our benchmark focuses exclusively on scientific papers. Extending it to broader domains (e.g., business reports, educational materials, advertising content) could benefit more fields. Second, while our training-free framework offers strong flexibility and adaptability, exploring end-to-end multimodal training for preference-guided slide generation is a promising direction. Third, although our MLLM-based evaluation shows general alignment with human judgment, a noticeable gap remains. We observe that MLLMs lack the fine-grained perception of humans and exhibit inherent self-bias, whereas cross-judge evaluations (e.g., Qwen judging GPT-based models) tend to yield results more consistent with human ratings. Designing more human-aligned evaluation protocols remains a valuable direction for future research.

Acknowledgments

This research is supported by the National Research Foundation Singapore under its AI Singapore Programme (Award Number: AISG3-RP-2022-030).

References

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; Zhong, H.; Zhu, Y.; Yang, M.; Li, Z.; Wan, J.; Wang, P.; Ding, W.; Fu, Z.; Xu, Y.; Ye, J.; Zhang, X.; Xie, T.; Cheng, Z.; Zhang, H.; Yang, Z.; Xu, H.; and Lin, J. 2025. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*.

Bandyopadhyay, S.; Maheshwari, H.; Natarajan, A.; and Saxena, A. 2024. Enhancing Presentation Slide Generation by LLMs with a Multi-Staged End-to-End Approach. In *Proceedings of the 17th International Natural Language Generation Conference*, 222–229.

Bartsch, R. A.; and Cobern, K. M. 2003. Effectiveness of PowerPoint presentations in lectures. *Computers & Education*, 41(1): 77–86.

Cachola, I.; Cucerzan, S.; Herring, A.; Mijovic, V.; Oveson, E.; and Jauhar, S. K. 2024. Knowledge-Centric Templatic Views of Documents. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 15460–15476.

Cao, J.; Jiao, D.; Yan, Q.; Zhang, W.; Tang, S.; and Zhuang, Y. 2024. IDEAL: Leveraging Infinite and Dynamic Characterizations of Large Language Models for Query-focused Summarization. *arXiv preprint arXiv:2407.10486*.

Cheng, X.; Gao, S.; Zhang, Y.; Wang, Y.; Chen, X.; Li, M.; Zhao, D.; and Yan, R. 2023. Towards Personalized Review Summarization by Modeling Historical Reviews from Customer and Product Separately. *arXiv preprint arXiv:2301.11682*.

Costa, M. J.; Amaro, H.; and Gonçalo Oliveira, H. 2023. SmartEDU: Accelerating Slide Deck Production with Natural Language Processing. In *International Conference on Applications of Natural Language to Information Systems*, 109–123.

Fan, A.; Grangier, D.; and Auli, M. 2018. Controllable Abstractive Summarization. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, 45–54.

Fu, T.-J.; Wang, W. Y.; McDuff, D.; and Song, Y. 2022. DOC2PPT: Automatic Presentation Slides Generation from Scientific Documents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 634–642.

Ge, J.; Wang, Z. Z.; Zhou, X.; Peng, Y.-H.; Subramanian, S.; Tan, Q.; Sap, M.; Suhr, A.; Fried, D.; Neubig, G.; et al. 2025. AutoPresent: Designing Structured Visuals from Scratch. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2902–2911.

Ghodratnama, S.; and Zakershahrak, M. 2024. SumRecom: A Personalized Summarization Approach by Learning from Users' Feedback. *arXiv preprint arXiv:2408.07294*.

Hong, F.-T.; Xu, Z.; Zhou, Z.; Zhou, J.; Li, X.; Lin, Q.; Lu, Q.; and Xu, D. 2025. Audio-visual Controlled Video Diffusion with Masked Selective State Spaces Modeling for Natural Talking Head Generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 12549–12558.

Hu, Q.; Moon, G.; and Ng, H. T. 2024. From Moments to Milestones: Incremental Timeline Summarization Leveraging Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 7232–7246.

Jiang, Z.; Ren, Y.; Li, R.; Ji, S.; Zhang, B.; Ye, Z.; Zhang, C.; Jionghao, B.; Yang, X.; Zuo, J.; et al. 2025. MegaTTS 3: Sparse Alignment Enhanced Latent Diffusion Transformer for Zero-Shot Speech Synthesis. *arXiv preprint arXiv:2502.18924*.

Li, D.-W.; Huang, D.; Ma, T.; and Lin, C.-Y. 2021. Towards Topic-Aware Slide Generation For Academic Papers With Unsupervised Mutual Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 13243–13251.

Li, J.; Li, H.; and Zong, C. 2019. Towards Personalized Review Summarization via User-Aware Sequence Network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 6690–6697.

LMDeploy Contributors. 2023. LMDeploy: A Toolkit for Compressing, Deploying, and Serving LLM. https://github.com/InternLM/lmdeploy.

Maheshwari, H.; Bandyopadhyay, S.; Garimella, A.; and Natarajan, A. 2024. Presentations are not always linear! GNN meets LLM for Text Document-to-Presentation Transformation with Attribution. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 15948–15962.

Mondal, I.; Li, Z.; Hou, Y.; Natarajan, A.; Garimella, A.; and Boyd-Graber, J. L. 2024a. SciDoc2Diagrammer-MAF: Towards Generation of Scientific Diagrams from Documents guided by Multi-Aspect Feedback Refinement. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 13342–13375.

Mondal, I.; Shwetha, S.; Natarajan, A.; Garimella, A.; Bandyopadhyay, S.; and Boyd-Graber, J. 2024b. Presentations by the Humans and For the Humans: Harnessing LLMs for Generating Persona-Aware Slides from Documents. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*, 2664–2684.

Mukherjee, S.; Jangra, A.; Saha, S.; and Jatowt, A. 2022. Topic-aware Multimodal Summarization. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, 387–398.

OpenAI. 2025a. ChatGPT. https://chatgpt.com/.

OpenAI. 2025b. GPT-4.1. https://platform.openai.com/docs/models/gpt-4.1.

OpenRouter Team. 2025. OpenRouter API Documentation. https://openrouter.ai/.

Pang, W.; Lin, K. Q.; Jian, X.; He, X.; and Torr, P. 2025. Paper2Poster: Towards Multimodal Poster Automation from Scientific Papers. *arXiv preprint arXiv:2505.21497*.

Qin, Z.; Zhao, W.; Yu, X.; and Sun, X. 2023. OpenVoice: Versatile Instant Voice Cloning. *arXiv preprint arXiv:2312.01479*.

Qorib, M. R.; Hu, Q.; and Ng, H. T. 2025. Just What You Desire: Constrained Timeline Summarization with Self-Reflection for Enhanced Relevance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 25065–25073.

Shi, J.; Zhang, Z.; Wu, B.; Liang, Y.; Fang, M.; Chen, L.; and Zhao, Y. 2025. PresentAgent: Multimodal Agent for Presentation Video Generation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 760–773.

Sun, E.; Hou, Y.; Wang, D.; Zhang, Y.; and Wang, N. X. 2021. D2S: Document-to-Slide Generation Via Query-Based Text Summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1405–1418.

Xu, H.; Liu, H.; Lv, Z.; Yang, Q.; and Wang, W. 2023a. Pretrained Personalized Review Summarization with Effective Salience Estimation. In *Findings of the Association for Computational Linguistics: ACL 2023*, 10743–10754.

Xu, R.; Wang, S.; Liu, Y.; Wang, S.; Xu, Y.; Iter, D.; He, P.; Zhu, C.; and Zeng, M. 2023b. LMGQS: A Large-scale Dataset for Query-focused Summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 14764–14776.

Xu, S.; and Wan, X. 2021. Neural Content Extraction for Poster Generation of Scientific Papers. *arXiv preprint arXiv:2112.08550*.

Xu, Y.; Ma, X.; Qiu, J.; and Zhao, H. 2025. Textual-to-Visual Iterative Self-Verification for Slide Generation. *arXiv preprint arXiv:2502.15412*.

Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; Lin, H.; Yang, J.; Tu, J.; Zhang, J.; Yang, J.; Yang, J.; Zhou, J.; Lin, J.; Dang, K.; Lu, K.; Bao, K.; Yang, K.; Yu, L.; Li, M.; Xue, M.; Zhang, P.; Zhu, Q.; Men, R.; Lin, R.; Li, T.; Xia, T.; Ren, X.; Ren, X.; Fan, Y.; Su, Y.; Zhang, Y.; Wan, Y.; Liu, Y.; Cui, Z.; Zhang, Z.; and Qiu, Z. 2024. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115*.

Yu, H.; and Han, J. 2022. Survey of Query-Based Text Summarization. *arXiv preprint arXiv:2211.11548*.

Yujian, L.; and Bo, L. 2007. A Normalized Levenshtein Distance Metric. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6): 1091–1095.

Zheng, H.; Guan, X.; Kong, H.; Zheng, J.; Lin, H.; Lu, Y.; He, B.; Han, X.; and Sun, L. 2025. PPTAgent: Generating and Evaluating Presentations Beyond Text-to-Slides. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 14413–14429.

Zheng, L.; Zhang, Y.; Guo, H.; Pan, J.; Tan, Z.; Lu, J.; Tang, C.; An, B.; and Yan, S. 2024. MEMO: Memory-Guided Diffusion for Expressive Talking Video Generation. *arXiv preprint arXiv:2412.04448*.

Zhu, Z.; Lin, K. Q.; and Shou, M. Z. 2025. Paper2Video: Automatic Video Generation from Scientific Papers. *arXiv preprint arXiv:2510.05096*.