Narrative-Driven Paper-to-Slide Generation via ArcDeck

Tarik Can Ozden*, Sachidanand VS*, Furkan Horoz*,
Ozgur Kara, Junho Kim†, and James M. Rehg†

University of Illinois Urbana-Champaign

Abstract. We introduce ArcDeck, a multi-agent framework that formulates paper-to-slide generation as a structured narrative reconstruction task. Unlike existing methods that directly summarize raw text into slides, ArcDeck explicitly models the source paper's logical flow. It first parses the input to construct a discourse tree and establish a global commitment document, ensuring the high-level intent is preserved. These structural priors then guide an iterative multi-agent refinement process, where specialized agents iteratively critique and revise the presentation outline before rendering the final visual layouts and designs. To evaluate our approach, we also introduce ArcBench, a newly curated benchmark of academic paper-slide pairs. Experimental results demonstrate that explicit discourse modeling, combined with role-specific agent coordination, significantly improves the narrative flow and logical coherence of the generated presentations. The ArcDeck codebase and ArcBench benchmark are available on our project webpage.

Keywords: Slide Generation · Agentic AI · Discourse Modeling

1 Introduction

An effective academic presentation conveys an academic paper's dense technical content in a concise, informative, and visually engaging manner while maintaining a clear narrative flow [34]. Crafting these presentations is a multi-faceted, complex process that requires deliberate decisions regarding content selection, structural organization, and visual design [4]. As a result of these complexities, the automatic generation of high-quality presentations (in the form of a slide deck) remains a significant challenge for modern AI systems.

The key challenge at the heart of slide generation is the need to identify the *narrative structure* of the source paper, which specifies how concepts are introduced, explained, and related to each other in order to tell the story of the work, and then map that structure into the visual form of a slide deck, while at the same time extracting and translating the concepts, results, graphics, and other elements that make up the paper content. Prior approaches to automatic paper-to-slide generation can be organized into three categories, as

*Equal contribution †Corresponding author

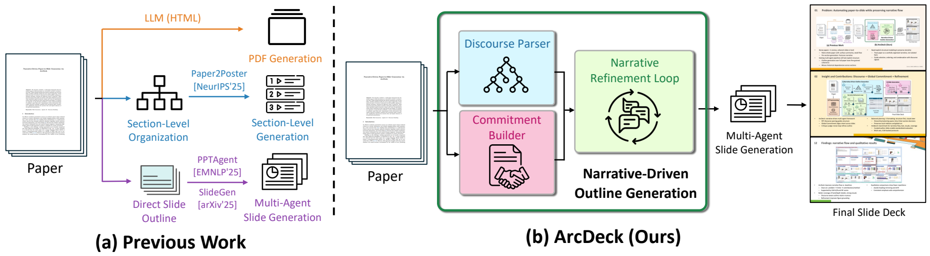

Fig. 1: Framework Comparison. (a) Prior methods rely on raw-text summarization, isolated section-level organization, or direct outline extraction, yielding condensed text without narrative guidance. (b) In contrast, *ArcDeck* explicitly models discourse structure and establishes global commitments to produce narrative-driven outlines through an iterative multi-agent refinement loop.

illustrated in Fig. 1(a), based on their approach to addressing this challenge. The most straightforward approach (orange arrow) is to feed the entire paper directly into a Large Language Model (LLM) to output a presentation-ready format (*e.g.*, HTML or XML) that can be converted into a PDF [3, 7, 36]. However, this approach struggles with long context windows— the model becomes overwhelmed by the volume of text, producing static, surface-level summaries that lack a cohesive flow. To address this concern, another category of works develop a section-level organization [30, 43] (blue arrow, Fig. 1(a)), processing each section independently before mapping them to pages and applying layout refinements [11, 44]. While this approach can identify the key organizing principles for the paper, it fundamentally fractures the narrative, as generating summaries in isolation creates temporally disconnected slide content that post-processing cannot repair. Recently, multi-agent frameworks [22, 39] have begun to address this limitation (purple arrow, Fig. 1(a)) by first generating a global paper-level slide outline consisting of titles and summarized paragraphs, which is then passed to a specialized multi-agent generation module. However, outline generation is itself a non-trivial task that requires hierarchical planning across multiple levels of abstraction, spanning both intra-section organization and overall paper structure. Moreover, because these models construct the outline by processing the entire paper at once, they again struggle to maintain fine-grained structural coherence in dense content. In summary, prior works have lacked an effective approach to structural modeling, which has limited their ability to preserve rhetorical dependencies and orchestrate the unified storyline required for a compelling presentation.

In this paper, we propose *ArcDeck*, a narrative-driven multi-agent framework that formulates paper-to-slide generation as a structured narrative reconstruction problem. *ArcDeck* addresses the structural and contextual shortcomings of prior approaches through three core mechanisms (Fig. 1(b)). First, to overcome the lack of explicit structural guidance, our *Discourse Parser* constructs a discourse tree inspired by Rhetorical Structure Theory (RST) [27]. By explicitly defining the relational connections between text units (*e.g.*, paragraphs), this hierarchical model guides initial structural planning and preserves local rhetorical relationships. Second, to prevent the overarching context loss typical of fragmented generation, a *Commitment Builder* establishes a *Global Commitment*. This module captures high-level intent and meta-information—such as the central thesis, scope, and overall narrative structure—and propagates them as persistent conditioning signals to ensure paper-level consistency. Third, recognizing that narrative coherence requires more than one-shot outlining, we incorporate a multi-agent *Narrative Refinement Loop*. Through a closed-loop critique–judge–revise process, the outline is iteratively refined to improve structural coherence, enabling a more effective translation from hierarchical planning to slide-level presentation. Together, these mechanisms allow *ArcDeck* to seamlessly balance content organization with visual formatting, producing a highly coherent narrative flow. An overview of *ArcDeck* framework is illustrated in Fig. 2

In addition, to support systematic evaluation, we introduce *ArcBench*, a curated paper–slide benchmark collected from multiple academic venues and spanning diverse research topics. *ArcBench* enables systematic comparison across multiple dimensions, including text coverage, narrative coherence, visual quality, and content retention, establishing a standardized evaluation protocol for paper-to-slide generation. We will open source *ArcDeck* and *ArcBench* to further research advances. To summarize, our main contributions are as follows:

- We propose *ArcDeck*, a narrative-oriented multi-agent framework for paper-to-slide generation that, for the first time, explicitly models discourse relations through hierarchical RST-based analysis.
- We introduce global commitments and a closed-loop critique–judge–revise refinement mechanism, enabling coherent structural planning and consistent narrative flow across the entire slide sequence. Ours is the first multi-agent solution and the first to support customization (*e.g.*, presentation length).
- We present *ArcBench*, a high-quality and open-source paper–slide paired dataset curated from author-prepared presentations at top-tier venues, providing the community with a reliable human reference standard for evaluating slide generation quality.

2 Related Work

Agentic Slide Generation. Early slide generation methods framed the task as content selection via regression models [15], while later deep learning approaches used RNNs to summarize sections into bullet points [9, 36]. The emergence of LLMs shifted the focus toward multi-agent and code-generation techniques [13, 22, 30, 39, 44, 45]. Beyond early zero-shot editing [13], frameworks like DocPres combined LLMs and VLMs for extraction and outlining [3]. Others improved organization using intermediate JSON format [7] or reframed the task as code synthesis using `python-pptx` scripts [11, 22, 30]. To enhance visual fidelity, models were fine-tuned on description-code pairs [11, 38] or utilized reference-based template editing [44]. Recent multi-agent pipelines iteratively refine generated code by coordinating specialized agents for content, layout, and figure alignment [22, 30, 39, 40, 42, 43, 45]. Some methods use a single agent to model narrative structure [22, 44], while others rely on section titles or predefined narrative templates, such as a Problem–Motivation–Results–Conclusion structure [30, 40]. In contrast to these prior works, we are the first to model source-paper discourse relations as a structured narrative, refine this narrative through a multi-agent loop, and construct a global commitment (see Fig. 2) that explicitly captures content-generation goals and supports customization to different presentation durations and target audience.

Rhetorical Structure Theory (RST). RST [27] models hierarchical discourse structure by defining rhetorical relations between text spans. It has been widely applied in downstream tasks such as summarization [23, 28], question answering [10], and sentiment analysis [5]. Existing RST parsers generally follow either top-down [19, 29] or bottom-up [8, 32] paradigms. As deep learning models gained prominence, end-to-end neural parsers have been proposed that learn discourse structures through representation learning [17, 20, 21, 41]. Other approaches formulate discourse parsing as a text generation problem, using LSTMs [6] or encoder–decoder architectures [14, 33] to generate discourse trees. More recent work [24, 26, 37] fine-tunes pretrained models for RST discourse tree generation. Other studies further leverage predicted discourse trees as structural guidance for text generation [1, 18]. Relative to this literature, our main contribution is demonstrating the value of RST-Parsing in slide generation for the first time. In addition, our approach to LLM-based RST-parsing differs from prior work in leveraging SOTA LLMs without finetuning and demonstrating their effectiveness. Using these RST trees as structural guidance, we enable the generation of highly coherent, narrative-driven content.

Paper-Slide Datasets. Prior work has introduced several paired datasets for paper-to-slide generation [9, 11, 22, 36, 44] (see Supp.). These datasets share several critical limitations. First, the datasets in [11, 44] provide broad domain coverage and lack a specific academic specialization. Given the difficulty of effective human-level slide generation, there is a need for a well-scoped dataset exposing the specific generation challenges of technical academic content such as dense equations, figures, and structured arguments. Second, prior publicly-available datasets [9, 11, 36, 44] were not curated for presentation quality, making it difficult to assess how effective slide-generation approaches are in achieving expert-level quality. Note that while the SlideGen [22] dataset also contains high-quality technical presentations, it has not been publicly-released. In contrast to these works, ArcBench is the result of curating 100 oral high-quality pairs from 994 paper-presentation pairs at top CV/ML conferences. We further validate our human reference slides by comparing them against generated baselines (see Tab. 4b), revealing a clear quality gap that motivates the benchmark.

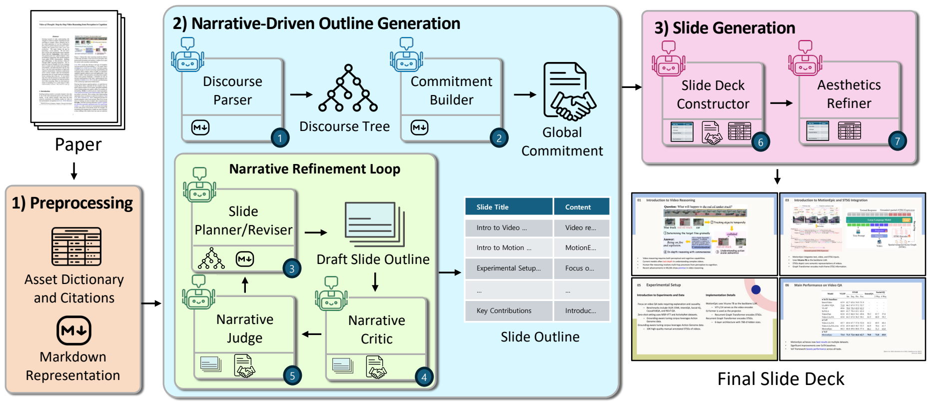

Fig. 2: ***ArcDeck* Overview.** 1) **Preprocessing** extracts textual and visual assets from the source paper; 2) **Narrative-Driven Outline Generation** combines discourse parsing and global commitment to guide the Narrative Refinement Loop, producing a structured Slide Outline; and 3) **Slide Generation** constructs the initial deck and refines its aesthetic layout to yield the final presentation. The numbered indices of each agent, *e.g.*, 1, denote their temporal execution order, while the icons at the bottom of each agent box indicate the specific structural inputs they receive.

3 ArcDeck: Narrative-Driven Slide Generation

3.1 Preprocessing

Preprocessing stage takes the raw PDF of the paper as input and distills it into structured textual and visual assets required for downstream generation. We utilize Docling [25] and Marker [31] to parse the paper, converting the text into a clean Markdown format while excluding the bibliography and appendix to focus on narrative content. Figures and tables are extracted into an asset dictionary, with each visual element stored alongside its caption and size information to provide layout context for the agents. In parallel, in-text citations are parsed to construct a mapping dictionary that links short-form citations to their full numbered references for slide generation.

3.2 Narrative-Driven Outline Generation

The Narrative-Driven Outline Generation stage translates the preprocessed text into a logically structured slide plan. To ensure the resulting slide is both structurally sound and thematically consistent, we orchestrate three core components that use the Markdown representation. First, the **Discourse Parser** 1 constructs a hierarchical discourse tree to explicitly model the rhetorical dependencies between text segments. Second, the **Commitment Builder** 2 establishes a Global Commitment to capture the overarching constraints and key messages of the target deck. Finally, guided by both these structural and thematic signals, the **Narrative Refinement Loop**—comprising the *Slide Planner/Reviser* 3, *Narrative Critic* 4, and *Narrative Judge* 5—iteratively drafts, evaluates,

Table 1: RST Relation Taxonomy. Definitions of the rhetorical relations used to construct our discourse trees (NS: Nucleus-Satellite; MN: Multinuclear).

Relation	Explanation	Type
Elaboration	Adds detail, examples, or implementation specifics.	NS
Explanation	Clarifies why or how a nucleus claim holds.	NS
Context	Background or definitions needed to interpret the nucleus.	NS
Purpose	Goal or motivation associated with the nucleus.	NS
Evaluation	Assessment, strength, or limitation of the nucleus.	NS
Organization	Roadmap or meta-structural framing.	NS
Joint	Parallel units at the same rhetorical level, equally central.	MN
Same-unit	Two EDUs forming one semantic unit split across boundaries.	MN

and polishes the slide outline. Together, this multi-agent process ensures the generated outline preserves the source paper's narrative flow and logical interdependencies before visual generation.

Discourse Parser 1. The Discourse Parser serves as a structural analysis agent that constructs a hierarchical RST discourse tree to explicitly model rhetorical dependencies between text segments, providing the guidance for content grouping and narrative ordering. Concretely, for each paper section, it generates a binary discourse tree by treating paragraphs as elementary discourse units (EDUs) that form the leaves. Inspired by the Rhetorical Structure Theory [27], the parser assigns relations between adjacent units as either (i) nucleus–satellite, where the nucleus carries the central claim and the satellite provides supporting detail, or (ii) multinuclear (symmetric), where both units are equally central (see Table 1 for the full relation taxonomy). At higher levels, the tree recursively groups EDUs based on the relations identified at lower levels, producing a hierarchy that encodes both local paragraph-level dependencies and the global narrative organization of the section: relations closer to the root span larger portions of the text and capture high-level rhetorical structure, while relations near the leaves reflect fine-grained narrative progression within subsections (see Fig. 3 for an example). The discourse trees are then serialized into JSON and passed to the Slide Planner for subsequent outline generation.

Commitment Builder 2. The Commitment Builder takes the Markdown data and user-provided 'target audience' along with 'presentation duration' as input, and generates a global commitment to establish an overall plan for slide generation. The generated commitment consists of key information describing the structure of the slide deck: a snapshot, core content that specifies the thesis and key takeaways, a talk contract that specifies the assumed prerequisites, a narrative spine, and an overall section plan (see Fig. 4). The global commitment serves as high-level guidance to the following Narrative Refinement Loop.

Narrative Refinement Loop 3, 4, 5. Once the Discourse Tree and Global Commitment are generated, the Narrative Refinement Loop iteratively refines

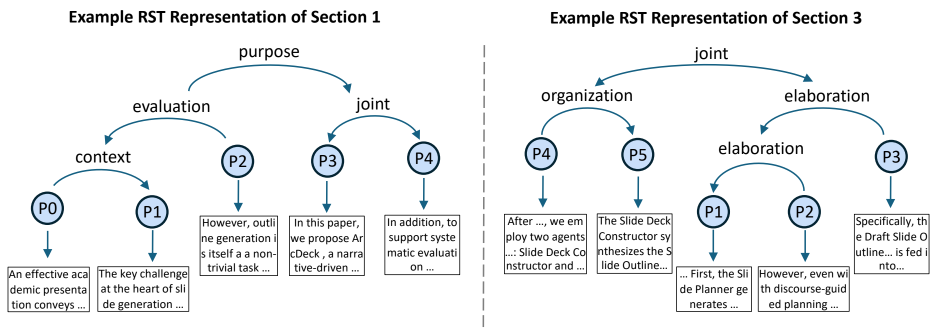

Fig. 3: Example Discourse Trees from Secs. 1 & 3 of this paper. Leaf nodes correspond to elementary discourse units (EDUs, corresponding to paragraphs), while internal nodes and arrows denote rhetorical relations that recursively group these spans. See Supp. for full examples.

the Markdown information to create a Slide Outline. First, the Slide Planner generates an initial slide outline guided by the structural signals from the Discourse Tree and the Global Commitment. For each section, it groups preprocessed paragraphs according to their rhetorical relations to determine which content should appear together, forming a logically structured narrative. The resulting Draft Slide Outline in JSON contains a section for each slide including a deck title, the grouped paragraphs for its content, and a rationale for the structural grouping decisions. However, even with discourse-guided planning, one-shot planning may fail to ensure alignment with the Global Commitment and a coherent narrative flow. We therefore introduce an iterative critique–judge–revise cycle to refine the outline. Within this loop, we assess the outline with the following criteria: (a) alignment with the Global Commitment, (b) global narrative flow, (c) section balance, (d) slide-level coherence, and (e) redundancy or missing content.

Specifically, the Draft Slide Outline generated by the Slide Planner is fed into the Narrative Critic together with the Global Commitment to obtain feedback, which is then passed to the Narrative Judge to determine whether the outline is ready or requires revision. The Judge then issues a decision, summarizes the rationale, identifies must-fix issues with severity levels (high, medium, low), and provides actionable guidance for the Reviser when necessary. If a revision is required, the Reviser updates the slide outline based on the combined feedback, after which it re-enters the evaluation cycle. This refinement process continues until the outline is *ready*, as determined by the Narrative Judge, or three refinement cycles have been completed, at which point the finalized outline is forwarded to the Slide Generation stage in JSON format.

3.3 Slide Generation

After preparing the narrative-driven Slide Outline, we employ two agents to generate an editable `.pptx` slide (Fig. 5): the **Slide Deck Constructor** 6 selects relevant figures and tables from the dictionary and assigns an appropriate template layout based on the visual elements, producing a Draft Slide Deck

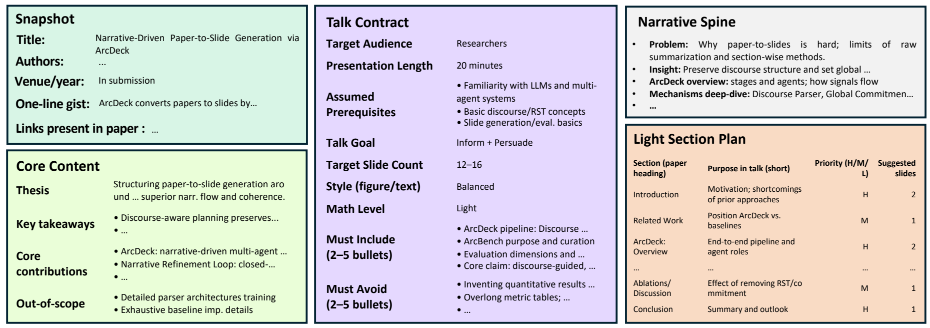

Fig. 4: Global Commitment Example. The global commitment establishes the high-level narrative progression of the presentation through five key components: a snapshot, core content, a talk contract, a narrative spine, and a light section plan. See Supp. for full examples.

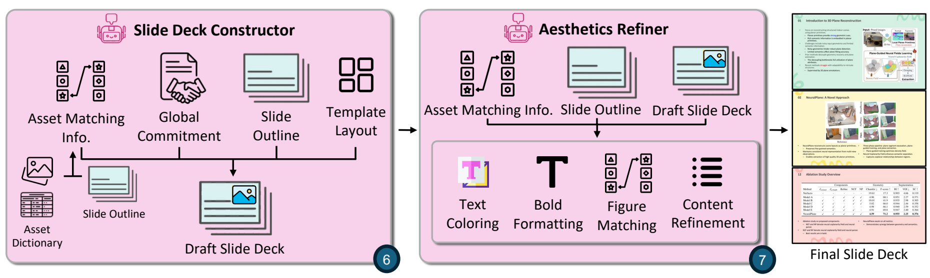

Fig. 5: Slide Generation Stage. Slide Deck Constructor 6 handles asset mapping, layout selection, and text generation, while the Aesthetics Refiner 7 polishes the deck via figure matching and content formatting.

conditioned on the Global Commitment, and the **Aesthetics Refiner** 7 refines the draft through content adjustment (*e.g.*, enriching sparse text), additional figure matching, and stylistic text formatting (details in Appendix).

Slide Deck Constructor 6. This agent synthesizes the Slide Outline, pre-processed visual assets, and the Global Commitment to assemble a Draft Slide Deck. For each slide, it first selects the most relevant figures and tables based on their extracted captions. It then assigns one of 14 available layout templates, conditioned on the text volume and the number of visual elements, its size and aspect ratio. Guided by the RST-derived outline and Global Commitment, the agent generates the slide text while explicitly highlighting key topics to preserve the intended narrative flow. Furthermore, it adapts the textual organization, toggling between bulleted hierarchies and paragraph narration based on information density, and notes down the short-form citation mentioned in the text, which will be later added as footnotes. The resulting draft deck is output as a structured JSON object (see Supp.) containing global metadata alongside slide-wise titles, content, matched visuals, and references.

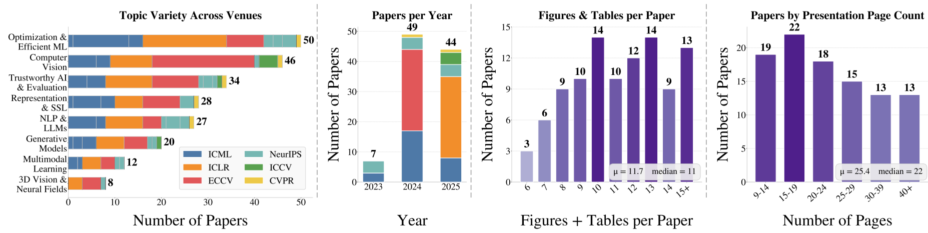

Fig. 6: ArcBench Statistics. Overview of our curated 100-pair dataset, showing the distribution of research topics across six major CV & ML venues, publication years, visual density, and author-prepared presentation lengths.

Aesthetics Refiner 7. The Aesthetics Refiner utilizes the Asset Matching Info., Slide Outline, and Draft Slide Deck to perform the final polish of the presentation. It executes four targeted operations: (i) **Figure Matching** conducts an additional review to insert relevant visual elements into slides lacking sufficient visual grounding; (ii) **Content Refinement** adjusts textual density by enriching sparse slides, condensing heavy text, or merging underutilized layouts to maintain a clear narrative pacing; (iii) **Text Coloring** dynamically applies a consistent theme color derived from the most frequently occurring colors in the deck's figures; and (iv) **Bold Formatting** adds targeted emphasis to key terminology to enhance overall readability. The resulting refined JSON is then rendered into the Final Slide Deck using the `python-pptx` library.

3.4 ArcBench Dataset

To establish a rigorous evaluation protocol for the paper-to-slide generation task, we introduce *ArcBench*, a newly curated paper-slide paired dataset. We extract source papers and their corresponding presentation materials directly from official venues, resulting in an initial 994 candidate pairs from six major computer vision and machine learning conferences (ICCV, CVPR, ECCV, ICML, ICLR, and NeurIPS) between 2022 and 2025. Because large-scale evaluation across multiple metrics and baselines is costly, running the full suite on all 994 pairs is prohibitive. Accordingly, we apply a post-processing step to curate a high-quality subset of 100 pairs for all subsequent experiments (Fig. 6).

To balance evaluation costs with the topical diversity, we select this 100-pair dataset through three strict filtering stages: we restrict the selection to *oral presentations* to ensure deliberate narrative design, and we require at least 3 figures and 3 tables per article to guarantee sufficient visual and quantitative material for evaluating content fidelity. A key novelty of *ArcBench* is paper-"author-prepared" (AP) presentation PDF pairs, providing a robust human reference standard to accurately measure how closely generated outputs approximate human-level preparation. While this initial dataset is limited to AI venues, we leave the expansion to other scientific disciplines for future work, and we release the curated 100-pair dataset of paper and slide links, along with a download script, to facilitate further research. Full details are available in the Supp.

Table 2: Summary of Evaluation Metrics.

Category	Specific Metric	Input	Output	High-Level Idea
VLM-based Q/A Quiz	Story	Pres. Text & 25 MCQs	Accuracy (%)	Storyline capture and high-level information retention.
	Visuals	Slide Images & 25 MCQs		Information conveyed through visual elements and slide layout.
	Hard	Pres. Text & 25 MCQs		Conceptual understanding of methodology and limitations.
	Depth	Pres. Text & 25 MCQs		Coverage of important technical/implementation details.
VLM-as-Judge	Text Quality (TQ)	Pres. Text	1–100 Score	Technical substance and specificity of the text.
	Narrative Flow (NF)			Logical progression and comprehensive section coverage.
	Visual Layout (VL)	Slide Images		Spatial geometry, balance, and arrangement of elements.
	Visual Thematic (VT)			Thematic consistency, design coherence, and visual elements.
Automated Metrics	ROUGE-L	Pres. Text & Source Paper	Overlap Score	Textual sequence overlap and content coverage.
	Perplexity	Pres. Text	Perplexity Score	Linguistic fluency and text coherence.
VLM Pairwise Preference	Narrative Flow	Two Decks & Source Paper	Win Rate (%)	Opening quality, narrative arc, and progressive revelation.
	Author-Prepared	Generated Deck & AP Deck		Overall presentation quality relative to human-authored slides.

4 Experiments

4.1 Experimental Setup

Baselines & Benchmark. We compare against four systems spanning three paradigms. **(i)** *Prompt-based generation:* **HTML** produces slides via a single LLM call that generates HTML+CSS, directly converted to PDF. **(ii)** *Multi-agent pipelines:* **Paper2Poster** [30] (P2Poster), a visual-in-the-loop multi-agent system adapted for slide generation; **PPTAgent** [44], a two-stage edit-based slide generation approach; **SlideGen** [22], a six-agent VLM framework with template-driven layout composition; and **(iii)** *Human reference:* the original *author-prepared* slides. To separate pipeline design from LLM capability, we run *ArcDeck* and all baselines with three generation backbones: GPT-4o [16], GPT-5 [35], and Qwen3-VL-32B [2]. All methods are evaluated on the 100 paper–slide pairs in *ArcBench*, ensuring identical source material and evaluation protocol.

Slide Generation Format. All generated slides follow a 13.33×7.5-inch layout. Our method uses a reference template with 14 distinct layout configurations covering common scenarios with multiple images and tables. For fairness, all methods are evaluated under fixed slide themes so that preference decisions reflect content and structure rather than stylistic variation.

4.2 Evaluation Setup

We evaluate *ArcDeck* along four axes: (i) *content coverage* (preservation of key information from the source paper), (ii) *narrative coherence* (logical slide progression), (iii) *textual quality* (fluency and technical specificity), and (iv) *visual quality* (clarity of layout and visual elements). Metrics are summarized in Tab. 2.

VLM-based Q/A Quiz. To measure how much of the source paper's content is preserved in the generated slides, we design a two-stage quiz protocol.

First, a VLM reads the source paper and generates 25 multiple-choice questions across four categories: *Story* (narrative arc), *Visuals* (information from figures and diagrams), *Hard* (deep conceptual understanding of methodology), and *Depth* (fine-grained technical details). Second, a separate VLM attempts to answer these questions using only the generated slides as input, with text-based categories answered from extracted slide text and the Visuals category answered from slide images. The resulting accuracy, reported in Tab. 3, reflects how effectively the slides communicate the paper's content at each level of granularity.

VLM-as-Judge. Rather than relying on generic metrics, we employ VLM judges with carefully designed rubrics to score each presentation on a 1-100 scale across four dimensions: *Text Quality* (technical substance and specificity), *Narrative Flow* (logical progression and section coverage), *Visual Layout* (spatial arrangement and readability), and *Visual Thematic* (design coherence and appropriate use of visual elements). Each dimension uses a 10-item checklist, scored by the number of satisfied criteria. Results are reported in Tab. 3.

Automated Metrics. We complement the VLM-based evaluations with two standard text metrics reported in Tab. 6: *ROUGE-L*, measuring sequence overlap between the generated slide text and the source paper to quantify content coverage, and *Perplexity* (computed via LLaMA-3-8B [12]), assessing linguistic fluency of the slide text.

VLM Pairwise Preference (A/B Test). We use two pairwise comparison protocols. For *baseline vs. ours*, a VLM judge receives two slide decks and the source paper, and selects which slide better preserves the source paper's logical structure and narrative arc, as reported in Tab. 4a. For *baseline vs. AP*, the judge instead evaluates holistic presentation quality, jointly considering content organization, visual design, information delivery, and slide structure, to assess how closely the generated slides approximate author-prepared presentations, as reported in Tab. 4b. We use both closed- (GPT-5) and open-source (Qwen3-VL) judges to mitigate single-model bias. Quiz and A/B evaluations are repeated 11 times with randomized order, using average scores or majority vote.

4.3 Analysis and Discussion

Narrative Flow. A central design goal of *ArcDeck* is to produce presentations with coherent, human-like narrative progression. As shown in Tab. 4a, *ArcDeck* consistently outperforms all baselines in pairwise comparisons, achieving the highest overall win rate across both open- and closed-source judges. The advantage is most pronounced against PPTAgent and HTML, whose outlines largely mirror source document headings without rhetorical restructuring. SlideGen presents a narrower gap, yet our method still maintains a consistent edge, which we attribute to the discourse-aware outline generation and the Narrative Refinement Loop. These findings are corroborated by the VLM-as-Judge NF scores in Tab. 3: across all generation backbones and judge models, *ArcDeck* achieves the highest or second-highest NF scores.

Table 3: VLM-Based Evaluation. The best value in each generation model is highlighted in green, and runner-up is underlined.

Method	VLM-based Q/A Quiz (Open) ↑ Story	Visuals	Hard	Depth	VLM-based Q/A Quiz (Closed) ↑ Story	Visuals	Hard	Depth	VLM-as-Judge (Open) ↑ TQ	NF	VL	VT	VLM-as-Judge (Closed) ↑ TQ	NF	VL	VT
Generation Model: GPT-5																
HTML	95.16	64.13	81.83	69.73	95.60	38.40	80.00	82.20	86.11	89.06	95.22	76.77	58.50	61.50	63.67	10.00
P2Poster	95.22	65.51	76.70	67.67	87.00	27.00	66.40	64.60	62.20	70.55	97.08	91.35	35.17	41.67	72.00	78.50
PPTAgent	92.96	65.08	69.49	65.67	75.00	12.80	32.00	36.20	35.95	55.94	84.86	81.95	16.50	37.67	71.33	57.50
SlideGen	95.26	68.37	78.39	69.43	97.60	35.80	86.00	82.40	69.02	88.18	97.36	95.28	36.17	56.67	88.33	84.83
ArcDeck	95.41	65.95	81.13	71.58	96.00	41.00	85.60	83.40	78.38	91.39	99.40	96.06	46.50	63.83	85.50	83.67
Generation Model: GPT-4o																
HTML	93.34	57.36	64.57	60.57	51.36	6.64	11.84	14.16	16.72	19.51	54.47	12.89	7.25	14.00	24.00	0.25
P2Poster	94.12	63.11	66.07	60.18	60.80	8.16	17.60	20.24	32.32	40.71	92.05	86.23	11.12	28.12	60.25	62.38
PPTAgent	93.35	61.90	66.01	60.84	59.04	7.52	15.12	18.56	16.43	25.75	82.59	71.49	12.62	26.75	59.12	34.75
SlideGen	93.75	62.79	66.15	61.00	60.00	8.08	17.84	21.36	19.28	41.40	93.30	89.40	9.62	30.75	84.75	78.25
ArcDeck	95.46	64.57	72.63	67.56	82.40	18.16	40.08	47.20	51.68	62.78	96.05	92.94	24.38	43.25	91.00	81.25
Generation Model: Qwen-3-VL-32B-Instruct																
HTML	94.07	61.53	75.24	66.68	85.67	36.65	63.26	61.77	67.55	71.02	82.05	66.74	48.62	51.88	64.00	24.38
P2Poster	94.28	63.36	75.96	66.75	86.14	21.49	51.26	54.51	65.96	74.24	95.90	89.60	36.50	49.50	72.25	75.88
PPTAgent	92.66	61.94	69.53	65.38	67.91	12.00	22.14	25.12	41.55	66.12	80.44	45.25	19.25	35.25	47.25	11.25
SlideGen	94.81	62.12	76.57	67.96	90.42	28.47	64.93	63.07	69.18	86.40	94.16	92.18	42.50	55.25	85.25	79.50
ArcDeck	95.69	66.58	83.22	71.39	93.12	34.79	68.28	66.70	80.06	89.00	98.31	94.68	53.62	63.00	87.25	79.50

Table 4: VLM Pairwise Preference (A/B Test). Holistic evaluation of generated presentations using two comparison protocols: assessing narrative flow against baselines (left), and overall quality against Author-Prepared (AP) slides (right). To mitigate single-model bias, both closed- (GPT-5) and open-source (Qwen3-VL) judges are used.

(a) Preference for Narrative Flow. Win rate (%) of **Ours** vs. baselines. Green indicates our method is preferred ($> 50\%$), red otherwise.

Generation Model	GPT-5		GPT-4o		Qwen3	
Evaluation Model	Qwen3	GPT-5	Qwen3	GPT-5	Qwen3	GPT-5
Ours vs. P2Poster	62.4	97.6	78.0	84.0	92.8	90.7
Ours vs. PPTAgent	100.0	100.0	99.0	95.0	100.0	97.9
Ours vs. SlideGen	55.3	50.6	81.0	68.0	57.7	48.5
Ours vs. HTML	82.4	84.7	79.0	92.0	76.3	84.5

(b) Preference vs. Author-Prepared (AP). Higher values indicate closer quality to AP slides. Green highlights the best per column.

Generation Model	GPT-5		GPT-4o		Qwen3	
Evaluation Model	Qwen3	GPT-5	Qwen3	GPT-5	Qwen3	GPT-5
P2Poster vs. AP	35.1	33.3	27.5	26.7	37.7	40.0
PPTAgent vs. AP	45.6	40.0	35.4	30.0	45.4	36.7
SlideGen vs. AP	**46.8**	45.8	32.1	33.3	46.0	45.8
HTML vs. AP	35.9	33.3	31.3	33.3	42.6	40.0
Ours vs. AP	35.7	**48.1**	**36.6**	**40.0**	**47.3**	**46.7**

Coverage of Content. We evaluate content coverage through quiz-based metrics as summarized in Tab. 3. *ArcDeck* achieves the highest or second-highest scores across all categories. On **Story**, our method leads across all three models, indicating that the discourse-driven outline better preserves the paper's high-level narrative arc. The largest gains appear on **Hard** and **Depth**, which reflect conceptual understanding and fine-grained detail. Under Qwen-3-VL, *ArcDeck* exceeds the next-best baseline by 6.65 on Hard and 3.43 on Depth, with similar margins under GPT-4o. We attribute this to the hierarchical discourse tree, which preserves elaboration and explanation relations and retains methodological nuances during condensation. On **Visuals**, which tests whether visually grounded questions can be answered from the slides alone, *ArcDeck* achieves the highest scores under GPT-4o and Qwen-3-VL and remains competitive under GPT-5, indicating strong visual coverage.

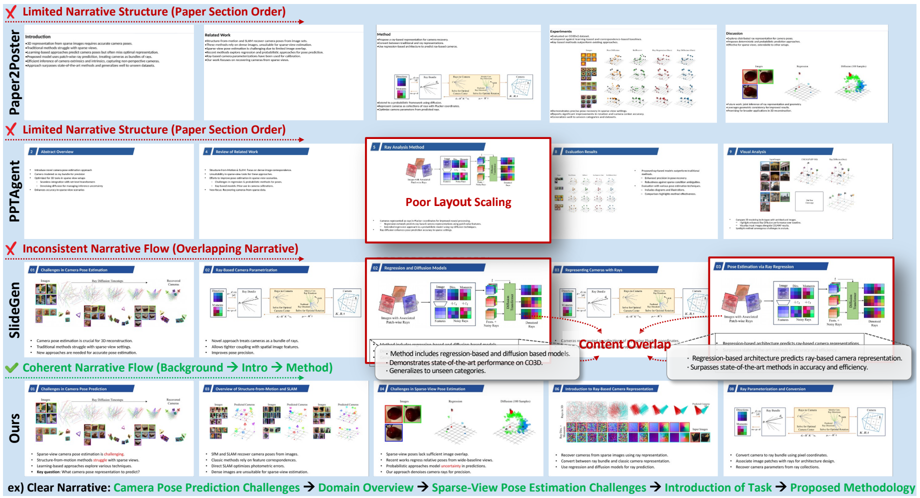

Fig. 7: Qualitative Comparison. Unlike baselines, *ArcDeck* shows clear narrative flow instead of mirroring paper section structure and avoids cross-slide content overlap.

Table 5: Outline Generation Ablation Study. Narrative Flow (NF) scores and ArcDeck win rate against ablated versions.

Method	NF	Win %
w/o Discourse Parser	7.50	89.1
w/o Global Comm.	7.52	94.5
w/o Narr. Ref. Loop	8.68	61.8
ArcDeck	9.70	–

Table 6: Automated Metrics Evaluation. Green indicates the best value in each column, and underline is runner-up. R-L values are scaled by 10^{-3}.

Generation Model	GPT-5		GPT-4o		Qwen3	
Automated Metrics	R-L ↑	PPL ↓	R-L ↑	PPL ↓	R-L ↑	PPL ↓
HTML	<u>64.7</u>	39.90	33.6	<u>31.51</u>	<u>108.5</u>	31.08
Paper2Poster	71.3	21.04	<u>53.8</u>	22.07	76.8	<u>16.96</u>
PPTAgent	33.6	91.43	34.6	57.21	64.6	23.33
SlideGen	55.1	80.46	45.2	43.84	83.9	23.09
ArcDeck	57.1	<u>46.01</u>	84.8	33.47	156.0	16.95

Textual & Visual Quality. Finally, we evaluate textual quality using automated overlap and fluency metrics (Tab. 6). Across generation backbones, *ArcDeck* achieves strong ROUGE-L (R-L) scores. It attains the highest ROUGE-L under GPT-4o and Qwen-3-VL, while remaining competitive under GPT-5. Perplexity (PPL) results indicate fluent slide text across models. Minor metric differences likely reflect the nature of slide writing, which favors concise bullet-style summaries over direct sentence reuse.

Qualitative examples under a fixed slide theme are shown in Fig. 7. Paper2Poster and PPTAgent mostly follow the original paper's section order, resulting in limited narrative arc (row 1-2). Paper2Poster also produces visually plain slides due to the lack of template support, while PPTAgent struggles to adapt layouts for varying figure sizes (col. 3). SlideGen achieves strong visual design but exhibits weaker narrative flow and content overlap (row 3). In contrast, *ArcDeck* produces a clearer narrative flow (row 4: from problem context to prior work and finally to the proposed method). Additional examples are in the Supp.

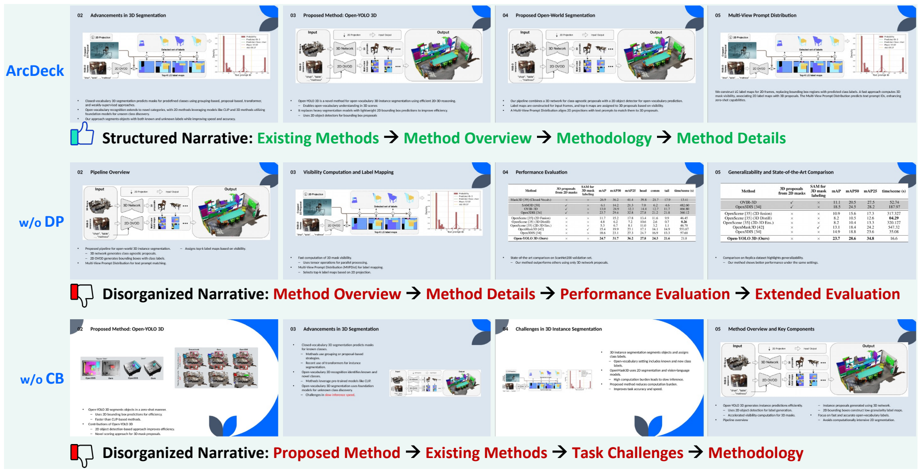

Fig. 8: Qualitative Ablation Study. Comparison between *ArcDeck* and variants without the Discourse Parser (DP) or Commitment Builder (CB).

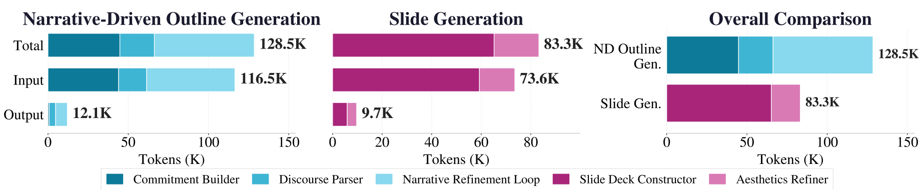

Fig. 9: Token Usage Analysis.

Discussion of Author-Prepared Slides. A key question in evaluating paper-to-slide generation is: *how close are automated methods to human-prepared presentations?* While pairwise A/B tests in Tab. 4a reveal relative strengths among systems, it cannot assess whether any method has reached a practically useful level of quality. To address this, we conduct holistic pairwise A/B comparisons against the original author-prepared slides, as reported in Tab. 4b, jointly evaluating content organization, visual design, information delivery, and slide structure. Among all methods, *ArcDeck* achieves the highest average selection rate against Author-Prepared slides, confirming that methods producing better slides among themselves also compete more effectively with human reference slides.

Ablation Study & Token Analysis. To demonstrate the effectiveness of our narrative-driven outline generation, we conduct qualitative and quantitative ablation studies (Fig. 8, Tab. 5). Without Discourse Parser, the slide becomes condensed and less structured, omitting key method details due to the lack of discourse-guided content grouping. Without Commitment Builder, slide ordering becomes disorganized, introducing the proposed method before prior work and task context. In contrast, the full pipeline produces a more coherent narrative progression—from existing methods to method overview and detailed methodology. Quantitatively (Tab. 5), removing each component individually lowers the

Narrative Flow, and ArcDeck consistently outperforms the ablated variants in pairwise comparisons. We further report token usage across agents in Fig. 9. Given the multi-agent design of *ArcDeck*, this breakdown provides how tokens are distributed across agents during outline generation and slide generation.

Human Evaluation. We conduct a human evaluation to assess the quality of generated slide decks in terms of content coverage and narrative flow. A total of 25 undergrad or MS/PhD students participated in the study. To ensure reliable evaluation of technical content, participants were asked to select the research area they were most familiar with among three options: Vision-Language/Multimodal Learning (V-L/MM), Generative Models, and Computer Vision Core. For each selected area, participants were presented with slide decks generated by ArcDeck, SlideGen, and PPTAgent for five oral papers, and were asked to rank them based on overall content quality and narrative coherence. ArcDeck achieves the best average ranking across the evaluated papers, indicating that the generated slide decks provide the strongest narrative flow and overall content coherence among the compared baseline frameworks, as shown in Tab. 7. The web interface of the experiment is shown in the Supp.

Table 7: Human Evaluation Results. Participants ranked slide decks based on content quality and narrative flow, where Rank 1 = 3 points, Rank 2 = 2 points, and Rank 3 = 1 point. The best value in each topic is highlighted in green, and runner-up is underlined.

Method	Human Evaluation Topic			
	V-L / MM	**Gen. Models**	**CV Core**	**Overall**
PPTAgent	1.03	1.13	1.20	1.10
SlideGen	<u>2.42</u>	<u>2.23</u>	<u>2.14</u>	<u>2.30</u>
ArcDeck	2.55	2.63	2.66	2.60

5 Conclusion

In this work, we introduce *ArcDeck*, a narrative-driven framework for automatic paper-to-slide generation that formulates the task as a structured narrative reconstruction problem. By explicitly modeling discourse structure through RST-based parsing, establishing a global commitment to preserve the paper's high-level intent, and refining outlines through a multi-agent critique–judge–revise loop, *ArcDeck* generates slide decks with improved narrative coherence and structural alignment. To support systematic evaluation, we also present *ArcBench*, a curated paper–slide paired dataset designed to assess slide generation across multiple dimensions, including content, narrative coherence, and visual quality. Experimental results demonstrate that explicit discourse modeling substantially improves both narrative structure and information retention, while remaining competitive in textual and visual quality.

Acknowledgements

Portions of this work were supported in part by the Health Care Engineering Systems Center in the UIUC Grainger College of Engineering, and by a grant from the UIUC Institute for Growth.

References

1. Adewoyin, R., Dutta, R., He, Y.: Rstgen: Imbuing fine-grained interpretable control into long-formtext generators. In: Proceedings of the 2022 conference of the north American chapter of the association for computational linguistics: human language technologies. pp. 1822–1835 (2022)
2. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., Zhu, K.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
3. Bandyopadhyay, S., Maheshwari, H., Natarajan, A., Saxena, A.: Enhancing presentation slide generation by llms with a multi-staged end-to-end approach. In: Proceedings of the 17th International Natural Language Generation Conference. pp. 222–229 (2024)
4. Bartsch, R.A., Cobern, K.M.: Effectiveness of powerpoint presentations in lectures. Computers & education **41**(1), 77–86 (2003)
5. Bhatia, P., Ji, Y., Eisenstein, J.: Better document-level sentiment analysis from rst discourse parsing. In: Proceedings of the 2015 conference on empirical methods in natural language processing. pp. 2212–2218 (2015)
6. Braud, C., Plank, B., Søgaard, A.: Multi-view and multi-task training of rst discourse parsers. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. pp. 1903–1913 (2016)
7. Cachola, I., Cucerzan, S., Herring, A., Mijovic, V., Oveson, E., Jauhar, S.K.: Knowledge-centric templatic views of documents. In: Findings of the Association for Computational Linguistics: EMNLP 2024. pp. 15460–15476 (2024)
8. Feng, V.W., Hirst, G.: Text-level discourse parsing with rich linguistic features. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 60–68 (2012)
9. Fu, T.J., Wang, W.Y., McDuff, D., Song, Y.: Doc2ppt: Automatic presentation slides generation from scientific documents. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 634–642 (2022)
10. Gao, Y., Wu, C.S., Li, J., Joty, S., Hoi, S.C., Xiong, C., King, I., Lyu, M.: Discern: Discourse-aware entailment reasoning network for conversational machine reading. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 2439–2449 (2020)
11. Ge, J., Wang, Z.Z., Zhou, X., Peng, Y.H., Subramanian, S., Tan, Q., Sap, M., Suhr, A., Fried, D., Neubig, G., et al.: Autopresent: Designing structured visuals from scratch. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 2902–2911 (2025)
12. Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al.: The llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024)
13. Guo, Y., Zhang, Z., Liang, Y., Zhao, D., Duan, N.: Pptc benchmark: Evaluating large language models for powerpoint task completion. In: Findings of the Association for Computational Linguistics: ACL 2024. pp. 8682–8701 (2024)

14. Hu, X., Wan, X.: Rst discourse parsing as text-to-text generation. IEEE/ACM Transactions on Audio, Speech, and Language Processing **31**, 3278–3289 (2023)
15. Hu, Y., Wan, X.: Ppsgen: Learning-based presentation slides generation for academic papers. IEEE transactions on knowledge and data engineering **27**(4), 1085–1097 (2014)
16. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
17. Ji, Y., Eisenstein, J.: Representation learning for text-level discourse parsing. In: Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 1: Long papers). pp. 13–24 (2014)
18. Kim, Z.M., Ramachandran, A., Tavazoee, F., Kim, J.K., Rokhlenko, O., Kang, D.: Align to structure: Aligning large language models with structural information. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 37519–37528 (2026)
19. Kobayashi, N., Hirao, T., Kamigaito, H., Okumura, M., Nagata, M.: Top-down rst parsing utilizing granularity levels in documents. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 8099–8106 (2020)
20. Kobayashi, N., Hirao, T., Kamigaito, H., Okumura, M., Nagata, M.: A simple and strong baseline for end-to-end neural rst-style discourse parsing. In: Findings of the Association for Computational Linguistics: EMNLP 2022. pp. 6725–6737 (2022)
21. Li, J., Li, R., Hovy, E.: Recursive deep models for discourse parsing. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 2061–2069 (2014)
22. Liang, X., Zhang, X., Xu, Y., Sun, S., You, C.: Slidegen: Collaborative multimodal agents for scientific slide generation. arXiv preprint arXiv:2512.04529 (2025)
23. Liu, Z., Chen, N.: Exploiting discourse-level segmentation for extractive summarization. In: Proceedings of the 2nd Workshop on New Frontiers in Summarization. pp. 116–121 (2019)
24. Liu, Z., Shi, K., Chen, N.: Dmrst: A joint framework for document-level multilingual rst discourse segmentation and parsing. In: Proceedings of the 2nd Workshop on Computational Approaches to Discourse. pp. 154–164 (2021)
25. Livathinos, N., Auer, C., Lysak, M., Nassar, A., Dolfi, M., Vagenas, P., Ramis, C.B., Omenetti, M., Dinkla, K., Kim, Y., et al.: Docling: An efficient open-source toolkit for ai-driven document conversion. In: AAAI Conference on Artificial Intelligence (2025)
26. Maekawa, A., Hirao, T., Kamigaito, H., Okumura, M.: Can we obtain significant success in rst discourse parsing by using large language models? In: Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 2803–2815 (2024)
27. Mann, W.C., Thompson, S.A.: Rhetorical structure theory: A theory of text organization. Tech. rep., University of Southern California, Information Sciences Institute Los Angeles (1987)
28. Marcu, D.: The theory and practice of discourse parsing and summarization. MIT press (2000)
29. Nguyen, T.T., Nguyen, X.P., Joty, S., Li, X.: Rst parsing from scratch. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 1613–1625 (2021)
30. Pang, W., Lin, K.Q., Jian, X., He, X., Torr, P.: Paper2poster: Towards multimodal poster automation from scientific papers. Advances in Neural Information Processing Systems **38** (2026)

31. Paruchuri, V.: Marker: Convert pdf to markdown + json quickly with high accuracy (2023), `https://github.com/datalab-to/marker`, accessed: 2026-02-12
32. Prendinger, H., et al.: A novel discourse parser based on support vector machine classification. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. pp. 665–673 (2009)
33. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of machine learning research **21**(140), 1–67 (2020)
34. Reynolds, G.: Presentation Zen: Simple ideas on presentation design and delivery. New Riders (2019)
35. Singh, A., Fry, A., Perelman, A., Tart, A., Ganesh, A., El-Kishky, A., McLaughlin, A., Low, A., Ostrow, A., Ananthram, A., et al.: Openai gpt-5 system card. arXiv preprint arXiv:2601.03267 (2025)
36. Sun, E., Hou, Y., Wang, D., Zhang, Y., Wang, N.X.: 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. pp. 1405–1418 (2021)
37. Thompson, K., Chaturvedi, A., Hunter, J., Asher, N.: Llamipa: An incremental discourse parser. In: Findings of the Association for Computational Linguistics: EMNLP 2024. pp. 6418–6430 (2024)
38. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
39. Xu, X., Xu, X., Chen, S., Chen, H., Zhang, F., Chen, Y.C.: Pregenie: An agentic framework for high-quality visual presentation generation. arXiv preprint arXiv:2505.21660 (2025)
40. Yang, Y., Jiang, W., Wang, Y., Song, Y., Wang, Y., Zhang, C.: Auto-slides: An interactive multi-agent system for creating and customizing research presentations. arXiv preprint arXiv:2509.11062 (2025)
41. Yu, N., Zhang, M., Fu, G.: Transition-based neural rst parsing with implicit syntax features. In: Proceedings of the 27th International Conference on Computational Linguistics. pp. 559–570 (2018)
42. Zeng, W., Ouyang, M., Cui, L., Ng, H.T.: Slidetailor: Personalized presentation slide generation for scientific papers. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 34584–34592 (2026)
43. Zhang, Z., Zhang, X., Wei, J., Xu, Y., You, C.: Postergen: Aesthetic-aware paper-to-poster generation via multi-agent llms. arXiv preprint arXiv:2508.17188 (2025)
44. Zheng, H., Guan, X., Kong, H., Zhang, W., Zheng, J., Zhou, W., Lin, H., Lu, Y., Han, X., Sun, L.: Pptagent: Generating and evaluating presentations beyond text-to-slides. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. pp. 14413–14429 (2025)
45. Zhu, Z., Lin, K.Q., Shou, M.Z.: Paper2video: Automatic video generation from scientific papers. arXiv preprint arXiv:2510.05096 (2025)