PresentAgent: Multimodal Agent for Presentation Video Generation

Jingwei Shi[1*] **Zeyu Zhang**[1*†] **Biao Wu**[2*] **Yanjie Liang**[1*]
Meng Fang[3] **Ling Chen**[2] **Yang Zhao**[4‡]
[1]AI Geeks, Australia
[2]Australian Artificial Intelligence Institute, Australia
[3]University of Liverpool, United Kingdom
[4]La Trobe University, Australia
[*]Equal contribution. [†]Project lead. [‡]Corresponding author: y.zhao2@latrobe.edu.au.

Abstract

We present PresentAgent, a multimodal agent that transforms long-form documents into narrated presentation videos. While existing approaches are limited to generating static slides or text summaries, our method advances beyond these limitations by producing fully synchronized visual and spoken content that closely mimics human-style presentations. To achieve this integration, PresentAgent employs a modular pipeline that systematically segments the input document, plans and renders slide-style visual frames, generates contextual spoken narration with large language models and Text-to-Speech models, and seamlessly composes the final video with precise audio-visual alignment. Given the complexity of evaluating such multimodal outputs, we introduce PresentEval, a unified assessment framework powered by Vision-Language Models that comprehensively scores videos across three critical dimensions: content fidelity, visual clarity, and audience comprehension through prompt-based evaluation. Our experimental validation on a curated dataset of 30 document–presentation pairs demonstrates that PresentAgent approaches human-level quality across all evaluation metrics. These results highlight the significant potential of controllable multimodal agents in transforming static textual materials into dynamic, effective, and accessible presentation formats. Code will be available at https://github.com/AIGeeksGroup/PresentAgent.

1 Introduction

Presentations are a widely used and effective medium for conveying complex ideas. By combining visual elements, structured narration, and spoken explanations, they enable information to unfold progressively and be more easily understood by diverse audiences (Fu et al., 2022). Despite their proven effectiveness, creating high-quality presentation videos from long-form documents—such as business reports, technical manuals, policy briefs, or academic papers—typically requires considerable manual effort (Li et al., 2023). This process involves identifying key content, designing slide layouts, writing scripts, recording narration, and aligning all elements into a coherent multimodal output.

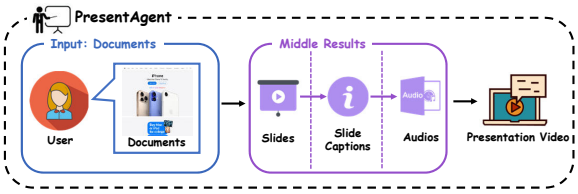

Figure 1: **Overview of PresentAgent.** It takes documents (e.g., web pages) as input and follows a generation pipeline: (1) document processing, (2) structured slide generation, (3) synchronized caption creation, and (4) audio synthesis. The final output is a presentation video combining visual slides with aligned narration. The purple-highlighted middle results emphasize the system's key transitional outputs during generation.

Although recent advancements in AI have enabled progress in related areas such as document-to-slide generation (Fu et al., 2022; Zheng et al., 2025a; Pang et al., 2025; Zhang et al., 2024) and text-to-video synthesis (Yang et al., 2024c; Li et al., 2023; Xue et al., 2025; Khachatryan et al., 2023; He et al., 2023; Solanki and Khublani, 2024), a critical gap remains: these methods either produce static visual summaries or generic video clips without structured narration, limiting their effectiveness for structured communication tasks like presentations.

To bridge this gap, we introduce the task of Document-to-Presentation Video Generation, which aims to automatically convert a structured or unstructured document into a narrated video presentation composed of synchronized slides and speech. This task presents unique challenges as it goes beyond traditional summarization (Lewis et al., 2019; Beltagy et al., 2020; Chen and Yang, 2021; Wang

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 760–773
November 4-9, 2025 ©2025 Association for Computational Linguistics

et al., 2024a) or text-to-speech (Tachibana et al., 2018; Ren et al., 2019; Popov et al., 2021; Ni et al., 2022) pipelines by requiring selective content abstraction, layout-aware planning (Wang et al., 2025), and precise multimodal alignment (Li et al., 2024) between visuals and narration. In contrast to prior work that focuses on either static slide and image generation (Zheng et al., 2025a; Deng et al., 2025; Xie et al., 2024) or audio summarization in isolation, our objective is to produce a fully integrated, viewer-ready video experience that closely mimics how human presenters deliver information in real-world scenarios.

To tackle these challenges, we propose a modular generation framework named PresentAgent, as shown in Figure 1. Given an input document, the system first segments it into semantic blocks through outline planning, then generates layout-guided slide visuals for each block and rewrites the key message into oral-style narration. Subsequently, these are then synthesized into audio and combined with the slide visuals to produce a time-aligned presentation video. Importantly, our pipeline is designed to be domain-adaptable and controllable, enabling broad applicability across document types and presentation styles.

Recognizing the need for rigorous evaluation of such complex multimodal outputs, we curate a test set of 30 human-authored document-video pairs spanning diverse domains, including education, finance, policy, and scientific communication. To comprehensively assess system performance, we further introduce a two-path evaluation strategy that combines fact-based comprehension assessment (via fixed multiple-choice quizzes) and preference-based scoring using vision-language models. This dual-pronged approach captures both objective correctness and subjective quality in video delivery.

Experiment results demonstrate that our method produces fluent, well-structured, and informative presentation videos, approaching human-level performance in both content delivery and viewer comprehension. These findings highlight the potential of combining language models, layout generation, and multimodal synthesis for creating explainable and scalable presentation systems from raw documents.

In general, our contributions are summarized as follows:

- We formulate and address the novel task of document-to-presentation video generation, which aims to produce narrated, slide-structured videos from long-form documents across diverse domains.
- We propose PresentAgent, a modular generation framework that integrates document parsing, layout-aware slide composition, narration planning, and audio-visual synchronization, enabling controllable and interpretable generation.
- We introduce PresentEval, a multi-dimensional evaluation framework powered by Vision-Language Models (VLMs), which scores videos along content, visual, and comprehension dimensions via prompt-based judging.
- We create a test set of 30 real-world document–presentation pairs and demonstrate through experiments and ablations that PresentAgent approaches human-level performance and significantly outperforms competitive variants.

2 Presentation Benchmark

The benchmark supports evaluation not only of fluency and fidelity, but also of downstream comprehension. Following the methodology introduced in Paper2Poster (Pang et al., 2025), we construct a quiz-style evaluation protocol (§5), where vision-language models are asked to answer factual content questions using only the generated video (slides + narration), simulating an audience's understanding. Human-authored videos are used as reference standards for both score calibration and upper-bound comparison. As shown in Figure 5, our benchmark encompasses four representative document types (academic papers, web pages, technical blogs, and slides) paired with human-authored videos, covering diverse real-world domains like education, research, and business reports.

We adopt a unified, model-based evaluation framework to assess the generated presentation videos. All evaluations are conducted using a vision-language model, guided by dimension-specific prompts tailored to different assessment objectives. The framework consists of two complementary components: (1) objective quiz evaluation, which measures factual accuracy through multiple-choice question answering; and (2) subjective scoring, which rates Content Quality, Visual or Audio Quality, and Comprehension Clarity on a 1–5 scale. Together, these metrics provide a comprehensive assessment of both the quality and informativeness

Figure 2: **Overview of our framework.** Our approach addresses the full pipeline of document-to-presentation video generation and evaluation. Left: Given diverse input documents—including papers, websites, blogs, slides, and PDFs—PresentAgent generates narrated presentation videos by producing synchronized slide decks with audio. Right: To evaluate these videos, we introduce PresentEval, a two-part evaluation framework: (1) Objective Quiz Evaluation (top), which measures factual comprehension using Qwen-VL; and (2) Subjective Scoring (bottom), which uses vision-language models to rate content quality, visual design, and audio comprehension across predefined dimensions.

of the generated videos.

2.1 Doc2Present Dataset

To support the evaluation of document to presentation video generation, we curate the Doc2Present Benchmark, a diverse dataset of document–presentation video pairs spanning multiple domains. Unlike prior benchmarks focused on research abstracts or slide generation, our dataset includes documents such as business reports, product manuals, policy briefs, and instructional texts, each paired with a human-crafted presentation video.We collect 30 high-quality video samples from public platforms, educational repositories, and professional presentation archives, further details regarding the data sources and statistical information of the dataset can be found in the appendix F.

2.2 PresentEval

To assess the quality of generated presentation videos, we adopt two complementary evaluation strategies: Objective Quiz Evaluation and Subjective Scoring, as shown in Figure 2. For each video, we provide the vision-language model with the complete set of slide images and the full narration transcript as a unified input—simulating how a real viewer would experience the presentation. In Objective Quiz Evaluation, the model answers a fixed set of factual questions to determine whether the video accurately conveys the key information from the source content. In Subjective Scoring, the model evaluates the video along three dimensions: the coherence of the narration, the clarity and design of the visuals, and the overall ease of understanding. All evaluations are conducted without ground-truth references and rely entirely on the model's interpretation of the presented content.

Objective Quiz Evaluation To evaluate whether a generated presentation video effectively conveys the core content of its source document, we use a fixed-question comprehension evaluation protocol. Specifically, we manually design five multiple-choice questions for each document, tailored to its content. These questions focus on key aspects such as topic recognition, structural understanding, and main argument extraction. As shown in Table 2, during evaluation, a vision-language model is given the video, including both visual frames and audio transcript, and asked to answer the five questions. Each question has four options, with one correct answer, annotated based on a human-created reference video. The final comprehension score (ranging from 0 to 5) reflects how many questions the model answered correctly, serving as a direct measure of how well the video communicates the original document.

Subjective Scoring To evaluate the quality of generated presentation videos, we adopt a prompt-based assessment using vision-language models. Instead of relying on human references or fixed metrics, we ask the model to evaluate each video from a viewer's perspective, using its own reasoning and preferences. The evaluation focuses on three aspects: coherence of narration, clarity and aesthetics of visuals, and overall ease of understanding. The model is shown the video and audio, and gives a score (1–5) with a brief explanation

for each aspect. This enables scalable, consistent, and human-aligned evaluation without manual references. As shown in Table 3, we design different prompts for different modalities and tasks to ensure targeted and effective assessment.

3 PresentAgent

To convert a long-form document into a narrated presentation video, we design a multi-stage generation framework that mirrors how human presenters prepare slides and talk tracks, as shown in Figure 3. Our method proceeds in four steps: segmenting the document into semantic units, composing slides with layout-aware structures, generating oral-style narration for each slide and assembling the visual and audio components into a synchronized video. This modular design supports controllability, interpretability, and multimodal alignment, enabling both high-quality generation and fine-grained evaluation. The following sections describe each component in detail.

3.1 Problem Formulation

Our method is designed to transform a long-form document into a structured presentation video through a multi-stage generation pipeline. We provide a formal description to highlight the key difference between our approach and conventional slide-based methods.

Conventional approaches often focus on generating slide elements S directly from a document chunk C, as in Equation 1, where each element includes text or image content, layout attributes, and visual style:

$$S = \{e_1, e_2, ..., e_n\} = f(C) \tag{1}$$

In contrast, we treat the entire document D as a globally structured input and generate a presentation in three steps: (1) a sequence of semantic segments $\{C_1, ..., C_K\}$ via outline planning, (2) a set of slides $\{S_1, ..., S_K\}$, each paired with a narrated audio track T_k generated by first producing a slide-specific script and then converting it to speech, and (3) a video V composed of visual and audio content aligned over time. This is defined as:

$$V = \texttt{Compose}(\{(S_1, T_1), ..., (S_K, T_K)\}) = g(D) \tag{2}$$

Rather than editing predefined templates or layouts, our system first identifies high-level structure in the document and then generates slide visuals and narration from scratch. This pipeline supports controllability, modular evaluation, and multimodal alignment for downstream comprehension and quality assessment.

3.2 Slide Planning and Composition

Our slide generation module is inspired by the editing-based paradigm proposed in PPTAgent (Zheng et al., 2025b), which formulates presentation construction as a structured editing process over HTML-like layouts. While PPTAgent focuses on producing editable `.pptx` slides, our goal is to generate visually coherent, narration-ready slide frames for downstream video synthesis. We re-implement the core idea in a self-contained pipeline tailored to multimodal synchronization.

We begin by segmenting the input document into coherent content blocks using a lightweight LLM-based parser. Each block is assigned a corresponding slide type such as bullet slide, figure-description, or title-intro, and matched with a predefined layout schema encoded in HTML. Unlike retrieval-based template matching, our system uses semantic and structural cues to map content to layout patterns in a rule-guided manner.

To populate the slide, we define a set of editable operations such as `replace_text`, `insert_image`, and `add_list`, which are applied to the layout structure. These instructions are generated by prompting a language model with the content block and layout constraints. Slides are then rendered into static visual frames using `python-pptx` or HTML-based renderers.

3.3 Narration and Audio Synthesis

To transform the static slides into an engaging presentation, we generate a spoken narration for each slide and synthesize it into audio. The process involves two components: narration script generation and text-to-speech synthesis.

For each content block corresponding to a slide, we prompt a language model to generate a concise, oral-style narration. The model is instructed to rewrite the key message of the slide into natural spoken language, avoiding dense text or technical jargon. We apply length control to ensure each narration falls within a target duration, typically between 30 and 150 seconds. Once the narration script is obtained, we synthesize the corresponding audio using a text-to-speech system. Each narration audio is paired with its slide and timestamped,

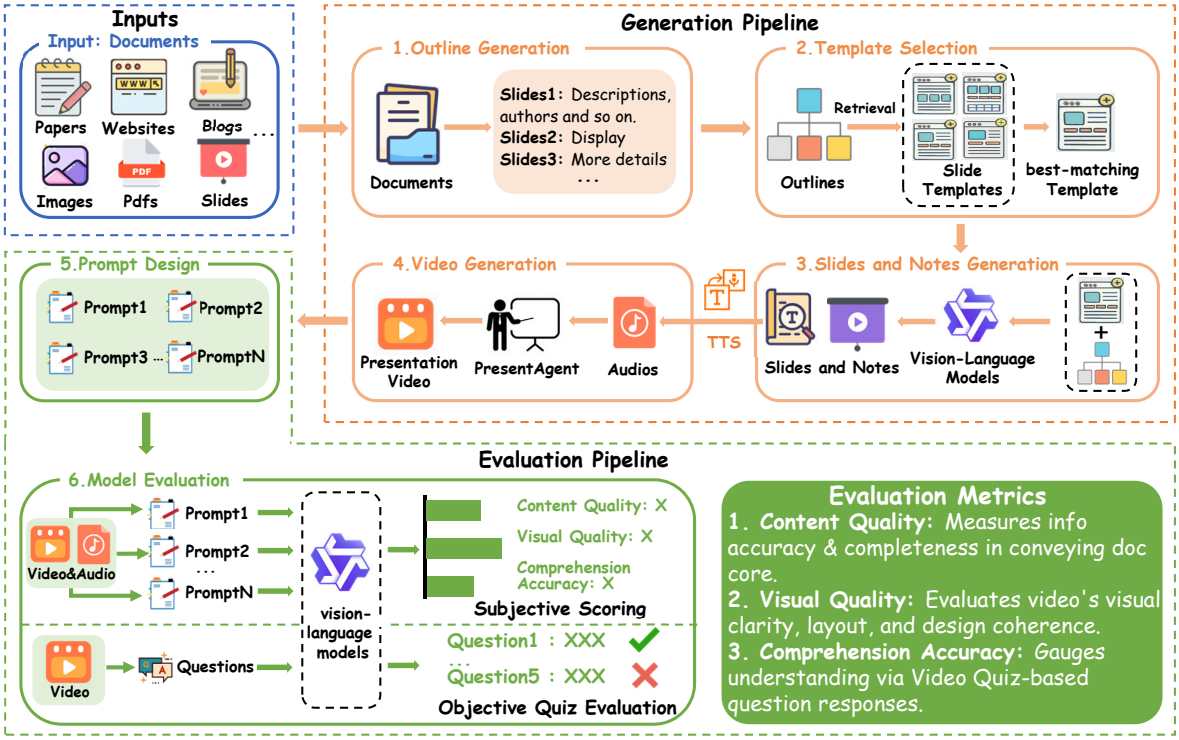

Figure 3: **Overview of the PresentAgent framework.** Our system takes diverse documents (e.g., papers, websites, PDFs) as input and follows a modular generation pipeline. It first performs outline generation (Step 1) and retrieves the most suitable template (Step 2), then generates slides and narration notes via a vision-language model (Step 3). The notes are converted into audio via TTS and composed into a presentation video (Step 4). To evaluate video quality, we design multiple prompts (Step 5) and feed them into a VLM-based scoring pipeline (Step 6) that outputs dimension-specific metrics.

forming the basis for synchronized video rendering in the next stage.

3.4 Video Assembly

In the final stage, we assemble the slide images and narration audio into a coherent, time-aligned presentation video. Each slide frame is displayed for the duration of its corresponding audio segment, with optional transitions between segments. We use video processing libraries such as `ffmpeg` to compose the visual and audio tracks. Each slide is rendered as a static frame, and the narration is added as synchronized voiceover audio. The output is a fully rendered video file in standard formats such as `.mp4`, suitable for presentation, sharing, or further editing. This stage completes the transformation from a raw document into a narrated, structured presentation video.

4 Experiments

We conduct experiments to evaluate the effectiveness of our proposed system in generating high-quality, narrated presentation videos. Given the novelty of the task, our focus is not on competing with existing baselines, but rather on assessing the performance of our full system relative to human-created presentations. Comprehension accuracy is determined based on performance in the PresentEval task. Evaluation setup can be found in appendix H.

4.1 Main Results

Table 1 presents evaluation results, covering both factual comprehension (Quiz Accuracy) and preference-based quality scores for video and audio outputs. In terms of quiz accuracy, most PresentAgent variants perform comparably to or better than the human reference (0.56), with Claude-3.7-sonnet (Anthropic, 2024) achieving the highest accuracy at 0.64, suggesting strong alignment between the generated content and the source document. Other models such as Qwen-VL-Max (Bai et al., 2025) and Gemini-2.5-flash (DeepMind, 2024) scored slightly lower (0.52), indicating room for improvement in factual grounding.

In terms of subjective quality, human-created

Method	Model	Quiz Accuracy	Video Score				Audio Score			
			Content	Visual	Comp.	Mean	Content	Audio	Comp.	Mean
Human	Human	0.56	4.0	4.6	4.8	4.47	4.8	4.6	5.0	4.80
PresentAgent	Claude-3.7-sonnet	0.64	4.0	4.0	4.0	4.00	4.2	4.6	4.8	4.53
PresentAgent	Qwen-VL-Max	0.52	4.2	4.8	4.4	4.47	4.6	4.2	5.0	4.60
PresentAgent	Gemini-2.5-pro	0.52	4.2	4.4	4.4	4.33	4.2	4.0	4.8	4.33
PresentAgent	Gemini-2.5-flash	0.52	4.2	5.0	3.8	4.33	4.2	4.2	4.8	4.40
PresentAgent	GPT-4o-Mini	0.64	4.8	4.6	4.6	4.67	4.0	4.4	4.8	4.40
PresentAgent	GPT-4o	0.56	4.0	4.2	3.6	3.93	4.2	4.4	4.8	4.47

Table 1: Detailed evaluation results on the 5-document test set. Fact-based evaluation includes accuracy on five fixed quiz questions (Q1–Q5). Preference-based evaluation includes 1–5 scale scores for content fidelity, visual design, and overall clarity. Each Quality Score group has a calculated mean column.

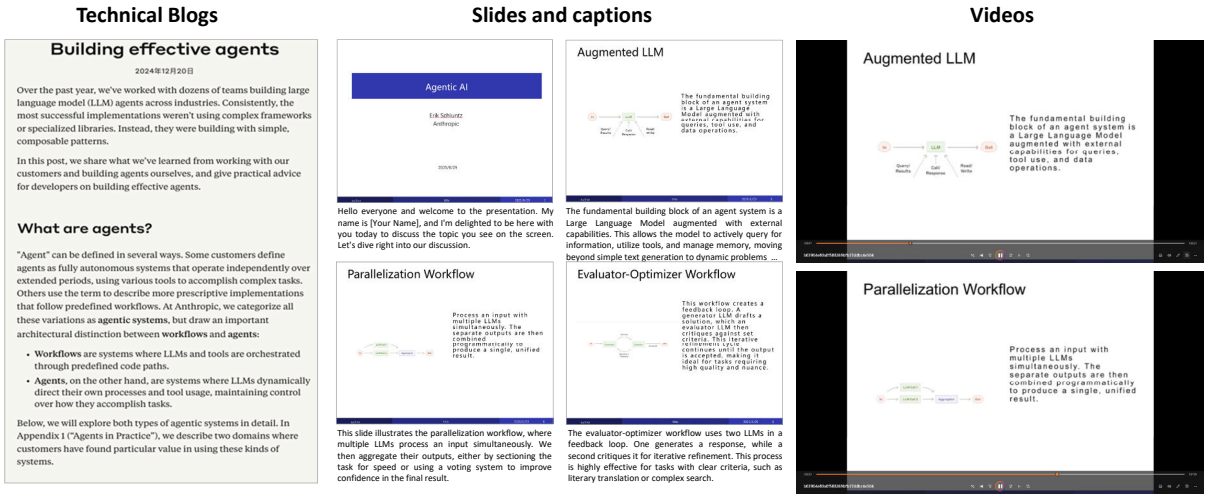

Figure 4: **PresentAgent Demo.** Automatically generates academic-style slides and narrated videos from research papers, streamlining the transformation from written content to engaging visual presentations.

presentations still lead with the highest video and audio scores overall. However, several PresentAgent variants show competitive performance. For example, GPT-4o-Mini (Achiam et al., 2023) achieves top scores in video content and visual appeal (both at or near 4.8), while Claude-3.7-sonnet (Anthropic, 2024) delivers the most balanced audio quality (mean 4.53). Interestingly, Gemini-2.5-flash (DeepMind, 2024) scores highest in visual quality (5.0) but lower in comprehension, reflecting a trade-off between aesthetics and clarity. These results highlight the effectiveness of our modular pipeline and the usefulness of our unified PresentEval framework in capturing diverse aspects of presentation quality.

4.2 Analysis

Figure 4 Presents a full example of a PresentAgent-auto-generated presentation video, showing a technical blog turned into a narrated presentation. The system identifies structural segments (e.g., introduction, technical explanations) and generates slides with oral-style captions and synchronized speech, covering topics like "parallelization workflow" and "agent system architecture" to demonstrate its ability to keep technical accuracy while delivering content clearly and conversationally.

5 Conclusion

In conclusion, we presented PresentAgent, a modular system for transforming long-form documents into narrated presentation videos. By addressing the challenges of slide planning, narration synthesis, and synchronized rendering, PresentAgent enables structured, controllable, and reusable multimodal outputs. To evaluate this novel task, we introduced a diverse benchmark and proposed complementary factual and preference-based metrics. Experimental results show that PresentAgent generates coherent, engaging, and informative presentations, approaching human quality. This work lays the groundwork for automated, explainable content

generation and opens new directions for research in multimodal communication across education, business, and accessibility.

References

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

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

Galen Andrew and Jianfeng Gao. 2007. Scalable training of L1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

Anthropic. 2024. Claude 3 technical overview. https://www.anthropic.com/news/claude-3. Accessed: 2025-06-30.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Jiaao Chen and Diyi Yang. 2021. Structure-aware abstractive conversation summarization via discourse and action graphs. *arXiv preprint arXiv:2104.08400*.

Google DeepMind. 2024. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. https://deepmind.google/technologies/gemini/. Accessed: 2025-06-30.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, and 1 others. 2025. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*.

Tsu-Jui Fu, William Yang Wang, Daniel McDuff, and Yale Song. 2022. Doc2ppt: Automatic presentation slides generation from scientific documents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 634–642.

Jiaxin Ge, Zora Zhiruo Wang, Xuhui Zhou, Yi-Hao Peng, Sanjay Subramanian, Qinyue Tan, Maarten Sap, Alane Suhr, Daniel Fried, Graham Neubig, and Trevor Darrell. 2025. Autopresent: Designing structured visuals from scratch. *arXiv preprint arXiv:2501.00912*.

Yingqing He, Menghan Xia, Haoxin Chen, Xiaodong Cun, Yuan Gong, Jinbo Xing, Yong Zhang, Xintao Wang, Chao Weng, Ying Shan, and 1 others. 2023. Animate-a-story: Storytelling with retrieval-augmented video generation. *arXiv preprint arXiv:2307.06940*.

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

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and 1 others. 2024. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Xin Li, Wenqing Chu, Ye Wu, Weihang Yuan, Fanglong Liu, Qi Zhang, Fu Li, Haocheng Feng, Errui Ding, and Jingdong Wang. 2023. Videogen: A reference-guided latent diffusion approach for high definition text-to-video generation. *arXiv preprint arXiv:2309.00398*.

Kevin Qinghong Lin, Linjie Li, Difei Gao, Qinchen Wu, Mingyi Yan, Zhengyuan Yang, Lijuan Wang, and Mike Zheng Shou. 2024a. Videogui: A benchmark for gui automation from instructional videos. *arXiv preprint arXiv:2406.10227*.

Kevin Qinghong Lin, Linjie Li, Difei Gao, Zhengyuan Yang, Shiwei Wu, Zechen Bai, Weixian Lei, Lijuan Wang, and Mike Zheng Shou. 2024b. Showui: One vision-language-action model for gui visual agent. *arXiv preprint arXiv:2411.17465*.

Pan Lu, Bowen Chen, Sheng Liu, Rahul Thapa, Joseph Boen, and James Zou. 2025. Octotools: An agentic framework with extensible tools for complex reasoning. *arXiv preprint arXiv:2502.11271*.

Shravan Nayak, Xiangru Jian, Kevin Qinghong Lin, Juan A. Rodriguez, Montek Kalsi, Rabiul Awal, Nicolas Chapados, M. Tamer Özsu, Aishwarya Agrawal, David Vazquez, Christopher Pal, Perouz Taslakian, Spandana Gella, and Sai Rajeswar. 2025. Ui-vision: A desktop-centric gui benchmark for visual perception and interaction. *arXiv preprint arXiv:2503.15661*.

Junrui Ni, Liming Wang, Heting Gao, Kaizhi Qian, Yang Zhang, Shiyu Chang, and Mark Hasegawa-Johnson. 2022. Unsupervised text-to-speech synthesis by unsupervised automatic speech recognition. *arXiv preprint arXiv:2203.15796*.

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

Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail Kudinov. 2021. Grad-tts: A diffusion probabilistic model for text-to-speech. In *International conference on machine learning*, pages 8599–8608. PMLR.

Yujia Qin, Yining Ye, Junjie Fang, Haoming Wang, Shihao Liang, Shizuo Tian, Junda Zhang, Jiahao Li, Yunxin Li, Shijue Huang, and 1 others. 2025. Ui-tars: Pioneering automated gui interaction with native agents. *arXiv preprint arXiv:2501.12326*.

Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. Yara parser: A fast and accurate dependency parser. *Computing Research Repository*, arXiv:1503.06733. Version 2.

Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2019. Fastspeech: Fast, robust and controllable text to speech. *Advances in neural information processing systems*, 32.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, and et al. 2023. Toolformer: Language models can teach themselves to use tools. *arXiv preprint arXiv:2302.04761*.

Shivam R Solanki and Drupad K Khublani. 2024. From script to screen: Unveiling text-to-video generation. In *Generative Artificial Intelligence: Exploring the Power and Potential of Generative AI*, pages 81–112. Springer.

Qiushi Sun, Kanzhi Cheng, Zichen Ding, Chuanyang Jin, Yian Wang, Fangzhi Xu, Zhenyu Wu, Chengyou Jia, Liheng Chen, Zhoumianze Liu, and 1 others. 2024. Os-genesis: Automating gui agent trajectory construction via reverse task synthesis. *arXiv preprint arXiv:2412.19723*.

Hideyuki Tachibana, Katsuya Uenoyama, and Shunsuke Aihara. 2018. Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4784–4788. IEEE.

Baode Wang, Biao Wu, Weizhen Li, Meng Fang, Yanjie Liang, Zuming Huang, Haozhe Wang, Jun Huang, Ling Chen, Wei Chu, and 1 others. 2025. Infinity parser: Layout aware reinforcement learning for scanned document parsing. *arXiv preprint arXiv:2506.03197*.

Guanghua Wang, Priyanshi Garg, and Weili Wu. 2024a. Segmented summarization and refinement: A pipeline for long-document analysis on social media. *Journal of Social Computing*, 5(2):132–144.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

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

Yuan Wang, Di Huang, Yaqi Zhang, Wanli Ouyang, Jile Jiao, Xuetao Feng, Yan Zhou, Pengfei Wan, Shixiang Tang, and Dan Xu. 2024d. Motiongpt-2: A general-purpose motion-language model for motion generation and understanding. *arXiv preprint arXiv:2410.21747*.

Biao Wu, Yanda Li, Meng Fang, Zirui Song, Zhiwei Zhang, Yunchao Wei, and Ling Chen. 2024. Foundations and recent trends in multimodal mobile agents: A survey. *arXiv preprint arXiv:2411.02006*.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. 2024. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, and 1 others. 2025. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*.

Qiyao Xue, Xiangyu Yin, Boyuan Yang, and Wei Gao. 2025. Phyt2v: Llm-guided iterative self-refinement for physics-grounded text-to-video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18826–18836.

John Yang, Carlos Jimenez, Alexander Wettig, Kilian Lieret, Shunyu Yao, Karthik Narasimhan, and Ofir Press. 2024a. Swe-agent: Agent-computer interfaces enable automated software engineering. *Advances in Neural Information Processing Systems*, 37:50528–50652.

Ke Yang, Jiateng Liu, John Wu, Chaoqi Yang, Yi R Fung, Sha Li, Zixuan Huang, Xu Cao, Xingyao Wang, Yiquan Wang, and 1 others. 2024b. If llm is the wizard, then code is the wand: A survey on how code empowers large language models to serve as intelligent agents. *arXiv preprint arXiv:2401.00812*.

Rui Yang, Lin Song, Yanwei Li, Sijie Zhao, Yixiao Ge, Xiu Li, and Ying Shan. 2023a. Gpt4tools: Teaching large language model to use tools via self-instruction. *Advances in Neural Information Processing Systems*, 36:71995–72007.

Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. 2023b. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, and 1 others. 2024c. Cogvideox: Text-to-video diffusion

models with an expert transformer. *arXiv preprint arXiv:2408.06072*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*.

Murong Yue, Wenlin Yao, Haitao Mi, Dian Yu, Ziyu Yao, and Dong Yu. 2024. Dots: Learning to reason dynamically in llms via optimal reasoning trajectories search. *arXiv preprint arXiv:2410.03864*.

Zeyu Zhang, Yiran Wang, Biao Wu, Shuo Chen, Zhiyuan Zhang, Shiya Huang, Wenbo Zhang, Meng Fang, Ling Chen, and Yang Zhao. 2024. Motion avatar: Generate human and animal avatars with arbitrary motion. *arXiv preprint arXiv:2405.11286*.

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

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

Zixiang Zhou, Yu Wan, and Baoyuan Wang. 2024. Avatargpt: All-in-one framework for motion understanding planning generation and beyond. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1357–1366.

A Related Work

A.1 Document-to-Multimodal Generation

Recent advances in large language models (LLMs) and multimodal generation have sparked growing interest in converting documents into diverse output formats, such as slides, posters, or audio summaries (Xu et al., 2025; Wang et al., 2025; Pang et al., 2025; Sun et al., 2024). Systems like PPTAgent (Zheng et al., 2025b) and Doc2PPT (Fu et al., 2022) treat document-to-slide generation as a structured summarization problem, focusing on layout-aware slide construction. Other works, such as Paper2Poster (Pang et al., 2025) extend this idea by producing single-page visual summaries using layout planning and visual feedback. However, these systems typically generate static outputs and do not model time-dependent delivery such as narration or slide progression. Our work builds upon these foundations, but further introduces temporal planning and audio-visual synchronization, enabling the generation of fully narrated presentation videos.

A.2 Vision-Language Agents

Recent advances have highlighted the expanding capabilities of vision language models (VLMs) beyond traditional language understanding. Techniques such as ReAct (Yao et al., 2023; Yang et al., 2023b; Yue et al., 2024) have shown that LLMs can operate as autonomous agents, capable of step-by-step reasoning and dynamic interaction through code execution (Wang et al., 2024c; Yang et al., 2024a,b), API function calls (Schick et al., 2023; Lu et al., 2025; Yang et al., 2023a), user interface manipulation (Lin et al., 2024b; Qin et al., 2025; Nayak et al., 2025; Wu et al., 2024), and motion generation (Zhang et al., 2024; Zhou et al., 2024; Wang et al., 2024d). Despite these developments, general-purpose agents still struggle with professional tasks that demand accuracy, domain-specific knowledge, and reliable interaction (Lin et al., 2024a). A closely related area is slide automation (Ge et al., 2025; Zheng et al., 2025a), which agents translate short text prompts into executable Python code to render presentation slides. In contrast, our proposed presentation video generation task is significantly more challenging: instead of taking a short prompt as input, the system processes an entire long-form document—such as a research paper, product manual, or technical report—and produces a well-structured presentation video with oral-style narration. This task imposes higher demands on content understanding, multimodal alignment, speech generation, and video synthesis. To address these challenges, we design a generation pipeline along with an automatic evaluation framework to systematically assess the generated videos in terms of information delivery, visual quality, and overall comprehensibility.

B Implementation Details

PresentAgent adopts a highly modular multimodal-generation architecture. At the language-understanding and generation layer, we run six primary LLM back ends in parallel—GPT-4o, GPT-4o-mini, Qwen-VL-Max, Gemini-2.5-Flash, Gemini-2.5-Pro, and Claude-3.7-Sonnet—and select or ensemble them on-the-fly with a dynamic routing policy that weighs input length, conversational complexity, and latency budget. For visual-language evaluation, we introduce the lightweight VLM Qwen-VL-2.5-3B-Instruct to score slide layout, chart readability, and cross-modal consistency, feeding its self-critique back into generation. Speech synthesis is unified on MegaTTS3, which outputs 24 kHz, 16-bit high-fidelity narration and supports prosody-tag controls for fine-grained rate, pitch, and emotion adjustment.

The experimental pipeline converts any input document—PDF, Markdown, DOCX, or web snapshot through three automated stages:

1. Structured parsing & re-ordering that maps content to a hierarchical topic–subtopic tree.

2. Per-slide generation with the chosen LLM, producing a PowerPoint deck containing titles, bullet points, graphic placeholders, and Alt-Text, while retrieving and inserting relevant images for key nouns.

3. Synchronized narration generation with MegaTTS3 in Chinese or English, followed by an FFmpeg script that assembles a 1080 p video with fade-in/out transitions and optional captions.

C Discussion

In this work, we synthesized presentation-style videos that integrate visual slides, textual narration, and spoken audio, simulating realistic multimodal communication scenarios. While our current evaluation focuses on the individual quality of each modality—such as visual clarity, textual relevance, and audio intelligibility—these dimensions are treated independently. However, in real-world

applications, the effectiveness of communication often hinges on the semantic and temporal coherence across modalities.

Future research should thus move beyond isolated assessments and aim toward fusion-aware understanding and evaluation. This entails not only modeling the interactions and alignment among image, audio, and text modalities, but also enabling the system to reason over their combined meaning. Existing models like ImageBind offer a unified embedding space for multiple modalities, but lack the capacity for high-level inference and semantic comprehension.

A promising direction lies in bridging representation alignment with multimodal reasoning, by integrating aligned modality encoders with powerful language models. This would allow the system to jointly perceive, interpret, and respond to complex multimodal inputs—such as explaining a visual concept based on both audio narration and visual cues, or identifying inconsistencies across modalities. Developing such reasoning-capable, fusion-aware models will be critical for advancing robust, coherent multimodal understanding in real-world applications.

D Limitations

Our work faces two key constraints: (1) Due to the high computational costs of commercial LLM/VLM APIs (e.g., GPT-4o and Gemini-2.5-Pro), evaluation was limited to five academic papers, potentially underrepresenting the document diversity shown in our benchmark (Figure 5); (2) PresentAgent currently generates static slides without dynamic animations/effects due to architectural constraints in video synthesis and trade-offs between generation speed and visual quality, as noted in ChronoMagic-Bench's temporal coherence studies. Future improvements could involve lightweight distillation models and physics-aware rendering engines.

E Evaluation Benchmark

As Shown in Figure 5, we showcase four of the representative document types in our benchmark: academic papers, web pages, technical blogs, and presentation slides. These documents cover a broad spectrum of real-world content domains, such as educational tutorials, research briefs, product manuals, scientific articles, news commentary, and business reports. Each document is paired with a manually authored presentation video, providing a diverse and realistic testbed for evaluating document-to-video generation systems in terms of multimodal coherence, content preservation, and presentation quality.

F Doc2Present Dataset Details

Data Source. We collect 30 high-quality video samples from public platforms, educational repositories, and professional presentation archives. Each video follows a structured narration format, combining slide-based visuals with synchronized voiceover. We manually align each video with its source document and ensure the following conditions are met: (1) the content structure of the video follows that of the document; (2) the visuals convey document information in a compact, structured form; and (3) the narration and slides are well-aligned temporally.

Data Statistics. The average document length is 3,000–8,000 words, while the corresponding videos range from 1 to 2 minutes and contain 5-10 slides. This setting highlights the core challenge of the task: transforming dense, domain-specific documents into effective and digestible multimodal presentations.

G PresentEval

G.1 Prompts of Objective Quiz Evaluation

Table 2 presents the prompting content for the evaluation method utilizing objective quiz-based assessment. Each set of questions included in this evaluation is crafted manually, with its creation firmly rooted in the actual content of the relevant documents. The formulation of these questions places a distinct emphasis on three key aspects: topic recognition, which involves the ability to accurately identify and grasp the central themes of the source material; structural understanding, referring to the comprehension of the organizational framework and logical arrangement of the document; and key argument identification, focusing on the capacity to pinpoint the core viewpoints and supporting arguments within the content. These carefully designed questions serve as a means to evaluate the extent to which the generated video successfully conveys the essential information, core ideas, and structural logic of the original source material, thereby assessing the effectiveness of the video in communicating the source content.

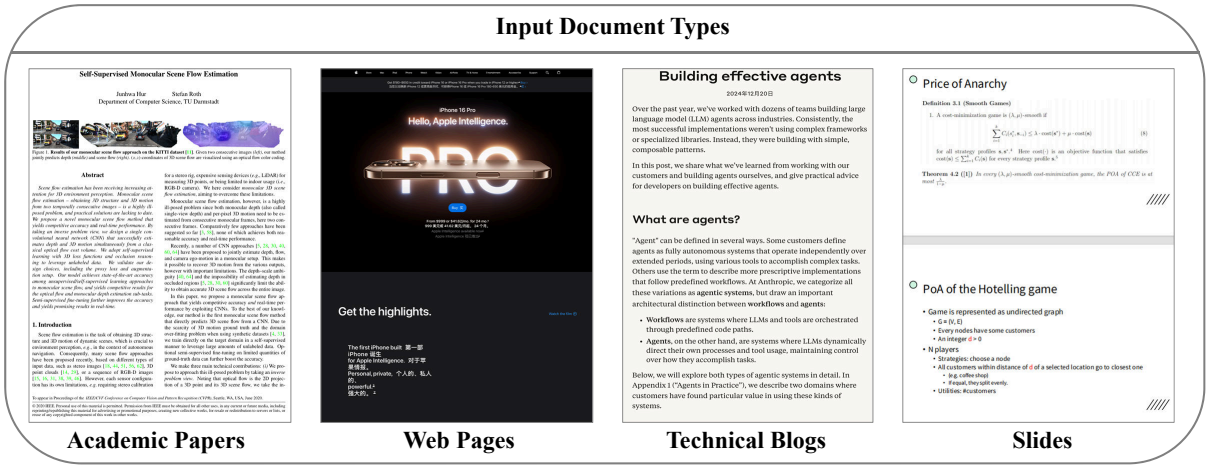

Figure 5: Document Diversity in Our Evaluation Benchmark.

Prensentation of Web Pages	What is the main feature highlighted in the iPhone's promotional webpage?
A.	A more powerful chip for faster performance
B.	A brighter and more vibrant display
C.	An upgraded camera system with better lenses
D.	A longer-lasting and more efficient battery
Prensentation of Academic Paper	What primary research gap did the authors aim to address by introducing the FineGym dataset?
A.	Lack of low-resolution sports footage for compression studies
B.	Need for fine-grained action understanding that goes beyond coarse categories
C.	Absence of synthetic data to replace human annotations
D.	Shortage of benchmarks for background context recognition

Table 2: Prompt of evaluation via Objective Quiz Evaluation. Each question set is manually created based on the actual document content, with a focus on topic recognition, structural understanding, and key argument identification. These questions evaluate how well the generated video communicates the source material.

G.2 Prompts of Subjective Scoring

Prompt of evaluation via subjective scoring is shown in table 3. This table showcases the prompting content employed in the subjective scoring-based evaluation approach. Each individual prompt within this set is precisely targeted at a specific evaluative dimension. These dimensions encompass narrative coherence, which pertains to the logical flow and consistency of the storytelling; visual appeal and audio appeal, focusing on the attractiveness and engaging nature of the visual elements and audio components respectively; and comprehension difficulty, referring to the level of ease or challenge in understanding the presented content. These prompts are meticulously designed to serve as a guiding framework for vision-language models, enabling them to assess presentations from a human-centric perspective. This means that the evaluation aligns with human perceptions, preferences, and ways of understanding, ensuring that the assessment results are more in line with how humans would judge the quality of the presentations.

H Evaluation Setup

We construct a test set consisting of 30 long-form documents, each paired with a manually created presentation video that serves as a human-level reference. These documents span a diverse range of topics, including education, product explanation, research overviews, and policy briefings. For each document, we generate a corresponding presentation video using our full generation pipeline.

All videos, both human-created and machine-generated, are evaluated using our unified evaluation framework, PresentEval. Each synthesized video is approximately two minutes in length. However, due to the current lack of a single multimodal model capable of jointly assessing visual and audio quality for videos longer than two minutes, we adopt a split evaluation strategy.

In the Objective Quiz stage, we use Qwen-VL-2.5-3B (Wang et al., 2024b) to evaluate the accuracy of the entire video using a fixed set of multiple-choice comprehension questions. In the Subjective Scoring stage, we extract short video/audio

Video	**Scoring Prompt**
Narr. Coh.	*"How coherent is the narration across the video? Are the ideas logically connected and easy to follow?"*
Visual Appeal	*"How would you rate the visual design of the slides in terms of layout, aesthetics, and overall quality?"*
Comp. Diff.	*"How easy is it to understand the presentation as a viewer? Were there any confusing or contradictory parts?"*
Audio	**Scoring Prompt**
Narr. Coh.	*"How coherent is the narration throughout the audio? Are the ideas logically structured and easy to follow?"*
Audio Appeal	*"How pleasant and engaging is the narrator's voice in terms of tone, pacing, and delivery?"*
Comp. Diff.	*"How easy is it to understand the spoken content? Were there any unclear or confusing parts in the audio?"*

Table 3: Prompt of evaluation via Subjective Scoring. Each prompt targets a specific dimension—narrative coherence, visual/audio appeal, or comprehension difficulty—and is designed to guide vision-language models in assessing presentations from a human-centric perspective. Abbreviations: Narr. Coh. = Narrative Coherence; Comp. Diff. = Comprehension Difficulty.

segments and evaluate them individually to assess quality in a more focused and scalable manner, using Qwen-Omni-7B (Xu et al., 2025).

Both models are guided by dimension-specific prompts and score each video or audio sample along three axes: Content Quality, Visual Quality, and Comprehension Accuracy.