GRAPHTALK: SLIDE-ANCHORED TOPIC GRAPHS WITH SUPPORTING DOCUMENT EVIDENCE FOR PRESENTATION NARRATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Generating spoken narration (a *talk track*) for an existing slide deck requires following the slides' topical structure while grounding each slide in evidence from long supporting documents. We present GRAPHTALK, a training-free pipeline built around an explicit, interpretable *structured planning layer* that replaces the flat, high-volume source context used by prior approaches. GRAPHTALK induces atomic topics from the slides, enriches them with retrieved document evidence, ranks them over a similarity graph with a PageRank-seeded variant of DivRank, and allocates them to slides by capacity-constrained matching in a shared embedding space. On two benchmarks (M3AV academic lectures and SEC EDGAR financial filings), two generation models, and a dual-ladder of deterministic and LLM-as-judge metrics, its central result is the strongest *document-grounded slide coverage* and *relevance*, holding across deterministic metrics and three LLM judges, at a token budget comparable to the strongest baselines. A small human study supports the coverage advantage. Automatic judges are unreliable for *supporting-document usage*, where our human study is more trustworthy and also favours GRAPHTALK. We validate judge reliability with multi-judge agreement, an image-grounded judge, and perturbation tests.

1 INTRODUCTION

Large language models (LLMs) (Brown et al., 2020; Achiam et al., 2023) have made automated content generation a high-impact frontier for both academic NLP and industry. One active direction is the automatic generation of presentation slide decks from source documents (Fu et al., 2022; Sun et al., 2021), driven by the need to distil research papers, earnings reports, and enterprise documentation into visual summaries. A fully autonomous presentation pipeline, however, demands a second capability that has received comparatively little attention: generating a high-quality *spoken narration* (a *talk track*) for every slide, so that a deck can be delivered asynchronously without a human presenter. We study exactly this problem: given an ordered slide deck and a corpus of supporting documents, produce a fluent, document-grounded per-slide narration.

Why narrate *existing* decks. We deliberately target the setting where the slides already exist and are fixed. A very large installed base of decks (enterprise and training material, investor and earnings presentations, archived talks) needs narration added after the fact, for accessibility, asynchronous video, e-learning, and auto-narrated briefings. This is distinct from *joint* slide-and-narration generation (for example PPTAgent (Zheng et al., 2025) and Paper2Video (Zhu et al., 2025)), which assumes the slides do not yet exist and can be co-designed with the script. Our setting instead treats each slide as a fixed, external artifact and produces narration faithful to *whatever* already appears on it, sourced from possibly-unaligned supporting documents. The output is a textual narration script (for display or as text-to-speech input), not synthesised audio.

Challenges. A talk track must satisfy two granularities: at the *local* (per-slide) level each narration should primarily address the slide's visible content without a verbatim read-off, enriching it with document evidence (a precise figure from a quarterly report; a detail from a cited paper), while a minority of slides (section dividers, background) are intentionally addressed only briefly; at the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

global (deck-wide) level narrations must cohere into a unified discourse (Barzilay & Lapata, 2008) with smooth transitions, diverse coverage, and equitable use of multiple documents. This is hard along four axes. First, the input is massive and heterogeneous: a deck may pair 25 slides with hundreds of pages of filings, which LLMs underutilise in long contexts (Liu et al., 2024). Second, the output is long and must stay globally consistent: a narration can exceed 3,000 words yet must avoid the surface repetition generators are prone to (Holtzman et al., 2020). Third, faithfulness under compression is fragile, and it invites hallucinated figures and claims (Maynez et al., 2020; Ji et al., 2023). Fourth, supervision is scarce, since paired datasets do not exist at scale, which favours training-free designs.

GRAPHTALK. We address these with GRAPHTALK, a training-free pipeline whose key component is an explicit, interpretable *structured planning layer* that the LLM consumes instead of raw source content (Figure 1). GRAPHTALK extracts atomic topics with initial evidence from the slides, enriches each with targeted evidence retrieved from the documents, fuses redundancies via embedding clustering, and ranks topics by global importance using a PageRank-seeded variant of DivRank (Mei et al., 2010) over a topic-similarity graph. A globally-constrained bipartite matching then assigns the ranked, evidence-tagged topics to slides, so each slide is narrated from a curated, source-attributed context. Because the pipeline needs no task-specific training data, it applies to any domain where a presentation is accompanied by supporting documents, and the structured planning layer is transparent enough to inspect at each stage.

Contributions and scope of claims. We are careful to claim only what the evidence robustly supports. **(1)** We propose GRAPHTALK (§3), a training-free structured planning layer for document-grounded slide narration combining atomic-topic induction, DivRank importance ranking, and globally-constrained slide–topic allocation in a shared embedding space. **(2)** We benchmark on two complementary datasets (M3AV academic lectures (Chen et al., 2024) and SEC EDGAR financial disclosures (Loukas et al., 2021; Chen et al., 2021)), spanning one- and two-document grounding and both reference-based and reference-free settings, under a dual-ladder metric suite. **(3)** Our headline empirical result is that GRAPHTALK delivers the strongest *document-grounded coverage* and *relevance*, a finding that is stable across deterministic metrics and three LLM judges, at a token budget on par with or below the strongest baselines; a human study supports the coverage advantage. Because automatic judges are less reliable for *supporting-document usage* (the content volume is large and the judge must check that every relevant document claim is used), we support that dimension with our human study, where GRAPHTALK also leads. **(4)** We audit judge reliability with multi-judge agreement, an image-grounded judge, and perturbation tests, isolating the dimensions that carry our headline claim.

2 RELATED WORK

Long-form generation from long context. Multi-document summarization is the canonical long-form task over heterogeneous sources; benchmarks such as Multi-News (Fabbri et al., 2019), pre-trained models (Xiao et al., 2022), and topic-guided variants (Cui & Hu, 2021) inject structure before generation. Yet LLMs underutilise evidence in the middle of very long inputs (Liu et al., 2024), motivating pipelines that pre-select evidence, with retrieval-augmented generation (Lewis et al., 2020) being the canonical instance and one of our baselines. GRAPHTALK instead pre-selects and ranks evidence through its structured planning layer, delivering curated, source-attributed context to the LLM.

Document-to-presentation generation, and narration from structured content. Automatic slide-deck generation from documents has attracted growing interest. D2S (Sun et al., 2021) uses slide titles as queries to retrieve and summarise paper passages; DOC2PPT (Fu et al., 2022) scales to hierarchical generation with joint layout prediction; more recent LLM-driven systems add structure, such as persona-aware slide discourse (Mondal et al., 2024), multi-staged end-to-end slide generation (Bandyopadhyay et al., 2024), and graph-neural architectures that model non-linear discourse and attribute each sentence to a source passage (Maheshwari et al., 2024). These target *slide-content* generation. The closest analogue to our task expands compact structured content (titles, outlines, attributed passages) into fluent prose; GRAPHTALK likewise uses structured intermediate representations and source attribution, but targets *spoken narration* for a fixed deck, which adds stronger

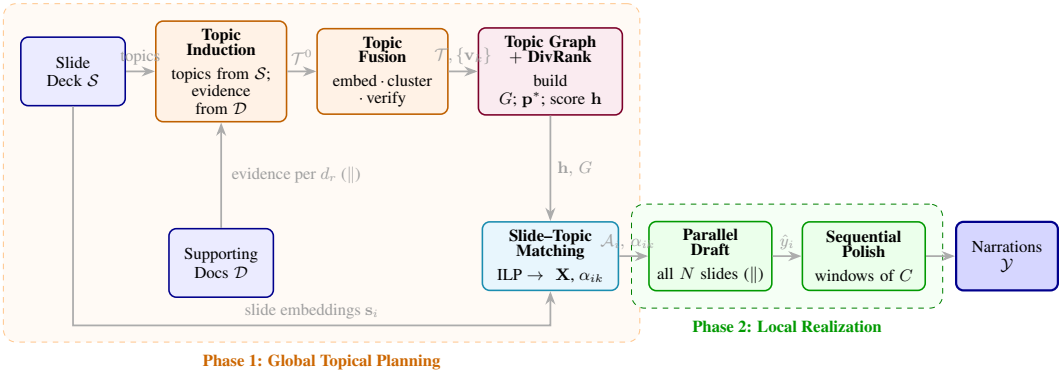


Figure 1: **Overview of the GRAPHTALK pipeline.** *Phase 1:* topics are induced from slides $\mathcal{S}$ and enriched with evidence from $\mathcal{D}$, fused, ranked via DivRank, and allocated to slides by a capacity-constrained ILP. *Phase 2:* narrations are drafted in parallel and refined in sequential windows for deck-wide coherence.

faithfulness requirements (each narration must cite verifiable facts) and explicit cross-slide coherence.

Presentation-style video generation. Generating complete presentation videos subsumes narration production. PPTAgent (Zheng et al., 2025) extracts patterns from reference decks and generates slide content and speaker notes; Paper2Video (Zhu et al., 2025) converts papers into videos with per-slide scripts from direct GPT-4 prompting. Both, and direct-prompting baselines broadly, treat narration as a single-step prompt without explicit topic planning, global diversity constraints, or multi-document grounding. GRAPHTALK instead adds a structured planning layer that performs global topic ranking and capacity-constrained allocation across the deck *before* any narration is produced.

3 THE GRAPHTALK FRAMEWORK

GRAPHTALK processes a slide deck $\mathcal{S}$ and a corpus of supporting documents $\mathcal{D}$ in two phases (Figure 1). *Global topical planning* builds a structured evidence layer: atomic topics are extracted from the slides and enriched with evidence from the documents, ranked over a topic-similarity graph, and allocated to slides. *Local slide realization* generates each narration grounded on the curated, evidence-tagged topics assigned to that slide.

3.1 PROBLEM FORMULATION

Let $\mathcal{S} = (s_1, \ldots, s_N)$ be an ordered slide deck, where each slide s_i is represented as $x_i = [u_i\,;\, v_i]$, a concatenation of its text content u_i and a natural-language visual description v_i obtained from a vision-language model. The supporting corpus is $\mathcal{D} = \{d_1, \ldots, d_R\}$, where each document d_r carries a short label ℓ_r. GRAPHTALK introduces a latent topical layer: a set of K atomic topics $\mathcal{T} = \{t_1, \ldots, t_K\}$, each carrying labeled evidence strings drawn from both the slides and the documents. A DivRank importance vector $\mathbf{h}$ and a binary assignment $\mathbf{X} \in \{0,1\}^{N\times K}$ determine which topics pair with which slides; $\mathcal{A}_i = \{t_k \mid X_{ik} = 1\}$ is the topic set for slide s_i. The narration for each slide is generated as $y_i \sim p_\theta(y_i \mid x_i, \{(t_k, \alpha_{ik}, \mathcal{E}_k)\}_{t_k\in\mathcal{A}_i}, y_{<i}, \mathrm{Prev}_{>i})$, where α_{ik} is the slide-local importance weight of topic t_k (§3.5), $y_{<i}$ are prior narrations, and $\mathrm{Prev}_{>i}$ previews upcoming slide content. Notation is summarised in Table 6 (Appendix A).

3.2 ATOMIC TOPIC INDUCTION AND EVIDENCE ENRICHMENT

An *atomic topic* is the smallest self-contained semantic unit that can be coherently narrated (a specific fact, claim, or concept). An LLM processes the full deck $\mathcal{S}$ and produces atomic topics, each with a title and initial evidence drawn near-verbatim from slide content, under two hard constraints,

non-overlap (each slide passage is evidence for at most one topic) and **full coverage** (the evidence union accounts for all substantive content), which together encourage a near-partition that pushes the LLM toward specific, citable material rather than generic summaries.

The initial topic set $\mathcal{T}^0$ is then *enriched* with evidence from $\mathcal{D}$: for each topic t_j and document d_r, relevant passages are appended as evidence strings tagged with ℓ_r. This introduces factual depth (figures, named entities, comparative claims) that slide text alone may omit; the evidence set $\mathcal{E}_k$ pools sources from both the presentation and the documents.

3.3 Topic Fusion

Overlapping slides may yield semantically redundant topics that compete spuriously downstream. GraphTalk resolves this in two stages.

Hierarchical topic embedding. Each topic is embedded by pooling its title and evidence in a shared ℓ_2-normalized space, $\mathbf{v}(t) = \ell_2\text{-norm}\big(\alpha\,\mathbf{e}(\text{title}(t)) + (1-\alpha)\,\frac{1}{L}\sum_{j=1}^{L}\mathbf{e}(e_j)\big)$, where $\mathbf{e}(\cdot) \in \mathbb{R}^d$ is a fixed pre-trained encoder and $\alpha = 0.5$ weights title and evidence equally. Critically, slides are embedded with the same formula (§3.5), placing topics and slides in a *common* vector space and enabling direct cosine comparisons between them.

Topic clustering and LLM verification. Topics with $\mathbf{v}(t_a)^\top \mathbf{v}(t_b) \geq \tau_{\text{merge}}$ are merge candidates; connected components are processed in parallel, and for each a second LLM call decides whether all members describe the same concept, fusing confirmed ones into a single topic with a unified title and combined, source-tagged evidence. This coarse-embedding-then-LLM-verification guards against false positives from thresholding alone, yielding $\mathcal{T} = \{t_1, \ldots, t_K\}$ ($K \leq |\mathcal{T}^0|$), each re-embedded to $\mathbf{v}_k$.

3.4 Topic Graph Construction and PageRank-Seeded DivRank

Topic graph. GraphTalk builds a weighted undirected graph $G = (\mathcal{T}, \mathcal{E}_G, \mathbf{W})$ with edge weights from pairwise cosine similarity: $(t_k, t_\ell) \in \mathcal{E}_G \Leftrightarrow \mathbf{v}_k^\top \mathbf{v}_\ell \geq \tau_g$ and $W_{k\ell} = \mathbf{v}_k^\top \mathbf{v}_\ell$. Isolated topics are connected to nearest neighbours so all topics participate in ranking.

PageRank prior. Let $P^{(0)}_{k\ell} = W_{k\ell}/\sum_j W_{kj}$ be the row-stochastic transition matrix. PageRank defines $\mathbf{p}^*$ as the stationary distribution of a damped walk, $p^*_\ell = (1-d)\frac{1}{K} + d\sum_k P^{(0)}_{k\ell}\,p^*_k$ for all ℓ, with damping $d \in (0,1)$; high-similarity edges accumulate higher p^*.

DivRank. DivRank (Mei et al., 2010) replaces the fixed $P^{(0)}$ with a score-modulated transition recomputed each iteration, $Q^{(t)}_{k\ell} = P^{(0)}_{k\ell}\,h^{(t-1)}_\ell \big/ \sum_j P^{(0)}_{kj}\,h^{(t-1)}_j$, and the score update, initialized at $\mathbf{h}^{(0)} = \mathbf{p}^*$, is $h^{(t)}_\ell \propto (1-\delta)\,p^*_\ell + \delta\sum_k Q^{(t)}_{k\ell}$, followed by ℓ_1-normalization. Destinations with higher current scores attract disproportionately more probability mass, a reinforcement dynamic that differentiates topic importance beyond fixed-transition PageRank, while the personalization term $(1-\delta)\,p^*_\ell$ anchors the distribution to graph structure.

Entropy-based early stopping. Unconstrained DivRank can collapse mass onto a few hubs, degrading matching. We monitor the normalized Shannon entropy $\mathcal{H}(\mathbf{h}^{(t)}) = -\frac{1}{\log K}\sum_k h^{(t)}_k \log h^{(t)}_k$ and halt once $\mathcal{H}(\mathbf{h}^{(t)}) < \eta$, returning $\mathbf{h}^{(t-1)}$. The final scores $\mathbf{h}$ encode global topic importance.

3.5 Globally Constrained Slide-Topic Matching

Shared embedding space. To compare topics and slides directly, each slide s_i is cast into topic form (its text as the title and its extracted keypoints as evidence) and embedded with the same pooler (Eq. 1), yielding $\mathbf{s}_i$; the similarity is $\sigma_{ik} = \mathbf{s}_i^\top \mathbf{v}_k$.

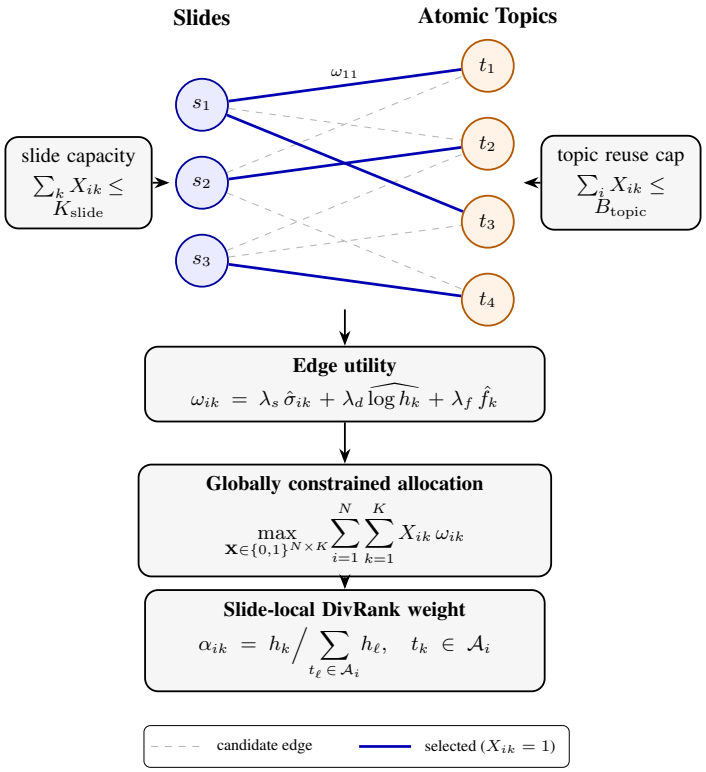


Figure 2: **Bipartite slide–topic allocation.** Dashed edges are candidates; bold edges are selected (X_{ik}=1). Edge utility ω_{ik} combines cosine similarity $\hat{\sigma}_{ik}$, log-DivRank importance $\widehat{\log h_k}$, and fact density $\hat{f}_k$. The ILP maximises $\sum X_{ik}\omega_{ik}$ subject to per-slide (K_{slide}) and per-topic (B_{topic}) caps, solved as min-cost max-flow.

Edge utility. For each candidate pair (s_i, t_k), an edge utility ω_{ik} combines three normalized signals: local similarity, global DivRank importance, and topic fact density. Full definitions and weights are in Appendix C.

Capacity-constrained assignment. The assignment $\mathbf{X}$ solves

$$\max_{\mathbf{X} \in \{0,1\}^{N \times K}} \sum_{i=1}^{N} \sum_{k=1}^{K} X_{ik}\, \omega_{ik} \quad \text{s.t.} \quad \sum_{k} X_{ik} \le K_{\text{slide}}\ \forall i, \quad \sum_{i} X_{ik} \le B_{\text{topic}}\ \forall k. \tag{2}$$

K_{slide} prevents overloading a slide; B_{topic} prevents one topic from dominating the deck. This integer program is equivalent to min-cost max-flow and solved exactly.

Slide-local topic weighting. Once assignment is fixed, topics in $\mathcal{A}_i$ are re-weighted by DivRank, projected to a slide-local simplex, $\alpha_{ik} = h_k / \sum_{t_\ell \in \mathcal{A}_i} h_\ell$ for $t_k \in \mathcal{A}_i$. This decouples *which* topics a slide receives (global matching) from *how prominently* each is weighted (DivRank).

3.6 Talk-Track Generation

Two-tier grounded generation, parallel draft, sequential polish. Each narration conditions on two tiers, the slide as a mandatory anchor (every visible element should be addressed) and the assigned topics as a grounding tier with evidence labelled by document provenance, and is instructed to incorporate any verifiable fact in a topic's evidence relevant to the slide, preventing generic narration when concrete evidence is available. All N drafts are generated concurrently; a subsequent sequential polish pass over non-overlapping windows of C slides, each conditioned on previously finalized narrations, then smooths transitions and removes redundancy. The full pipeline is Algorithm 1 (Appendix B); prompts are in Appendix C.

Table 1: M3AV LLM-as-judge metrics (1–5, higher better; primary judge Llama-3.3-70B; n=151, except Faith./SDUsage on $n\approx$104–105). Cells show the mean; standard errors are reported in the text where cited. Scores come from one comparative judge pass, so absolute values are pass-specific and not directly comparable across tables. **Bold**=best, underlined=second-best per column within each generator block. ‡Ablation.

		Cov.	Rel.	Div.	SDUsage	Faith.	Trans.	ReadAv.	NoLeak.	DecIm.	ChartEx.
Gemma-3-27B	Direct	3.39	3.72	3.87	3.01	4.01	**3.73**	3.41	3.60	3.46	4.19
	Summ	3.73	3.74	**3.90**	**3.14**	**4.04**	3.61	3.61	3.66	3.65	4.28
	RAG	3.68	3.89	3.83	3.13	3.97	3.53	3.69	3.74	3.71	4.37
	GraphTalk	**4.59**	**4.26**	3.88	3.03	4.01	3.23	**4.33**	**4.34**	**4.20**	**4.69**
	GT-SD‡	3.53	3.55	3.55	3.07	3.81	3.20	3.36	3.34	3.31	4.18
Pixtral-12B	Direct	2.89	3.28	3.70	**3.03**	**4.20**	3.50	2.89	3.10	2.91	4.15
	Summ	3.68	3.66	3.83	2.96	3.98	3.44	3.57	3.72	3.55	4.39
	RAG	3.91	3.93	**3.85**	2.93	3.89	**3.51**	3.88	3.92	3.80	4.52
	GraphTalk	**4.29**	**4.12**	3.75	2.86	3.94	3.18	**4.06**	**4.11**	**3.99**	**4.65**
	GT-SD‡	3.37	3.49	3.61	2.87	3.88	3.21	3.28	3.35	3.22	4.32

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets. **M3AV** (Chen et al., 2024) is an academic-lecture corpus of 151 presentations (≈2,114 slides) with ground-truth narrations transcribed from recordings, enabling reference-based and judge-based evaluation. **SEC EDGAR** (Loukas et al., 2021; Chen et al., 2021) is a financial-disclosure set of 45 earnings presentations (22 companies, 8 GICS sectors, 1,137 slides), each paired with a 10-K and a 10-Q report, the challenging two-document grounding condition, with no ground-truth narration so evaluation there is judge-only.[1]

Baselines and parity. All baselines use the *same* generation models, slide text and image inputs, and global polish pass, so differences isolate the planning layer. **Direct** concatenates the presentation with the full text of all documents and generates in one parallel pass; **Summ** first summarises each document and then generates as in Direct; **RAG** chunks every document, retrieves the top chunks per slide by semantic similarity, and generates per slide. We additionally evaluate a family of GRAPHTALK ablations (§4.5).

Implementation. Generation uses Gemma-3-27B and Pixtral-12B (local vLLM (Kwon et al., 2023)); the primary judge is Llama-3.3-70B-Instruct (AWQ-INT4) (Zheng et al., 2023), with GPT-4o and GPT-4.1 added in §4.4. Per-metric statistics use only samples for which the metric is available; all hyperparameters follow Table 7, with no dataset-specific tuning.

4.2 EVALUATION FRAMEWORK

We adopt a dual-ladder framework pairing deterministic reference-based metrics with LLM-as-judge assessments. **Deterministic (M3AV):** ROUGE-L (Lin, 2004) and METEOR (Banerjee & Lavie, 2005) for lexical overlap with ground truth; Keyword Recall (KW-Rec) for domain-term recall against sources; Numeric Fact F1 (FactF1) for numeric-fact F1 against sources; Self-BLEU (Holtzman et al., 2020) and Distinct-2 for deck-wide diversity; Adjacent-slide cosine (AdjCos) for local continuity. Because per-slide M3AV narrations are short (typically 1–3 sentences), ROUGE-L is applied at the granularity it is normally validated at; we treat it as one rung of a deliberately multi-metric ladder rather than a standalone proxy for long-form quality. **LLM-as-judge:** ten dimensions scored 1–5 (higher better): Slide Coverage (Cov.), Relevance (Rel.), Diversity (Div.), Supporting-Document Usage (SDUsage), Source Faithfulness (Faith.), Transition Quality (Trans.), Readoff Avoidance (ReadAv.), No Leakage (NoLeak.), Decorative Image Handling (DecIm.), Chart Explanation (ChartEx.). Definitions are in Appendix E.

[1] Both datasets were cleaned; all reported sizes reflect counts after filtering low-quality samples.

Table 2: SEC EDGAR LLM-as-judge metrics (1–5; Pixtral-12B; primary judge Llama-3.3-70B). Cells show the mean; standard errors are reported in the text where cited. Scores come from one comparative judge pass, so absolute values are pass-specific and not directly comparable across tables. **Bold**=best, underlined=second-best per column. ‡Ablation.

Method	Cov.	Rel.	Div.	Trans.	ReadAv.	NoLeak.	DecIm.	ChartEx.
Direct	2.00	2.52	3.17	2.81	2.09	2.30	2.11	3.07
Summ	3.64	3.60	3.70	3.33	3.58	3.67	3.77	4.00
RAG	4.02	3.93	**4.05**	**3.63**	3.90	3.94	**4.10**	4.34
GraphTalk	**4.27**	**4.28**	3.60	3.33	**3.98**	**4.05**	4.02	**4.46**
GT-SD‡	3.46	3.69	3.59	3.42	3.49	3.46	3.50	3.99

Table 3: M3AV deterministic metrics (n=151). Cells show the mean; standard errors are reported in the text where cited. **Bold**=best, underlined=second-best per column within each generator block (SelfBLEU↓: lower is better). ‡Ablation (§4.5).

		Reference-based				Diversity / Fluency		
Gen	Method	ROUGE-L	METEOR	KW-Rec	FactF1	SelfBLEU↓	Dist-2	AdjCos
Gemma-3-27B	Direct	**0.18**	0.26	0.61	0.10	**0.15**	0.85	0.67
	Summ	0.17	0.26	0.62	0.15	0.18	0.83	**0.68**
	RAG	**0.18**	**0.27**	0.66	0.13	**0.15**	0.84	0.66
	GraphTalk	0.15	**0.27**	**0.72**	**0.22**	0.18	0.81	0.66
	GT-SD‡	0.17	0.26	0.67	0.19	**0.15**	**0.86**	0.60
Pixtral-12B	Direct	**0.20**	0.26	0.55	0.10	0.17	0.84	0.63
	Summ	0.18	0.26	0.60	0.14	0.24	0.77	**0.66**
	RAG	0.18	**0.27**	0.68	0.14	0.18	0.81	0.64
	GraphTalk	0.16	**0.27**	**0.73**	**0.17**	0.20	0.77	0.62
	GT-SD‡	0.18	0.25	0.62	0.13	**0.15**	**0.85**	0.56

4.3 Main Results

Tables 1 and 2 report the main LLM-as-judge results, and Table 3 the full M3AV deterministic metrics.

Coverage and relevance: the robust win. GraphTalk achieves the highest Slide Coverage and Relevance across both generators on M3AV (Cov. 4.59 / 4.29; Rel. 4.26 / 4.12), the only method to lead both, with a +0.86 Coverage margin over the next-best baseline (Summ) on Gemma-3-27B. On two-document SEC EDGAR it again leads Coverage (4.27 ± 0.15) and Relevance (4.28 ± 0.13), while Direct collapses on Coverage (2.00 ± 0.08), indicating that global topic allocation matters most when evidence spans multiple long documents. The human study (§4.7) independently confirms it is preferred on Coverage and, on the dimension where automatic judges are unreliable, on Supporting-Document Usage.

Deterministic metrics. The deterministic ladder (Table 3) reinforces the grounding result: GraphTalk attains the best Keyword Recall (0.72 / 0.73) and Numeric-Fact F1 (0.22 / 0.17) on both generators, so its narrations recover more of the domain terms and exact source figures than any baseline. METEOR matches the strongest baseline (0.27 on both generators), and the remaining lexical-overlap and diversity columns stay close across methods, as expected for the short per-slide reference narrations these metrics score against.

Is the gain from the planner or from more tokens? We add two controls on SEC EDGAR (n=45 per generator; Table 11). GT-LLMPlan keeps the full pipeline but lets the generator LLM assign topics to slides in place of DivRank and the ILP. It falls below GraphTalk on Coverage (3.63 vs 4.03 on Pixtral, 4.13 vs 4.34 on Gemma) and on Relevance, so the structured planner drives the gain, not the LLM. RAG-TokenMatched runs the RAG baseline at a larger retrieval budget, so its prompt token count matches GraphTalk (ratio 1.01). Even with matched tokens it trails

Table 4: Inter-judge Spearman ρ on SEC EDGAR (45 decks, both generators). Agreement is strong on Coverage/Relevance, the dimensions carrying our headline claim, and lower on the more saturated dimensions.

Judge pair	Cov.	Rel.	Faith.	SDUsage	Div.
Llama vs GPT-4o	0.83	0.79	0.10	0.16	0.20
Llama vs GPT-4.1	0.82	0.80	0.08	0.13	0.13
GPT-4o vs GPT-4.1	0.92	0.85	0.27	0.18	0.12

GRAPHTALK on Coverage (3.45 vs 4.03 on Pixtral, 3.35 vs 4.34 on Gemma), so the gain does not come from a larger token budget. Paired tests confirm both results. Pooled over both datasets and both generators, the Coverage and Relevance gains over both controls are significant (Wilcoxon signed-rank, all $p < 10^{-7}$; Table 12).

4.4 JUDGE RELIABILITY

Because SEC EDGAR has no ground truth, the LLM judge is the sole signal there; we therefore audit it directly rather than assume its validity.

Multi-judge agreement. We re-scored the SEC EDGAR generations with two additional judges (GPT-4o, GPT-4.1) under the identical protocol and measured per-dimension Spearman agreement against the primary Llama-3.3-70B judge (Table 4; per-judge method scores in Table 13). Agreement is strong on *Coverage* (0.82–0.92) and *Relevance* (0.79–0.85), the dimensions carrying our headline claim, and the ranking is preserved: GRAPHTALK is #1 on Coverage under every judge for both generators, and #1 on Relevance in 5 of 6 judge$\times$generator cells. Agreement is lower on the more saturated dimensions (Faithfulness, SDUsage, Diversity, all $\leq$0.28), so we anchor our headline claims on Coverage and Relevance. Supporting-Document Usage is the clearest case: the large content volume and the need to check every relevant document claim make it hard to score automatically, so we rely on the human study (§4.7) for that dimension.

Image-grounded judge. Addressing the concern that a text-only judge sees a slide *description* rather than the slide, a GPT-4o judge shown the actual rendered slide image on a subset (8 decks/generator; Appendix Table 9) preserves the ranking: top methods stay on top and Direct stays last.

Perturbation sensitivity. Injecting cascading corruption (0–50%) into narrations, the judge score falls near-monotonically on the two dimensions most exposed to hallucination (mean Spearman ρ: Source Faithfulness 0.971, Supporting-Document Usage 0.970). Details are in Appendix I.

4.5 COMPONENT ABLATION

We isolate every component of the planning layer on the full M3AV set (n=151; Appendix Table 10). **GT-SD** extracts topics from the documents only (no slide-anchoring); **GT-PR** replaces DivRank with vanilla PageRank; **GT-Greedy** replaces the global ILP with per-slide top-K assignment (no reuse cap); **GT-NoEnrich** drops cross-document evidence enrichment. All five are scored together in one comparative judge pass per sample.

Reading the ablation honestly. The dominant factor by far is *slide-anchoring*: removing it (GT-SD) collapses Coverage from $\approx$3.9–4.1 to $\approx$2.8–2.9, with parallel drops on Relevance, Readoff Avoidance, No-Leakage, and Chart Explanation. The other three components move the headline Coverage/Relevance metrics only modestly; their value is clearest on grounding (enrichment) and deck-wide balance (DivRank, ILP). We give the full per-component breakdown, including the full SEC EDGAR set (n=45), in Appendix G.

Table 5: Human evaluation on M3AV (6 presentations, 2 annotators; Gemma-3-27B). Cells show the mean (1–5); standard errors are reported in the text where cited. **Bold**=best per column. GRAPHTALK is preferred on Coverage and Supporting-Document Usage.

Method	Cov.	Rel.	Div.	SDUsage	Faith.	Trans.	ReadAv.	NoLeak.	DecIm.	ChartEx.
Direct	3.17	**4.67**	**4.33**	3.33	**4.33**	**4.83**	4.67	4.33	**4.33**	4.00
RAG	4.17	4.50	4.17	4.33	**4.33**	4.00	**4.83**	**4.50**	**4.33**	**4.50**
GraphTalk	**5.00**	3.33	2.67	**4.83**	3.00	3.67	3.67	2.67	3.17	3.00

4.6 COST AND EFFICIENCY

A structured pipeline could be dismissed as buying quality with extra compute; it does not. GRAPHTALK's total token budget is comparable to or below the strongest baselines: on SEC EDGAR it uses fewer total tokens than Summ on both generators, and fewer than Direct with Pixtral (close to Direct with Gemma), because it distils long documents into a compact topic layer rather than feeding their full text to the generator. Its extra wall-clock is spent in CPU planning stages, not on a larger LLM budget (Table 8, Appendix D).

4.7 HUMAN EVALUATION

We conducted a small blind human study on M3AV (6 presentations, 2 annotators who are AI researchers familiar with the task), comparing Direct, RAG, and GRAPHTALK with all narrations produced by the same generator (Gemma-3-27B). Systems were shown under anonymous labels with an independently randomised label-to-system mapping per presentation, applied only after scoring. Annotators rated each presentation on all ten rubric dimensions used by the LLM judge. We report all ten rubric dimensions (Table 5).

What the human study supports. Humans support our two central claims: GRAPHTALK is best on Coverage (5.00) and on Supporting-Document Usage (4.83). Supporting-Document Usage is exactly the dimension automatic judges score poorly (§4.4), because the content volume is large and the judge must check that every relevant document claim is used; the human study is more trustworthy here, and it confirms GRAPHTALK performs well on this metric. Judge–human rank agreement is strong on Coverage (Spearman $+0.75$), and we treat the study as targeted human evidence for the Coverage and Supporting-Document Usage claims.

5 CONCLUSION

We presented GRAPHTALK, a training-free pipeline that replaces raw, high-volume source context with an explicit, interpretable structured planning layer built from slides and documents via DivRank-guided ranking and capacity-constrained slide–topic allocation. Its robust advantage on *document-grounded coverage* and *relevance*, across deterministic metrics and three LLM judges, holds at a token budget on par with the strongest baselines, and a human study supports the coverage advantage. Automatic judges are unreliable for supporting-document usage, so we support that dimension with the human study, where GRAPHTALK also leads. Promising extensions include richer visual slide representations, learned fact-density signals, and tighter judge–human calibration.

ETHICS STATEMENT

This work uses two public datasets (M3AV academic lectures; SEC EDGAR financial filings) and introduces no new human-subjects data collection beyond a small voluntary annotation by researchers on the team, who rated anonymised system outputs and no personal data. Automatically generated narration can contain factual errors or hallucinations; we study Source Faithfulness explicitly and caution that outputs intended for accessibility, e-learning, or investor communication should be verified before use. Financial narration in particular must not be treated as investment advice.

Reproducibility Statement

The method is specified in §3 and Algorithm 1; all hyperparameters, model deployments, and per-stage implementation choices are in Appendix C (Table 7), with judge-metric rubrics in Appendix E. Both datasets are public. We report exact sample sizes for every table, including the sub-sampled analyses (Tables 4, 9, 8, 5), and describe filtering and failed-sample handling in Appendix C. The method and all settings are specified in full (§3, Appendix C), including the generation and judge prompt design, to support independent reimplementation.

Use of Large Language Models

LLMs are the object of study in this paper (as generators and as judges) and are described in §4.1. Separately, we used a general-purpose AI assistant to aid drafting and to help implement and aggregate experiment scripts; all such code was reviewed and tested by the authors, and all reported numbers were produced by the authors' own experimental runs. We take responsibility for the final content of this work.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, et al. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Sambaran Bandyopadhyay, Himanshu Maheshwari, Anandhavelu Natarajan, and Apoorv Saxena. 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.

Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pp. 65–72, 2005.

Regina Barzilay and Mirella Lapata. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34, 2008.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

Zhe Chen, Heyang Liu, Wenyi Yu, Guangzhi Sun, Hongcheng Liu, Ji Wu, Chao Zhang, Yu Wang, and Yanfeng Wang. M3AV: A multimodal, multigenre, and multipurpose audio-visual academic lecture dataset. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9041–9060, 2024.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. FinQA: A dataset of numerical reasoning over financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3697–3711, 2021.

Peng Cui and Le Hu. Topic-guided abstractive multi-document summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 1463–1472, 2021.

Alexander R. Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1074–1084, 2019.

Tsu-Jui Fu, William Yang Wang, Daniel McDuff, and Yale Song. DOC2PPT: Automatic presentation slides generation from scientific documents. In *Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence*, pp. 634–642, 2022.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2020.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):248:1–248:38, 2023.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, pp. 611–626, 2023.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pp. 74–81, 2004.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024.

Lefteris Loukas, Manos Fergadiotis, Ion Androutsopoulos, and Prodromos Malakasiotis. EDGAR-CORPUS: Billions of tokens make the world go round. In *Proceedings of the Third Workshop on Economics and Natural Language Processing (ECONLP), EMNLP 2021*, pp. 13–18, 2021.

Himanshu Maheshwari, Sambaran Bandyopadhyay, Aparna Garimella, and Anandhavelu Natarajan. Presentations are not always linear! GNN meets LLM for text document-to-presentation transformation with attribution. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 15948–15962, 2024.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1906–1919, 2020.

Qiaozhu Mei, Jian Guo, and Dragomir Radev. DivRank: The interplay of prestige and diversity in information networks. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1009–1018, 2010. doi: 10.1145/1835804.1835931.

Ishani Mondal, Shwetha S, Anandhavelu Natarajan, Aparna Garimella, Sambaran Bandyopadhyay, and Jordan Boyd-Graber. Presentations by the humans and for the humans: Harnessing LLMs for generating persona-aware slides from documents. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2664–2684, 2024.

Edward Sun, Yufang Hou, Dakuo Wang, Yunfeng Zhang, and Nancy X. R. Wang. 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.

Wen Xiao, Iz Beltagy, Giuseppe Carenini, and Arman Cohan. PRIMERA: Pyramid-based masked sentence pre-training for multi-document summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5245–5263, 2022.

Hao Zheng, Xinyan Guan, Hao Kong, Wenkai Zhang, Jia Zheng, Weixiang Zhou, Hongyu Lin, Yaojie Lu, Xianpei Han, and Le Sun. PPTAgent: Generating and evaluating presentations beyond text-to-slides. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 14402–14418, 2025.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Zeyu Zhu, Kevin Qinghong Lin, and Mike Zheng Shou. Paper2video: Automatic video generation from scientific papers. In *Workshop on Scaling Environments for Agents*, 2025. URL https://openreview.net/forum?id=LvRHonr4gv.

A NOTATION

Symbol	Meaning
$\mathcal{S} = (s_1, \ldots, s_N)$	Ordered slide deck (N slides)
$x_i = [u_i; v_i]$	Slide text and visual description
$\mathcal{D} = \{d_r\}_{r=1}^{R}$	Supporting documents, labels ℓ_r
$\mathcal{T}^0$	Initial topic pool extracted from slides
$\mathcal{T} = \{t_1, \ldots, t_K\}$	Merged atomic topic set
$\mathcal{E}_k = \{(e_{k,j}, \ell_{k,j})\}$	Evidence strings for topic t_k
$\mathbf{v}_k \in \mathbb{R}^d$	Hierarchical embedding of topic t_k
$\mathbf{s}_i \in \mathbb{R}^d$	Hierarchical embedding of slide s_i
$G = (\mathcal{T}, \mathcal{E}_G, \mathbf{W})$	Weighted topic graph
$\mathbf{p}^*$	PageRank prior over topics
$\mathbf{h} = (h_1, \ldots, h_K)$	DivRank importance scores
σ_{ik}	Cosine similarity of s_i and t_k
$\mathbf{X} \in \{0,1\}^{N \times K}$	Slide-topic assignment
$\mathcal{A}_i$	Topics assigned to slide s_i
α_{ik}	Slide-local DivRank weight of t_k
ω_{ik}	Bipartite edge utility

Table 6: Notation used in the GRAPHTALK methodology.

B END-TO-END ALGORITHM

Algorithm 1: GRAPHTALK

Input : Slide deck $\mathcal{S} = (s_1, \ldots, s_N)$; documents $\mathcal{D} = \{d_r, \ell_r\}$; thresholds $\tau_{\text{merge}}, \tau_g, \eta$; capacities $K_{\text{slide}}, B_{\text{topic}}$

Output: Narrations $\mathcal{Y} = (y_1, \ldots, y_N)$

Parse each s_i into $x_i = [u_i; v_i]$

Extract atomic topics $\mathcal{T}^0 = \{t_j\}$ with initial evidence from $\mathcal{S}$

foreach $t_j \in \mathcal{T}^0$ (in parallel) **do**

  Retrieve and append supporting evidence from $\mathcal{D}$ to t_j

Embed $\mathcal{T}^0$ with Eq. equation 1; cluster by cosine sim. (τ_{merge})

foreach *cluster* (in parallel) **do**

  LLM-verify and fuse $\rightarrow$ merged topic

$\mathcal{T} \leftarrow$ fused set; re-embed to obtain $\{\mathbf{v}_k\}$

Build topic graph G (τ_g); $\mathbf{p}^* \leftarrow \text{PageRank}(G)$; $\mathbf{h} \leftarrow \text{DivRank}(G, \mathbf{p}^*, \eta)$

Embed slides as topics; compute $\sigma_{ik} = \mathbf{s}_i^\top \mathbf{v}_k$

Solve Eq. equation 2 for $\mathbf{X}$; compute α_{ik}

foreach $s_i \in \mathcal{S}$ (in parallel) **do**

  $\hat{y}_i \leftarrow \text{LLM}(x_i, \mathcal{A}_i, \{\mathcal{E}_k\})$

for *each window of C consecutive slides* **do**

  Polish drafts conditioned on preceding narrations

return $\mathcal{Y} = (y_1, \ldots, y_N)$

C REPRODUCIBILITY DETAILS

Models. Generation uses `mistralai/Pixtral-12B-2409` and `google/gemma-3-27b-it` (both vision-capable, bfloat16, vLLM); the primary judge is `Meta-Llama-3.3-70B-Instruct` (AWQ-INT4), temperature 0; the secondary judges are GPT-4o and GPT-4.1. Embeddings use `all-MiniLM-L6-v2` (d=384). Generation temperature is 0.7; max draft/polish tokens are 1024/512.

Parameter	Value	Description
$K_{\max}$	50	Max atomic topics per document
Doc truncation	70/30	Head/tail split at 120K chars
α	0.5	Title/evidence weight (Eq. 1)
τ_{merge}	0.85	Cosine threshold for merge candidates
τ_g	0.65	Topic graph edge threshold
κ	3	kNN fallback degree for isolates
δ	0.85	DivRank damping factor
$T_{\min}/T_{\max}$	5 / 30	Min/max DivRank iterations
η	0.35	Entropy early-stop threshold
K_{slide}	7	Max topics per slide
B_{topic}	3	Max slides a topic can serve
$\lambda_s/\lambda_d/\lambda_f$	0.45 / 0.35 / 0.20	Edge-utility weights
C	10	Polish window size (slides)

Table 7: Hyperparameter values used in all experiments. No dataset-specific tuning was applied.

Edge utility and fact density. For each candidate pair (s_i, t_k), $\omega_{ik} = \lambda_s \hat{\sigma}_{ik} + \lambda_d \widehat{\log h_k} + \lambda_f \hat{f}_k$, where $\hat{\cdot}$ is min-max normalization over candidate edges, $\widehat{\log h_k}$ is log-compressed DivRank importance, and $\hat{f}_k$ is normalized fact density (indicators for percentages, currency, fold-change, plus-minus ranges, and multi-digit integers). The optimization is min-cost max-flow on a source→slide→topic→sink graph, solved with NetworkX.

Filtering and failed samples. Reported sizes reflect counts after filtering low-quality samples; per-metric statistics exclude samples whose judge call failed to parse or timed out (marked with a sentinel and dropped from that metric only), which is why some cells carry slightly different n.

D Additional Result Tables

This appendix collects additional result tables whose headline numbers are already reported in the prose of §4: the generation-cost breakdown (Table 8), the image-grounded judge scores (Table 9), the planner and token-matched comparison (Table 11), and the paired significance tests (Table 12).

Table 8: Generation cost: total tokens (thousands) and SEC wall-clock per slide. M3AV n=151; SEC EDGAR n=23 (Pixtral) / 24 (Gemma).

	M3AV total tok		SEC total tok		SEC s/slide	
Method	Pixtral	Gemma	Pixtral	Gemma	Pixtral	Gemma
Direct	286k	237k	843k	704k	18	39
Summ	159k	137k	935k	975k	22	46
RAG	148k	119k	492k	556k	13	20
GraphTalk	306k	250k	583k	762k	24	37

Table 9: Image-grounded GPT-4o judge (shown the rendered slide image), Coverage / Relevance on a SEC EDGAR subset (8 decks per generator). The ranking is stable: top methods stay on top and Direct stays last. **Bold**=best per column.

	Pixtral		Gemma	
Method	Cov.	Rel.	Cov.	Rel.
Direct	2.97	3.67	3.59	4.03
Summ	4.09	4.26	4.01	4.24
RAG	**4.45**	**4.50**	4.36	**4.54**
GraphTalk	4.34	4.35	**4.51**	4.43
GT-SD	3.98	4.27	4.00	4.18

Table 10: Component ablation on M3AV (n=151, comparative judge pass; mean). Slide-anchoring is the dominant factor for Coverage/Relevance; enrichment most affects SDUsage/Faithfulness; DivRank and the ILP contribute smaller, diversity- and stability-oriented effects. Absolute scores are pass-specific and not directly comparable across tables. **Bold**=best per column within each generator block.

Gen	Variant	Cov.	Rel.	Div.	SDUsage	Faith.	Trans.
Pixtral	GraphTalk (full)	**3.94**	**3.98**	3.75	2.66	3.91	3.38
	GT-PR (−DivRank)	3.89	3.97	3.57	2.65	3.98	**3.40**
	GT-Greedy (−ILP)	3.89	3.94	**3.79**	**2.67**	**4.00**	**3.40**
	GT-NoEnrich (−enrich)	3.81	3.93	3.68	2.65	3.93	**3.45**
	GT-SD (−slide-anchor)	2.84	3.07	**3.83**	**2.77**	**4.02**	3.17
Gemma	GraphTalk (full)	**4.13**	4.09	3.75	2.80	3.98	3.46
	GT-PR (−DivRank)	4.11	**4.09**	3.64	2.82	4.00	3.50
	GT-Greedy (−ILP)	4.05	4.05	3.81	2.79	3.97	**3.52**
	GT-NoEnrich (−enrich)	4.02	**4.09**	**3.86**	2.79	**4.03**	3.50
	GT-SD (−slide-anchor)	2.94	3.10	3.85	**2.97**	3.97	3.21

Table 11: Planner control (GT-LLMPlan) and token-matched baseline (RAG-TokenMatched) on SEC EDGAR, scored in one comparative judge pass together with the reused RAG and GRAPHTALK narrations. LLM-judge scores (1–5), mean ± SE. Absolute scores are pass-specific and not directly comparable across the other tables. Cell n ranges from 34 to 45, because a few reused GRAPHTALK narrations returned an invalid judge score; the paired significance in Table 12 uses the common samples.

Gen	Method	Coverage	Relevance	Faithfulness	SD-Usage
Pixtral	GraphTalk	4.03±0.05	3.91±0.05	3.83±0.17	3.12±0.09
	GT-LLMPlan	3.63±0.09	3.68±0.07	3.73±0.14	2.98±0.09
	RAG	3.58±0.06	3.68±0.06	3.87±0.15	3.09±0.09
	RAG-TokenMatched	3.45±0.07	3.58±0.07	4.09±0.15	3.07±0.07
Gemma	GraphTalk	4.34±0.04	4.16±0.05	4.25±0.13	2.99±0.09
	GT-LLMPlan	4.13±0.05	3.93±0.06	4.18±0.12	2.99±0.09
	RAG	3.51±0.07	3.63±0.07	4.08±0.14	2.89±0.08
	RAG-TokenMatched	3.35±0.05	3.53±0.06	4.10±0.11	2.94±0.08

E JUDGE METRIC DEFINITIONS

Each of the ten dimensions is scored 1–5 (higher better) per slide with a self-contained rubric prompt, then averaged to a deck-level value. Source Faithfulness and Supporting-Document Usage require the full source text. On the main M3AV run they are computed on a partial subset (n≈104–105), which is why some M3AV cells carry a smaller n. They are omitted from the main SEC EDGAR table (Table 2) but are computed on SEC EDGAR in the comparative control and ablation passes (Tables 11, 14), where the full document text is available. Transition is scored at slide boundaries; Chart Explanation and No Leakage are skipped (score 5) when inapplicable (no chart present; final slide). When multiple systems are compared, narrations for the same slide are scored together and randomly shuffled to reduce position bias, with the mapping restored after scoring.

F PER-JUDGE METHOD SCORES

Table 13 reports Coverage and Relevance under the two additional judges (GPT-4o, GPT-4.1) on the full SEC EDGAR set, complementing the primary Llama-3.3-70B judge in Table 2. Under both

Table 12: Paired significance of GRAPHTALK against the two controls, pooled over both datasets and both generators. Wilcoxon signed-rank (two-sided); the 95% CI is a 10k-sample bootstrap of the mean paired difference (GRAPHTALK − baseline). Stars: $^{***}p < 0.001$, $^{**}p < 0.01$; ns means not significant.

Metric	Baseline	n	mean Δ	95% CI	p
Slide Coverage	GT-LLMPlan	371	$+0.170$	$[+0.119, +0.222]$	1.1×10^{-10} (***)
	RAG-TokenMatched	372	$+0.697$	$[+0.643, +0.750]$	2.6×10^{-54} (***)
Relevance	GT-LLMPlan	371	$+0.143$	$[+0.092, +0.192]$	6.1×10^{-8} (***)
	RAG-TokenMatched	372	$+0.426$	$[+0.373, +0.479]$	1.6×10^{-37} (***)
SD-Usage	GT-LLMPlan	371	$+0.038$	$[-0.015, +0.091]$	0.17 (ns)
	RAG-TokenMatched	372	-0.098	$[-0.158, -0.039]$	1.2×10^{-3} (**)

additional judges and both generators, GRAPHTALK ranks first on Coverage and Relevance, so the headline ordering does not depend on any single judge.

Table 13: Per-judge method scores on SEC EDGAR (45 decks per generator; the same narrations scored by each judge), for the two additional judges beyond the primary Llama-3.3-70B. GRAPHTALK ranks first on Coverage and Relevance under both judges and both generators. **Bold**=best per column within each generator block.

		Coverage		Relevance	
Gen	Method	GPT-4o	GPT-4.1	GPT-4o	GPT-4.1
Gemma	Direct	2.74	2.45	3.14	2.65
	Summ	3.13	2.77	3.30	2.81
	RAG	3.52	3.30	3.65	3.42
	GraphTalk	**4.32**	**4.30**	**4.01**	**4.06**
Pixtral	Direct	2.05	1.80	2.41	2.06
	Summ	3.40	3.27	3.47	3.26
	RAG	3.68	3.53	**3.83**	3.49
	GraphTalk	**3.97**	**3.79**	**3.83**	**3.73**

G COMPONENT ABLATION ON SEC EDGAR

We now run the full SEC EDGAR set (n=45 per generator; one comparative judge pass per deck; Table 14). Removing slide-anchoring (GT-SD) again causes the largest drop, and it is the only large one: Coverage falls from 3.88 to 2.76 on Pixtral and from 4.12 to 2.71 on Gemma, with matching drops on Relevance, Readoff Avoidance, No-Leakage, and Chart Explanation. The other three components move Coverage and Relevance only slightly, within 0.15 of the full model on both generators, which confirms on the full sample that slide-anchoring is the dominant factor. Their effects are smaller than the earlier partial-sample run suggested: at n=45, dropping enrichment (GT-NoEnrich) leaves Faithfulness essentially unchanged (3.72 vs 3.69 on Pixtral, 4.05 vs 4.12 on Gemma), and its clearest cost is a lower Coverage on Gemma (4.12 to 3.97) at the smallest token budget of all variants. We report SEC EDGAR separately rather than pooling it with M3AV.

H HYPERPARAMETER SENSITIVITY (STRUCTURAL)

We swept the six consequential hyperparameters around their defaults on 40 presentations (both datasets, both generators), measuring the structural response of the planning layer. For each hyperparameter we average a structural metric over presentations at each grid value and report the coefficient of variation (CV) of those per-setting means (Table 15). Outcome-relevant structure is stable (CV $\leq$ 6%); quantities a hyperparameter is designed to control respond as intended, but none destabilise coverage, so the single reported configuration is representative rather than finely tuned. This is a structural analysis of the planning layer.

Table 14: Component ablation on SEC EDGAR (n=45 per generator; one comparative LLM-judge pass per deck). LLM-judge dimensions (1–5, higher is better), mean ± SD. Absolute scores are pass-specific and not directly comparable across the other tables. Removing slide-anchoring (GT-SD) is the only change that moves Coverage or Relevance by a large margin.

Gen	Method	Cov.	Rel.	Faith.	SDUsage	Trans.	Div.
Pixtral-12B	GraphTalk	3.88±0.25	3.90±0.23	3.72±1.11	3.08±0.62	3.88±0.67	3.76±0.80
	GT-PR	3.87±0.31	3.92±0.25	3.74±0.98	3.08±0.50	3.25±0.85	3.60±0.81
	GT-Greedy	3.80±0.31	3.89±0.34	3.77±0.98	3.05±0.58	3.61±0.89	3.71±0.92
	GT-NoEnrich	3.78±0.39	3.84±0.32	3.69±1.22	3.09±0.47	3.17±0.89	3.64±0.88
	GT-SD	2.76±0.48	2.97±0.47	3.73±1.06	3.15±0.49	2.92±0.96	3.64±0.86
Gemma-3-27B	GraphTalk	4.12±0.19	4.09±0.23	4.05±1.08	3.14±0.49	3.28±1.10	3.69±0.73
	GT-PR	4.10±0.23	4.08±0.22	4.21±0.97	3.12±0.55	3.55±0.86	3.73±0.69
	GT-Greedy	4.11±0.21	4.11±0.26	4.06±1.11	3.18±0.47	3.38±0.98	4.00±0.80
	GT-NoEnrich	3.97±0.26	4.08±0.22	4.12±1.04	3.22±0.40	3.31±0.95	3.87±0.73
	GT-SD	2.71±0.34	2.95±0.30	4.17±1.01	3.13±0.54	2.95±0.94	3.49±0.94

Hyperparameter (grid)	Key structural metric	Variation (CV)
τ_g (0.55–0.75)	DivRank concentration	4.1%
τ_{merge} (0.80–0.90)	# merged topics	5.8%
δ (0.75–0.95)	DivRank concentration	1.5%
η (0.35–0.75)	DivRank concentration	2.6%
K_{slide} (5–9)	slide coverage	0.0%
B_{topic} (2–4)	slide coverage	0.0%

Table 15: Structural sensitivity of the planning layer (40 presentations).

I Perturbation Sensitivity Details

For each of 3 document sets we build cascadingly corrupted variants of the original narration at 0/10/20/30/40/50% severity, score all levels together in one comparative judge call, and compute the Spearman ρ between severity and $-$score, averaged over sets. Headline dimensions respond near-monotonically (Source Faithfulness 0.971, Supporting-Document Usage 0.970), evidence that the judge is sensitive to corruption on these two dimensions.

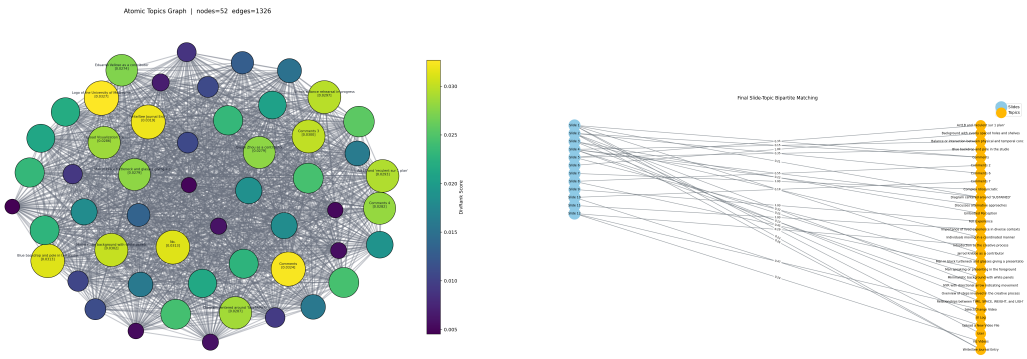
J Case Studies

All examples use Pixtral-12B, chosen because its intermediate pipeline artefacts are fully available for inspection.

Qualitative narration comparison. On M3AV sample CHI-004BD, only GraphTalk incorporates a direct source quotation from the supporting CHI paper (a Pina Bausch quotation on dance and technique), whereas Direct and RAG produce generic framings absent in the source. On SEC EDGAR sample AIG_Q2_2023, GraphTalk grounds the narration in a specific transaction date (May 2023) and broader strategic context from the 10-K, both absent in the competing outputs.

Topic graph and assignment. Figure 3 visualises the topic-similarity graph and bipartite assignment for CHI-004BD (20 topics, 12 slides); DivRank concentrates mass in three hub topics ($0.282 + 0.277 + 0.233 = 79\%$ of total score), with a long tail of lower-ranked topics that matching distributes across slides.

DivRank evolution. Figure 4 traces how DivRank concentrates importance on the CHI-1EE11 topic graph (15 topics), with normalised entropy falling from the near-uniform PageRank prior toward a concentrated distribution before entropy-based early stopping.



(a) Topic similarity graph (τ_g=0.65); node size $\propto$ DivRank.

(b) Bipartite assignment: 58 edges, 12 slides, 20 topics.

Figure 3: Topic graph and slide-topic assignment for M3AV sample CHI-004BD (GRAPHTALK, Pixtral-12B).

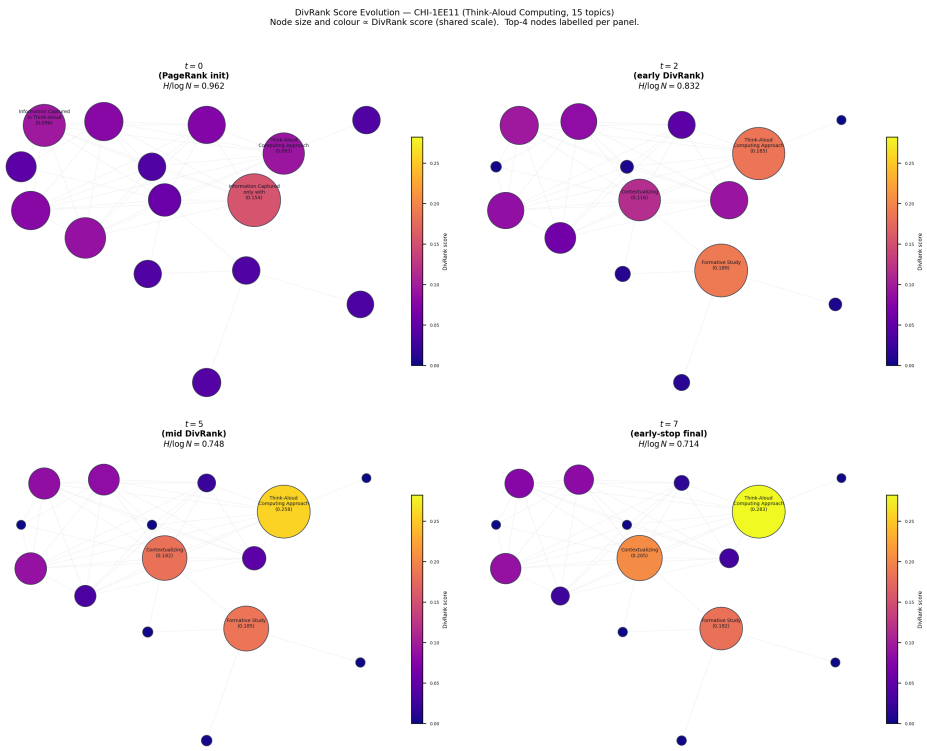


Figure 4: DivRank score evolution on the CHI-1EE11 topic graph across four snapshots; node size is normalised per panel and colour is on a shared scale. Normalised entropy $H/\log N$ falls from the near-uniform PageRank prior toward a concentrated distribution, stopping early once it approaches η=0.35.

Qualitative observations. A close read of outputs on M3AV sample CHI-903ED (GRAPHTALK, Pixtral-12B, 12 slides) surfaces two small, cosmetic side effects of drafting slides in parallel, both easily smoothed by a light consistency pass. Because slides are drafted independently, a few openings can echo one another (some begin with a fresh greeting), and a coined entity can vary slightly in surface form across slides (here the product name appears as *Priview*, *PriView*, and *PrivaView*). The sequential polish pass already reconciles most local flow; a short deck-level normalisation step

would tidy these remaining surface details while preserving the document grounding that the planning layer contributes.