# SCISLIDE: SCIENTIFIC SLIDE GENERATION AS VERIFIABLE ARGUMENT RECONSTRUCTION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Preparing slides for a scientific talk is a document-to-multimodal generation task: text, figures, and equations must be placed on a visual canvas to guide an audience through the paper's causal argument. We formulate this task as argument reconstruction rather than extractive summarization, with explicit verification at each step. SciSlide builds a multi-slide argument graph in which every node carries a verbatim source span and every figure attaches to its supporting claim, enabling string-based verification of coverage and grounding. We score decks with sixteen computable metrics that operationalize the assertion–evidence paradigm for technical presentations, and learn only the layout component (coordinate prediction and geometric composition) from 7,500 conference decks paired with their papers. Against four published systems and a prompted frontier LLM, SciSlide leads on assertion headlines, argument-driven slide order, and figure–claim alignment; human raters prefer its decks in $65\%$ of pairwise comparisons. For layout, autoregressive coordinate generation outperforms classification by $6.2$ percentage points in strict exact match.

## 1 INTRODUCTION

A conference talk on a research paper is usually given with a deck of about a dozen slides. Preparing that deck is a task separate from writing the paper: the deck must present an argument that the audience can follow in real time. Automatic paper-to-slide systems (Hu & Wan, 2014; Sun et al., 2021; Fu et al., 2022; Zheng et al., 2025; Liang et al., 2025; Zheng et al., 2026; Cui et al., 2026) aim to reduce this effort. We focus on scientific talks, whose success depends on the audience following the causal argument and locating evidence for each claim. Machine-learning and computer-vision venues serve as the testbed: they supply the largest public paper–slide collection, and the task requires document layout analysis, visual asset grounding, and coordinate prediction—core capabilities of both communities.

Published systems either edit a reference slide (Zheng et al., 2025) or synthesize slides without one (Liang et al., 2025; Cui et al., 2026). In both cases the paper text is the unit of transfer and a template the unit of placement, so the task reduces to extractive summarization followed by template filling. This formulation does not model three properties of a presentation deck. (i) Design attributes: font size, alignment, emphasis, color, and theme are inherited from the template rather than predicted, although the source decks in our corpus use a median of three font-size bins per slide. (ii) Argument structure: selecting salient sentences keeps the sentences but drops the causal relations among problem, method, evidence, and claim, so slide order is not constrained by the argument. (iii) The assertion–evidence form: for technical talks, the assertion–evidence paradigm (Alley, 2003; Garner & Alley, 2013), which rests on multimedia-learning and cognitive-load principles (Mayer, 2021; Sweller, 1988), prescribes a one-sentence assertion headline above visual evidence instead of a topic phrase above a bullet list; systems optimized for summary fidelity produce the second form.

We therefore formulate the task as argument reconstruction and implement this formulation in SciSlide, which builds an explicit, source-grounded representation of the paper's argument and derives the deck from it (Figure 1). This formulation enables deterministic verification: the constraints listed above can each be checked by a program over the source paper and the rendered deck. Assertion headlines, evidence-bearing slide bodies, argument order, figure-to-claim placement, and geometric validity each admit a deterministic program over the source paper and the rendered deck, and one



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.

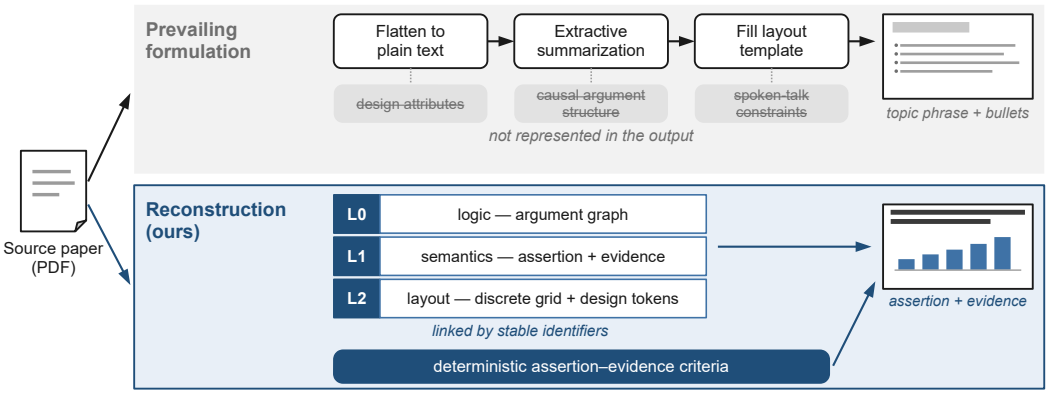



Figure 1: Two formulations of the same task. Flattening the paper and filling a template (top) leaves design attributes, causal argument structure, and the constraints of a spoken talk unrepresented. SciSlide reconstructs the paper in three layers linked by stable identifiers (bottom) and scores the result with deterministic criteria. Both thumbnails are drawn at the same scale.

such program can serve as the evaluation protocol, as the acceptance test of a correction loop, and as a filter on training data. This extends the finding of Pan et al. (2026), that rule-based rewards are more reliable than VLM judgment for slide aesthetics, to content and argument structure. Our contributions are the following.

- A multi-slide argument graph in which every node stores a verbatim source span, every content item the node it realizes, and every figure the node it supports, so that argument coverage and figure placement are checked against the source PDF rather than asserted by a prompt (Sec. 3.3).
- Sixteen deterministic metrics that operationalize the assertion–evidence criteria, with a validation against human ratings on 180 decks (30 papers $\times$ 6 systems) that identifies which metrics track human judgment and which are optimized by degenerate outputs (Secs. 3.2 and 5.3).
- A corpus of 7,500 conference decks paired with their source papers across 24 venues ($\sim$130k pages), re-expressed in a three-layer schema that retains five design attributes as discrete tokens (Sec. 4).
- A layout model that emits quantized coordinates autoregressively. On a matched 7B backbone it raises strict exact match from $8.33\%$ to $14.53\%$ over a discriminative head, and at 14B it reaches $22.7\%$; two auxiliary training signals reduce accuracy (Sec. 5.4).

## 2 Related Work

**Slide generation.** Early work treats slide generation as extractive summarization (Kan, 2007; Hu & Wan, 2014; Wang et al., 2017; Sefid et al., 2021; Qiang et al., 2016; Cachola et al., 2024; Sun et al., 2021; Fu et al., 2022; Lev et al., 2019). Current systems edit reference slides (Zheng et al., 2025; Jung et al., 2026; Shi et al., 2026), generate slides as programs (Ge et al., 2025; Xu et al., 2025), or add agent self-refinement (Liang et al., 2025; Ma et al., 2026; Cui et al., 2026; Liu et al., 2026; Zheng et al., 2026). In all cases the paper text is the transfer unit and a template the placement unit. Available datasets contain generic presentations without paired papers (Zheng et al., 2025), few targets (Zeng et al., 2026), or aesthetic annotations (Cui et al., 2026; Liu et al., 2026); earlier pairs (Kan, 2007; Sun et al., 2021; Fu et al., 2022) lack design tokens and argument structure.

**Argument structure and communication principles.** Deck structure has been modeled as discourse trees (Ozden et al., 2026; William & Thompson, 1988), entity-preserving graph indices (Meier et al., 2025), and logic chains (Yang et al., 2026), and sentences of scientific text are known to carry recoverable rhetorical roles (Teufel & Moens, 2002; Jin & Szolovits, 2018; Lawrence &

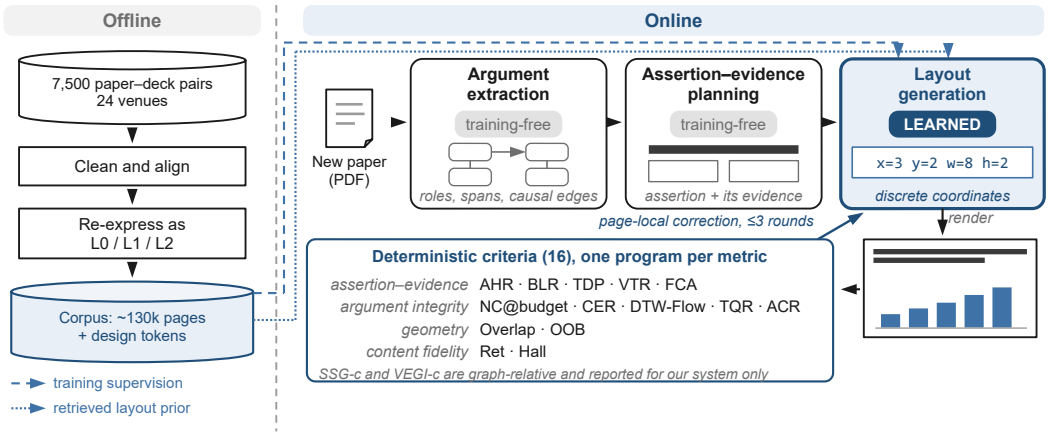



Figure 2: Overview of SciSlide. Offline (left), real decks are cleaned and re-expressed in the three-layer schema of Sec. 3.1; the resulting corpus supplies the training data and the retrieval memory for layout. Online (right), argument extraction and content planning (Sec. 3.3) are training-free, and layout generation (Sec. 3.4) is the only learned component. The same deterministic criteria score the rendered deck and trigger page-local correction.

Reed, 2020). In these systems the structure is produced by a prompt and is neither checked against the source nor reported as a metric. Our graph stores a verbatim source span per node and a node reference per content item and asset, so coverage, edge retention, and figure placement are computed by matching. The assertion–evidence paradigm (Alley, 2003; Garner & Alley, 2013) and the signaling, modality, and contiguity principles (Mayer, 2021; Mayer & Moreno, 2003) have entered automatic systems only as prompt text; we map each of them to a metric.

**Evaluation and layout.** Model judges show position, verbosity, and self-preference biases (Zheng et al., 2023; Wang et al., 2024; Dubois et al., 2024; Panickssery et al., 2024) that prompting reduces but does not remove (Liu et al., 2023), and $n$-gram and embedding metrics (Lin, 2004; Zhang et al., 2019) measure similarity to a reference rather than suitability for a talk. Recent slide benchmarks therefore use computable checks (Pan et al., 2026; Chen et al., 2026b; Zhao et al., 2026; Chen et al., 2026a; Jang et al., 2026), in line with the move from learned to verifiable rewards in post-training (Lambert et al., 2024; Guo et al., 2025); we apply the same approach to argument structure. Layout generation has been studied with element-sequence models (Li et al., 2019; Gupta et al., 2021; Kong et al., 2022), discrete diffusion (Inoue et al., 2023), constraint serialization (Jiang et al., 2023), language models that emit layouts as text (Lin et al., 2023; Tang et al., 2024), and content-aware placement (Zhou et al., 2022; Hsu et al., 2023), whereas slide systems rely on templates, HTML, or reinforcement learning on aesthetics (Zheng et al., 2025; Liang et al., 2025; Pan et al., 2026; Liu et al., 2026). We adopt the quantize-and-emit formulation of Chen et al. (2021).

# 3 Reconstructing a Paper as a Presentation

## 3.1 Task and Three-Layer Representation

Given a source paper $P$ in PDF form, the task is to produce a deck $D = (s_1, \ldots, s_K)$ that satisfies the criteria $R(D, P) \in \mathbb{R}^{16}$ of Sec. 3.2 and is geometrically valid. $R$ is not aggregated into a scalar, because some of its coordinates are optimized by degenerate decks (Sec. 5.3). Figure 2 shows an overview of SciSlide.

The deck is represented in three layers linked by stable identifiers. L0 (logic) is a multi-slide argument graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{A})$. Each node $v \in \mathcal{V}$ has one of seven roles (problem, gap, method, evidence, claim, contribution, limitation) and a source span $\mathrm{span}(v)$ that occurs verbatim in $P$; each edge has one of six causal relations (`motivates`, `addressed_by`, `evaluated_by`, `supports`, `contrasts`, `follows_from`); each asset $a \in \mathcal{A}$ is a figure or table extracted from

$P$ and attached to the node it supports. L1 (semantics) holds, per slide, one assertion headline—a declarative sentence with a finite main verb—and an ordered list of evidence items, each with a role, a `node_id`, and for visual items an `asset_id`, together with an optional presenter-note script. L2 (layout) places each item on a $24 \times 13$ integer grid as a bounding box with design tokens (font-size bin, alignment, emphasis, color) and assigns a deck-level theme.

L0 and L1 are discrete structures whose constraints—acyclicity, role typing, span validity, one assertion per slide—can be checked directly, so we build them with training-free procedures that apply these constraints as hard filters. L2 is a conditional distribution over geometry given content, for which no closed-form rule exists, and is the only learned component. Because L1 is fixed first, the layout model does not have to decide what a slide says.

**Generation uses a model; scoring does not.** A language model participates in building L0 and L1, so these layers are not deterministic functions of $P$. Everything downstream of them is: a node enters $\mathcal{G}$ only if its span occurs verbatim in $P$, the topology is forced to a connected directed acyclic graph by explicit checks, and every metric is computed from $(P, \mathcal{G}, D)$ by string and index matching without a model. A hallucinated node therefore cannot enter the graph, and the system being evaluated cannot influence its own score.

### 3.2 Assertion–Evidence Criteria as Deterministic Metrics

The criteria are defined before the procedures that satisfy them, because the procedures are designed against them. Three criteria follow from Alley (2003) and Mayer (2021). The assertion criterion (signaling principle; Mayer, 2021) requires each slide to be headed by one declarative sentence rather than a topic phrase and is scored by the assertion-headline ratio (AHR). The evidence criterion (modality and redundancy principles; Mayer, 2021) requires the body to carry figures, tables, or equations rather than text restating the headline, and is scored by the bullet-list ratio (BLR), a text-density penalty (TDP), and the visual-to-text area ratio (VTR). The grounding criterion (contiguity principle; Mayer, 2021) requires each displayed asset to appear on the slide whose assertion it supports; it is scored by the visual-evidence grounding index (VEGI-c) over $\mathcal{G}$ and, for systems that produce no graph, by figure–caption alignment (FCA). Three additional metric families address properties not covered by the assertion–evidence criteria. Argument integrity comprises budgeted node coverage (NC@budget), causal-edge retention (CER), a dynamic-time-warping distance between the slide role sequence and a topological order of $\mathcal{G}$ (DTW-Flow; Sakoe & Chiba, 1978), source-span grounding (SSG-c) with its graph-free counterpart, the traceable quote rate (TQR), and the asset citation rate (ACR). Geometry comprises the pairwise overlap rate (Overlap) and the out-of-bounds rate (OOB). Content fidelity comprises source-token retention (Ret) and the hallucination rate (Hall). Appendix A defines all sixteen. The headline test uses a dependency parser (Qi et al., 2020) and the grounding tests use string matching; no metric invokes a generative model.

**Thresholds and scope.** Three thresholds have no principled derivation from Alley (2003) or Mayer (2021): the body-word budget $\tau$=30 in TDP, the bullet-list size of three items in BLR, and IoU $> 0.30$ in Overlap. These and the thresholds of TQR and FCA were fixed on the validation split before any baseline was run and are not tuned per system. TDP is monotone in $\tau$, so the ranking of decks whose bodies all exceed $\tau$ does not depend on its value; the BLR and Overlap thresholds are conventions of this paper. The criteria encode one paradigm, validated on technical talks (Alley, 2003; Garner & Alley, 2013), and do not apply to keynotes or design portfolios, which are judged partly on grounds that the paradigm deliberately excludes.

**Three uses of one program.** The same programs are the evaluation protocol (Sec. 5), the acceptance test of the correction loop (Sec. 3.4), and a filter that keeps for layout training only pages whose ground-truth geometry satisfies the geometry criteria. They are not used as a differentiable objective; the one differentiable geometric objective we tried reduced accuracy (Sec. 5.4).

### 3.3 Argument Extraction and Content Planning

L0 is built from $P$ in three training-free stages. A layout-analysis front end (a SURYA-based `marker` pipeline (datalab-to, 2024) with a PyMuPDF fallback) converts $P$ into typed blocks: section headers, paragraphs, figures, and tables. We use it because it recovers borderless tables

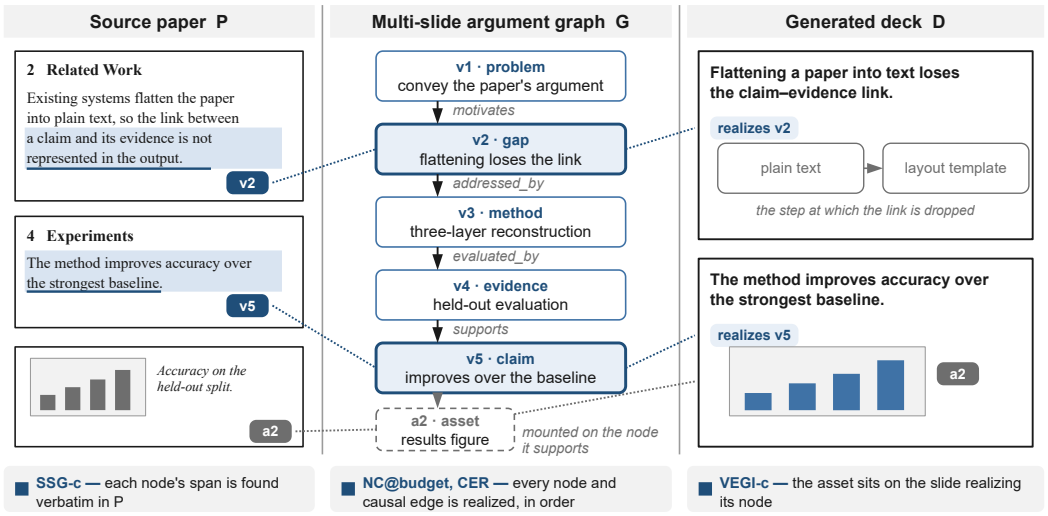



Figure 3: Verifiable grounding, shown schematically. Each node (centre) stores a verbatim source span (left, underlined), each slide (right) records the node it realizes, and each asset records the node it supports, so SSG-c, NC@budget, CER, and VEGI-c are computed by string and index matching rather than judged by a model. The text is illustrative; Appendix G shows rendered slides from a real paper.

and separates pseudo-code from tables; other document-analysis front ends could be substituted (Lopez, 2009; Blecher et al., 2024). A language model then reads each section and emits nodes, with roles initialized from the rhetorical function of the section (Introduction → problem/gap, Method → method, Results → evidence, Conclusion → claim/contribution). A node whose span is not found verbatim in $P$ is rejected and re-extracted, making source grounding a construction invariant rather than a measured quantity. Edges are built rule-first: `addressed_by` and `evaluated_by` follow from section adjacency, the language model proposes the remaining cross-section relations, and cycle and connectivity checks make $\mathcal{G}$ a connected directed acyclic graph. Each figure and table is attached to a node using caption–node token overlap, page proximity to the node's span, and role keywords. Attaching assets to nodes in the graph enables a computable test of whether each figure appears on the slide that presents the claim it supports (Figure 3).

L1 is derived from $\mathcal{G}$ by a second training-free procedure. Slides follow a topological order of $\mathcal{G}$ (problem → method → evidence → claim), with ties broken by source position, so that slide order follows the argument rather than the section order of the paper. For each slide the procedure writes one declarative assertion headline and fills the body with the supporting evidence items, each tagged with its `node_id` and, for visual items, its `asset_id`. A node for which no single declarative assertion can be written is merged into its parent instead of receiving a topic-phrase headline. A presenter-note script is generated per slide from the assertion, the node text, and the asset captions.

### 3.4 Layout as Autoregressive Discrete Coordinates

L2 is learned. We fine-tune Qwen2.5-14B-Instruct (Yang et al., 2025) with LoRA (Hu et al., 2022) and 4-bit quantization (Dettmers et al., 2023) by supervised fine-tuning on the corpus of Sec. 4, without a custom head or auxiliary loss. Following the quantize-and-emit formulation of Chen et al. (2021), the geometry of each element is written as four grid-index tokens in the target sequence (`x=3 y=2 w=8 h=2`), interleaved with its design tokens, and the model is trained with cross-entropy over the full sequence. The alternative is a discriminative head that reads one hidden state per element and applies four independent classifiers over coordinate bins. That hidden state must encode both the identity and the geometry of the element, and the four argmax decisions are made independently, so their errors compound under a criterion that requires all four coordinates to match. Autoregressive generation conditions each coordinate on the content, the design tokens, and the

coordinates already emitted, so elements are placed relative to one another through the self-attention of the backbone.

**Retrieved layouts.** The corpus also serves as a layout memory (Lewis et al., 2020). Each page is indexed by a 20-dimensional content signature, and at inference the $K$ layouts nearest to the L1 content of the target page are supplied as in-context exemplars. Whenever retrieved layouts are used in training, a leave-one-paper-out rule prevents the model from retrieving the layout it is asked to predict.

**Inspection and page-local correction.** After L2 is emitted, the deck is rendered and scored with the criteria. When a metric on a page falls below its threshold, a language-model corrector rewrites that page only, for at most three rounds. Which pages are corrected and when the loop stops are decided by the program, not by a vision-language model reading a screenshot as in Zheng et al. (2026), so the trigger and the stopping rule are reproducible, and a page that already meets its thresholds is never modified.

## 4 A Corpus of Paper-Paired Academic Decks

We collect 7,500 conference decks from 24 venues, each paired with its source paper through a venue-specific submission identifier, following Lo et al. (2020). Each deck is parsed into text runs, images, and shapes with their geometry. Decorative elements below an area threshold are removed, paragraphs split by the converter within one text box are merged, bullet hierarchy is recovered from indentation and font size, font names are normalized, and colors are quantized to a fixed palette. L0 and L1 are obtained by applying the procedures of Sec. 3.3 to the source paper and aligning their output with the real slides; L2 is obtained by discretizing slide geometry onto the $24 \times 13$ grid and recovering five design attributes (font-size bin, alignment, emphasis, color, and theme). Each headline is labeled as an assertion or a topic phrase, and each content item receives a role and, where applicable, a `node_id` and an `asset_id`. The result is $\sim$130k annotated pages. The design tokens allow a model to predict design attributes instead of inheriting them from a template. The source decks have a mean AHR of $0.34$, that is, most headlines in human-made academic decks are topic phrases; this value is the reference point for the generated decks in Sec. 5.2.

## 5 Experiments

We examine whether the argument-grounded reconstruction of SciSlide improves the comparable metrics relative to published systems (Sec. 5.2), whether the metrics track human judgment (Sec. 5.3), and whether the output representation or the backbone accounts for layout accuracy (Sec. 5.4).

### 5.1 Setup

**Data.** The corpus is split 80/10/10 by venue, so no paper occurs in two splits. System-level comparisons use a sample of 300 papers from the test split, spanning 3 venues (NeurIPS, CVPR, ICCV); human ratings are collected on 30 papers sampled from the test split.

**Models and hardware.** The layout generator (Qwen2.5-14B-Instruct, LoRA, 4-bit) is the only trained component; it and all 7B configurations of Sec. 5.4 were fine-tuned on 8 NVIDIA A800-SXM4 GPUs (approximately 420 GPU-hours for the 14B run). Argument extraction and content planning (Sec. 3.3) use Gemini 3 Pro (temperature $1.0$, max tokens $8192$), and the page-local corrector (Sec. 3.4) uses Claude Sonnet 4.6 (temperature $0.7$, max tokens $4096$); both are accessed through their public APIs. The layout generator decodes greedily. No model is involved in computing any metric.

**Baselines.** We compare SciSlide with four published systems with released code: AutoPresent (Ge et al., 2025), program generation; PPTAgent (Zheng et al., 2025), reference-slide editing; SlideTailor (Zeng et al., 2026), Chain-of-Speech generation; and PreGenie (Xu et al., 2025), an agentic Slidev framework. We also prompt Gemini 3 Pro (temperature $1.0$, max tokens $8192$) to produce

Table 1: Results on the metrics that one program computes for all systems (Appendix A): AutoPresent (Ge et al., 2025), PPTAgent (Zheng et al., 2025), SlideTailor (Zeng et al., 2026), PreGenie (Xu et al., 2025), prompted Gemini 3 Pro, and SciSlide (ours). Bold marks the best value in each row.

| Metric | AutoPresent | PPTAgent | SlideTailor | PreGenie | Gemini 3 Pro | SciSlide |
|---|---|---|---|---|---|---|
| AHR ↑ | 0.12 | 0.70 | 0.66 | 0.08 | 0.03 | **0.86** |
| BLR ↓ | 0.020 | **0.003** | 0.199 | 0.195 | 0.026 | 0.011 |
| TDP ↓ | 88.5 | 35.3 | **1.0** | 22.0 | 9.2 | 6.8 |
| VTR ↑ | 0.04 | 0.15 | 0.07 | 0.19 | 0.86 | **0.91** |
| FCA ↑ | 0.09 | 0.27 | 0.35 | 0.04 | 0.33 | **0.84** |
| NC@budget ↑ | 0.62 | 0.57 | 0.41 | 0.48 | 0.39 | **0.85** |
| CER ↑ | 0.72 | 0.51 | 0.36 | 0.36 | 0.36 | **0.76** |
| DTW-Flow ↓ | 3.3 | 2.4 | 3.3 | 3.7 | 4.3 | **1.4** |
| TQR ↑ | **0.81** | 0.61 | 0.52 | 0.51 | 0.45 | 0.63 |
| ACR ↑ | 0.24 | 0.47 | 0.50 | 0.20 | **0.95** | 0.83 |
| Overlap ↓ | 0.060 | 0.003 | **0.000** | 0.007 | 0.017 | 0.002 |
| OOB ↓ | 0.102 | 0.008 | 0.005 | 0.019 | 0.007 | **0.002** |

the deck directly, because it requires no code or system setup and is therefore the baseline most readily available to a researcher. All systems receive the same PDF and the same extracted figures (Appendix I).

**Comparable metrics.** A metric enters Table 1 only if one program computes it identically for every system. SSG-c and VEGI-c are defined over $\mathcal{G}$, which only SciSlide produces, and equal $1.00$ for it by construction (Appendix C); Table 1 reports their graph-free counterparts instead. TQR is the fraction of content-bearing text spans that match some 8-gram of $P$ verbatim, and FCA is the fraction of placed assets whose caption shares idf-weighted content words with the headline of the same slide (weighted Jaccard $\geq 0.25$). Because node coverage increases with the amount of text on a slide, NC@budget truncates every deck to the same number of content items per slide before matching. Overlap and OOB are computed per slide over leaf elements, with one leaf per text frame or picture and groups contributing their children; Appendix B gives the procedure.

### 5.2 System-level Comparison

**SciSlide leads on the argument-grounded metrics.** SciSlide reaches NC@budget $0.85$ against $0.62$ for the best baseline, AutoPresent (Ge et al., 2025), lowers DTW-Flow from $2.4$ (PPTAgent (Zheng et al., 2025)) to $1.4$, and reaches FCA $0.84$ against $0.35$ (SlideTailor (Zeng et al., 2026)). To ensure that the reference graph does not favor SciSlide, we re-extracted it using GPT-4o with the same procedure of Sec. 3.3; on this independently extracted graph, SciSlide achieves NC@budget $0.82$ and the best baseline $0.60$, maintaining a margin of $0.22$. A system without a causal graph has no mechanism for ordering slides by argument or for placing a figure next to the claim it supports. SciSlide also has the highest AHR ($0.86$ against $0.70$ for PPTAgent), above the $0.34$ of the human-made source decks. The CER margin is small ($0.76$ against $0.72$ for AutoPresent): edge retention is not budget-controlled, and AutoPresent, which has the highest text density (TDP $88.5$), realizes both endpoints of many edges by reproducing most of the paper's text. Prompted Gemini 3 Pro, which requires no code or system setup, produces mostly topic-phrase headlines (AHR $0.03$) despite being the same frontier language model that performs our argument extraction. When we explicitly prompt it with the assertion–evidence rules of Alley (2003), AHR rises to $0.68$, but NC@budget remains at $0.43$, DTW-Flow at $3.7$, and FCA at $0.48$, which indicates that the prompt can improve headline syntax but does not provide the causal structure needed for argument-driven slide order and figure placement; Appendix G shows rendered examples.

**Where SciSlide does not lead.** AutoPresent (Ge et al., 2025) reaches the highest TQR ($0.81$ against $0.63$) because copied sentences are traceable by construction, whereas SciSlide rewrites nodes as assertions. SlideTailor (Zeng et al., 2026) reaches TDP $1.0$ by producing slides with an average of $1.0$ words per body (excluding headlines); SciSlide's TDP of $6.8$ is higher, though still below most baselines. On geometry, SciSlide reaches the lowest OOB rate ($0.002$) after page-local correction, tied with the best template-based systems on Overlap ($0.002$).

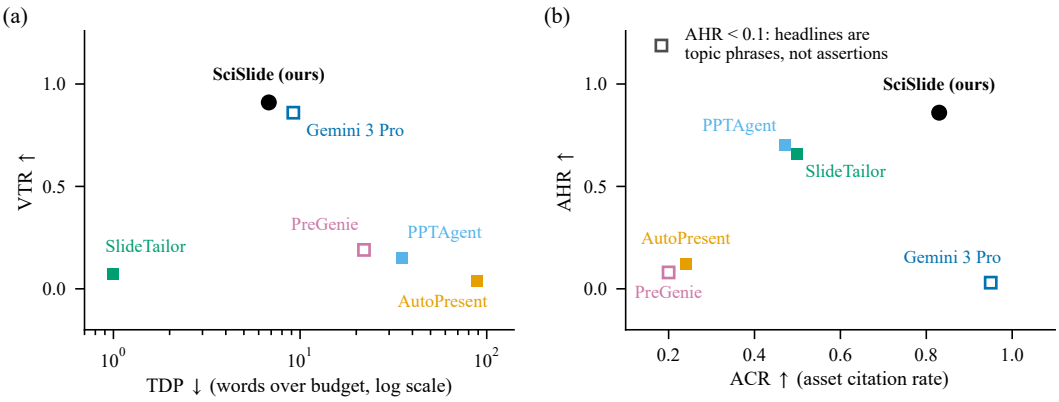



Figure 4: Two metrics that can be optimized by outputs no reader would accept. (a) Text-density penalty (TDP): SlideTailor produces the lowest value by generating nearly empty slides (mean 1.0 words per body). (b) Asset citation rate (ACR): Gemini 3 Pro reaches the highest value by placing every figure, while producing topic-phrase headlines (AHR 0.03). In both panels, hollow markers denote systems whose headlines are almost all topic phrases (AHR $< 0.1$). See Appendix F for the correlation between each metric and human judgment.

**Metrics that do not track quality.** Two metrics can be optimized by outputs no reader would accept. TQR rewards verbatim copying: AutoPresent (Ge et al., 2025) reaches 0.81 by reproducing most of the paper's text (TDP 88.5). ACR rewards placing all figures regardless of relevance: Gemini 3 Pro reaches 0.95 by placing every extracted figure on slides whose headlines are topic phrases (AHR 0.03). Figure 4 illustrates these two cases: SlideTailor (Zeng et al., 2026) minimizes TDP to 1.0 with nearly empty slides, and Gemini 3 Pro maximizes ACR while producing no assertion headlines. We therefore read TQR, ACR, and TDP only jointly with AHR and FCA; Sec. 5.3 tests whether they correlate with human judgment.

### 5.3 Do the Metrics Track Human Judgment?

**Preference study.** Graduate students in machine learning or computer vision rated 180 decks (30 per system, drawn from 30 papers sampled from the test split), blinded to system identity, with two raters per deck, on one 5-point Likert scale (1 = poor, 5 = excellent) rating the suitability of the deck for a conference presentation. We use a single overall score because sub-scales for design, coherence, and fidelity are highly correlated and the correlation analysis below needs sample size more than sub-scale resolution. Mean ratings are: SciSlide 3.8, PPTAgent (Zheng et al., 2025) 3.5, SlideTailor (Zeng et al., 2026) 3.3, Gemini 3 Pro 3.1, AutoPresent (Ge et al., 2025) 2.9, and PreGenie (Xu et al., 2025) 2.7. In pairwise comparisons formed by matching SciSlide with each of the five baselines on the same 30 papers (150 pairs), SciSlide is preferred in 65% of cases (95% CI $[57\%, 72\%]$; inter-rater weighted $\kappa = 0.68$).

**Per-deck correlation.** For each metric we compute the Spearman correlation $\rho$ between its per-deck value and the human score. Pooled correlation is computed over all 180 decks (30 papers $\times$ 6 systems); within-system correlation uses system-controlled partial correlation on the 150 baseline decks (30 papers $\times$ 5 systems, excluding SciSlide), and compares each metric with AHR by a Steiger test (numerical values in Appendix F). Pooled across systems, AHR ($\rho = 0.58$, 95% CI $[0.47, 0.67]$, $n = 180$) and VTR (0.44, $[0.31, 0.56]$) correlate most strongly with the human score. Within systems ($n = 150$, Fisher-$z$ transformation), AHR remains the strongest (0.38, $[0.23, 0.51]$), followed by VTR (0.29, $[0.13, 0.43]$) and FCA (0.27, $[0.11, 0.42]$). TQR (0.08 pooled, 0.05 within) and TDP (0.02 pooled, $-0.02$ within) do not correlate with human judgment: the upper ends of their within-system intervals (0.21 and 0.14) lie below the lower end of AHR's (0.23), and both differences are significant ($z = 3.3$ and 4.0, $p < .01$). These are the two metrics that can be optimized by outputs no reader would accept, TQR by verbatim copying and TDP by empty slides, which supports the vector reading of Sec. 5.2 empirically: a scalar aggregate would average metrics

that track human judgment with metrics that do not. ACR correlates weakly (0.20 pooled, 0.13 within) because Gemini 3 Pro attains the highest ACR by placing figures regardless of relevance; FCA, which checks caption–headline agreement, correlates nearly twice as strongly (0.40 pooled, 0.27 within). At $n = 150$ a true $\rho$ of 0.20 is detected with probability of about 0.70, so we state the TQR and TDP results as differences from AHR rather than as an absence of correlation.

### 5.4 Layout: Output Representation, Training Signals, and Correction

On a matched 7B backbone, replacing the four-way classification head with autoregressive coordinate generation raises strict exact match from 8.33% to 14.53%, a factor of 1.74 (Table 4 in Appendix D). The 14B configuration ($r$=32, 2.0 epochs) reaches 18.1%; raising rank to 64 and epochs to 3.0 brings it to 22.7%. The first comparison isolates the effect of model scale (14B vs. 7B) at matched rank and training budget, yielding a gain of 3.6 percentage points; the remaining 4.6 points come from the higher-capacity LoRA and longer training. Relaxing each coordinate to a tolerance of one grid cell raises accuracy to 52.3% at 7B and 61.8% at 14B, so about two fifths of the elements that miss exact match at 7B, and half at 14B, are within one cell on all four coordinates; the remaining errors are larger.

**Two auxiliary training signals reduce accuracy.** Both are compared with plain cross-entropy on the 7B backbone at matched rank and epochs (8.96%). A differentiable overlap and out-of-bounds penalty on relaxed coordinates ($\lambda = 0.5$) lowers exact match to 3.38%. Cross-entropy has a single minimizer, the ground-truth layout, whereas the penalty is minimized by every non-overlapping in-bounds layout, and the training layouts already satisfy it; the two objectives therefore disagree over most of the output space (Appendix D). Prepending retrieved layouts during training gives 8.71%: the model learns to follow the retrieved layout rather than the content, and its accuracy drops further when no close reference exists at inference. We therefore use retrieval only at inference; retrieval without generation reaches 5.2% (Appendix E).

**Inspection and page-local correction.** The corrector raises AHR by 0.07 in the first round and by 0.03 in the second, and reaches a fixed point by the third (median two rounds). The first round rewrites 18% of pages; OOB drops from 0.032 to 0.002, and Overlap from 0.011 to 0.002. No round lowered a metric that was above its threshold at the start of that round: among 1,127 page rewrites across all rounds and metrics, zero induced a regression. Because the loop optimizes the metrics it inspects, its gains are not independent evidence of better decks; the human study of Sec. 5.3 provides that evidence.

## 6 Conclusion

We posed scientific paper-to-slide generation as reconstruction of the paper's argument under assertion–evidence constraints rather than as extractive summarization into a template, instantiated this formulation in SciSlide, and showed that the properties it targets can be computed by deterministic programs. A source-grounded argument graph makes coverage and figure placement checkable against the paper; the same metrics score a deck, trigger its correction, and filter training data; and the only learned component, layout, gains more from an autoregressive discrete output representation than from auxiliary geometric or retrieval signals. The metrics are informative as a vector but not as a scalar, since some of them are optimized by degenerate outputs; designing metrics that are not optimized by such outputs is the main open problem for this form of evaluation.

**Limitations.** Each metric is a heuristic implementation of its definition: AHR misclassifies imperative and elliptical headlines, ACR counts assets without checking their placement, and the $n$-gram metrics (TQR, Ret, Hall) score a faithful paraphrase as untraceable; Appendix B reports agreement with hand-computed values. The metrics do not cover speaking time, audience adaptation, or delivery. Extraction and planning quality is bounded by the language model, which we have not varied. The corpus consists of English papers from machine-learning and computer-vision venues, and transfer to other fields is untested. Metric redundancy and pretraining contamination of the layout backbone are discussed in Appendix J.

### AI USE STATEMENT

In this work, we used generative AI tools to aid and polish writing, to draft portions of sections of the paper, for retrieval and discovery of related work and the verification of related-work claims, and for research execution, namely code development including the baseline evaluation pipeline. We have not used generative AI tools for research ideation: the research questions, method, corpus design, and metric definitions were determined by the authors, and no reported experimental result was produced by a generative AI tool. Generative AI is not used as a judge anywhere in our evaluation, which is computed entirely by deterministic programs. Generative AI is a component of the described system: Gemini 3 Pro performs argument extraction and content planning (Sec. 3.3), and Claude Sonnet 4.6 performs page-local correction (Sec. 3.4). We have reviewed all AI-assisted work: AI-assisted text was checked line by line by the authors, every related-work claim was verified against the cited source, and AI-assisted code was reviewed and tested by the authors, with the metric implementations checked against hand-computed values (Appendix B.2). We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

### ETHICS STATEMENT

The corpus consists of conference slide decks authored by third parties and released on public venue or author pages. We use them for research on presentation generation and do not redistribute any deck unless its license permits it; we will release the derived annotations and extraction code unconditionally, and the slide content only for decks whose license allows redistribution. The human study involved volunteer graduate-student raters who gave informed consent, were told the purpose of the study, and rated anonymized decks; no personal data was collected. Automatic slide generation could be used to produce presentations that misrepresent the findings of a paper; the source-grounding metrics (SSG-c, TQR, Hall) make such misrepresentation detectable but do not prevent it.

### REPRODUCIBILITY STATEMENT

Sec. 4 describes corpus construction and Sec. 5.1 the data split, models, and hardware. Appendix A defines every metric and lists every threshold, all fixed on the validation split before any test run, and Appendix B gives pseudocode for the metrics whose implementation is not determined by the definition, together with their agreement with hand-computed values. Appendix D lists all layout training configurations, and Appendix I the baseline settings. All metrics are computed by deterministic programs, so the reported values are reproducible given the same decks. Generation calls commercial APIs (Gemini 3 Pro, Claude Sonnet 4.6) whose outputs are not guaranteed to be bitwise reproducible across model versions. We will release the corpus annotations, the extraction and generation code, and the evaluation scripts.

## REFERENCES

Michael Alley. *The craft of scientific presentations: Critical steps to succeed and critical errors to avoid*. Springer, 2003.

Lukas Blecher, Guillem Cucurull Preixens, Thomas Scialom, and Robert Stojnic. Nougat: Neural optical understanding for academic documents. In *International Conference on Learning Representations*, volume 2024, pp. 37646–37663, 2024.

Isabel Cachola, Silviu Cucerzan, Allen Herring, Vuksan Mijovic, Erik Oveson, and Sujay Kumar Jauhar. Knowledge-centric templatic views of documents. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 15460–15476, 2024.

Haodong Chen, Xuanhe Zhou, Wei Zhou, Xinyue Shao, Yanbing Zhu, Bo Wang, Jiawei Hong, Anya Jia, and Fan Wu. X+ slides: Benchmarking audience-conditioned slide generation. *arXiv preprint arXiv:2606.19256*, 2026a.

Ting Chen, Saurabh Saxena, Lala Li, David J Fleet, and Geoffrey Hinton. Pix2seq: A language modeling framework for object detection. *arXiv preprint arXiv:2109.10852*, 2021.

Xin-Sheng Chen, Jiayu Zhu, Pei-lin Li, Hanzheng Wang, Shuojin Yang, and Meng-Hao Guo. Presentbench: A fine-grained rubric-based benchmark for slide generation. *arXiv preprint arXiv:2603.07244*, 2026b.

Zhiyao Cui, Chenxu Wang, Shuyue Hu, Yiqun Zhang, Wenqi Shao, Qiaosheng Zhang, and Zhen Wang. Design first, code later: Aesthetically pleasing template-free slides generation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 30470–30490, 2026.

datalab-to. Marker, 2024. URL `https://github.com/datalab-to/marker`.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115, 2023.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*, 2024.

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

Joanna Garner and Michael Alley. How the design of presentation slides affects audience comprehension: A case for the assertion-evidence approach. *International Journal of Engineering Education*, 29(6):1564–1579, 2013.

Jiaxin Ge, Zora Zhiruo Wang, Xuhui Zhou, Yi-Hao Peng, Sanjay Subramanian, Qinyue Tan, Maarten Sap, Alane Suhr, Daniel Fried, Graham Neubig, et al. Autopresent: Designing structured visuals from scratch. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2902–2911. IEEE, 2025.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025.

Kamal Gupta, Justin Lazarow, Alessandro Achille, Larry S Davis, Vijay Mahadevan, and Abhinav Shrivastava. Layouttransformer: Layout generation and completion with self-attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1004–1014, 2021.

Hsiao Yuan Hsu, Xiangteng He, Yuxin Peng, Hao Kong, and Qing Zhang. Posterlayout: A new benchmark and approach for content-aware visual-textual presentation layout. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6018–6026, 2023.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Yue Hu and Xiaojun Wan. Ppsgen: Learning-based presentation slides generation for academic papers. *IEEE transactions on knowledge and data engineering*, 27(4):1085–1097, 2014.

Naoto Inoue, Kotaro Kikuchi, Edgar Simo-Serra, Mayu Otani, and Kota Yamaguchi. Layoutdm: Discrete diffusion model for controllable layout generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10167–10176. IEEE, 2023.

Daesik Jang, Morgan Lindsay Heisler, Linzi Xing, Yifei Li, Edward Wang, Ying Xiong, Yong Zhang, and Zhenan Fan. Deckbench: Benchmarking multi-agent frameworks for academic slide generation and editing. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 9159–9170, 2026.

Zhaoyun Jiang, Jiaqi Guo, Shizhao Sun, Huayu Deng, Zhongkai Wu, Vuksan Mijovic, Zijiang James Yang, Jian-Guang Lou, and Dongmei Zhang. Layoutformer++: Conditional graphic layout generation via constraint serialization and decoding space restriction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18403–18412, 2023.

Di Jin and Peter Szolovits. Hierarchical neural networks for sequential sentence classification in medical scientific abstracts. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 3100–3109, 2018.

Kyudan Jung, Hojun Cho, Jooyeol Yun, Soyoung Yang, Jaehyeok Jang, and Jaegul Choo. Talk to your slides: High-efficiency slide editing via language-driven structured data manipulation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 3370–3399, 2026.

Min-Yen Kan. Slideseer: A digital library of aligned document and presentation pairs. In *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, pp. 81–90, 2007.

Xiang Kong, Lu Jiang, Huiwen Chang, Han Zhang, Yuan Hao, Haifeng Gong, and Irfan Essa. Blt: Bidirectional layout transformer for controllable layout generation. In *European Conference on Computer Vision*, pp. 474–490. Springer, 2022.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, et al. Tulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*, 2024.

John Lawrence and Chris Reed. Argument mining: A survey. *Computational linguistics*, 45(4): 765–818, 2020.

Guy Lev, Michal Shmueli-Scheuer, Jonathan Herzig, Achiya Jerbi, and David Konopnicki. Talksumm: A dataset and scalable annotation method for scientific paper summarization based on conference talks. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2125–2131, 2019.

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.

Jianan Li, Jimei Yang, Aaron Hertzmann, Jianming Zhang, and Tingfa Xu. Layoutgan: Generating graphic layouts with wireframe discriminators. *arXiv preprint arXiv:1901.06767*, 2019.

Xin Liang, Xiang Zhang, Yiwei Xu, Siqi Sun, and Chenyu You. Slidegen: Collaborative multimodal agents for scientific slide generation. *arXiv preprint arXiv:2512.04529*, 2025.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Jiawei Lin, Jiaqi Guo, Shizhao Sun, Zijiang Yang, Jian-Guang Lou, and Dongmei Zhang. Layoutprompter: Awaken the design ability of large language models. *Advances in Neural Information Processing Systems*, 36:43852–43879, 2023.

Chengzhi Liu, Yuzhe Yang, Kaiwen Zhou, Zhen Zhang, Yue Fan, Yanan Xie, Peng Qi, and Xin Eric Wang. Presenting a paper is an art: Self-improvement aesthetic agents for academic presentations. In *The Fourteenth International Conference on Learning Representations*, 2026. URL `https://openreview.net/forum?id=8NXCwNjFNR`.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 2511–2522, 2023.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel S Weld. S2orc: The semantic scholar open research corpus. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 4969–4983, 2020.

Patrice Lopez. Grobid: Combining automatic bibliographic data recognition and term extraction for scholarship publications. In *ECDL*, volume 9, pp. 473–474, 2009.

Qianli Ma, Jipeng Xiao, Siyu Wang, Zhiheng Tian, Wangyu Feng, Shibo Wang, Chang Guo, Shuochen Chang, Qingyang Liu, and Zhipeng Zhang. Omnipresent: Generating coherent presentation suites from scientific papers. *arXiv preprint arXiv:2607.02590*, 2026.

Richard E. Mayer. *Multimedia learning*. Cambridge University Press, 3 edition, 2021.

Richard E Mayer and Roxana Moreno. Nine ways to reduce cognitive load in multimedia learning. *Educational psychologist*, 38(1):43–52, 2003.

Alexander Meier, Mahei Manhai Li, and Roman Rietsche. Developing a hybrid vector-graph retrieval system for entity-preserving and inspiring storyline creation of presentation slides. In *ECIS*, 2025.

Tarik Can Ozden, Sachidanand VS, Furkan Horoz, Ozgur Kara, Junho Kim, and James Matthew Rehg. Narrative-driven paper-to-slide generation via arcdeck. *arXiv preprint arXiv:2604.11969*, 2026.

Yiming Pan, Chengwei Hu, Xuancheng Huang, Can Huang, Mingming Zhao, Yuean Bi, Xiaohan Zhang, Aohan Zeng, and Linmei Hu. Aeslides: Incentivizing aesthetic layout in llm-based slide generation via verifiable rewards. *arXiv preprint arXiv:2604.22840*, 2026.

Arjun Panickssery, Samuel R Bowman, and Shi Feng. Llm evaluators recognize and favor their own generations. *Advances in Neural Information Processing Systems*, 37:68772–68802, 2024.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th annual meeting of the association for computational linguistics: system demonstrations*, pp. 101–108, 2020.

Yuting Qiang, Yanwei Fu, Yanwen Guo, Zhi-Hua Zhou, and Leonid Sigal. Learning to generate posters of scientific papers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

Hiroaki Sakoe and Seibi Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE transactions on acoustics, speech, and signal processing*, 26(1):43–49, 1978.

Athar Sefid, Prasenjit Mitra, Jian Wu, and C Lee Giles. Extractive research slide generation using windowed labeling ranking. In *Proceedings of the Second Workshop on Scholarly Document Processing*, pp. 91–96, 2021.

Chengxin Shi, Qinnan Cai, Zeyuan Chen, Long Zeng, Yibo Zhao, Jing Yu, Jianxiang Yu, and Xiang Li. Apex: Academic poster editing agentic expert. *arXiv preprint arXiv:2601.04794*, 2026.

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

John Sweller. Cognitive load during problem solving: Effects on learning. *Cognitive science*, 12(2): 257–285, 1988.

Zecheng Tang, Chenfei Wu, Juntao Li, and Nan Duan. Layoutnuwa: Revealing the hidden layout expertise of large language models. In *International Conference on Learning Representations*, volume 2024, pp. 38369–38388, 2024.

Simone Teufel and Marc Moens. Summarizing scientific articles: experiments with relevance and rhetorical status. *Computational linguistics*, 28(4):409–445, 2002.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, et al. Large language models are not fair evaluators. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 9440–9450, 2024.

Sida Wang, Xiaojun Wan, and Shikang Du. Phrase-based presentation slides generation for academic papers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

Mann William and Sandra Thompson. Rhetorical structure theory: Towards a functional theory of text organization. *Text*, 8(3):243–281, 1988.

Xiaojie Xu, Xinli Xu, Sirui Chen, Haoyu Chen, Fan Zhang, and Ying-Cong Chen. PreGenie: An agentic framework for high-quality visual presentation generation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 3045–3063, 2025.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengpeng Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2025.

Ming Yang, Zhiwei Zhang, Jiahang Li, Haoseng Liu, Yuzheng Cai, and Weiguo Zheng. Deepslide: From artifacts to presentation delivery. *arXiv preprint arXiv:2605.15202*, 2026.

Wenzheng Zeng, Mingyu Ouyang, Langyuan Cui, and Hwee Tou Ng. Slidetailor: Personalized presentation slide generation for scientific papers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 34584–34592, 2026.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

Bo Zhao, Maosheng Pang, Chen Zhang, Huan Yang, Yixin Cao, and Wei Ji. Unipptbench: A unified benchmark for presentation generation across diverse input settings. *arXiv preprint arXiv:2605.17356*, 2026.

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.

Hao Zheng, Guozhao Mo, Xinru Yan, Qianhao Yuan, Wenkai Zhang, Xuanang Chen, Yaojie Lu, Hongyu Lin, Xianpei Han, and Le Sun. Deeppresenter: Environment-grounded reflection for agentic presentation generation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 31545–31558, 2026.

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. *Advances in neural information processing systems*, 36:46595–46623, 2023.

Min Zhou, Chenchen Xu, Ye Ma, Tiezheng Ge, Yuning Jiang, and Weiwei Xu. Composition-aware graphic layout gan for visual-textual presentation designs. *arXiv preprint arXiv:2205.00303*, 2022.

## A Metric Definitions

For the comparable metrics defined over $\mathcal{V}$ or $\mathcal{E}$ (NC@budget, CER, DTW-Flow, Ret), the graph is the one extracted from the source paper $P$ by the procedure of Sec. 3.3; it is the same for every system, and a node counts as realized when some content item of the deck contains an 8-gram of its span (Algorithm 5), so no system needs to emit node identifiers. The two invariant metrics take the value $1.00$ for SciSlide by construction (Appendix C); their comparable counterparts are TQR and FCA. Every free parameter ($\tau = 30$ words; $n = 8$ for TQR and Hall; idf-weighted Jaccard $\geq 0.25$ for FCA; IoU $> 0.30$ for Overlap; $m = 5$ for NC@budget, the median number of content items per slide in the human-made source decks of the validation split) was fixed on the validation split before any test-set run.

## B Metric Implementation and Validation

This appendix gives the computation for the five metrics whose implementation is not determined by the definition in Appendix A (Sec. B.1) and reports each implementation against values computed by hand (Sec. B.2). The remaining metrics—BLR, TDP, VTR, DTW-Flow, OOB, and the two construction invariants—follow their definitions directly.

Table 2: Deterministic evaluation metrics, computed from the source paper $P$, the argument graph $\mathcal{G}$, and the rendered deck $D$. "Scope" records whether the metric is computed by one program for every system (CMP) or is defined over $\mathcal{G}$ and therefore reported only for SciSlide as a construction invariant (INV). All are heuristic implementations of the stated definition; Sec. 6 lists their known failure modes and Appendix B their validation. ↑/↓: higher/lower is better; $K$ is the number of slides.

| Metric | Scope | Definition |
|---|---|---|
| *Assertion–evidence criteria* | | |
| AHR ↑ | CMP | fraction of headlines parsed as a finite declarative clause |
| BLR ↓ | CMP | fraction of slides containing a $\geq$3-item bullet list |
| TDP ↓ | CMP | $\frac{1}{K}\sum_k \max(0,\ w_k - \tau)$, $w_k$ body words, $\tau$=30 |
| VTR ↑ | CMP | area(figures+tables+equations) / area(all content) |
| FCA ↑ | CMP | fraction of placed assets whose caption tokens overlap the same slide's headline tokens with idf-weighted Jaccard $\geq 0.25$ |
| VEGI-c ↑ | INV | fraction of placed assets whose mounted node is realized by the slide carrying the asset |
| *Argument integrity* | | |
| NC@budget ↑ | CMP | fraction of $\mathcal{V}$ realized by $\geq$1 content item, after truncating every deck to $m$ content items per slide |
| CER ↑ | CMP | fraction of $\mathcal{E}$ with both endpoints realized, in order |
| DTW-Flow ↓ | CMP | DTW distance, slide role sequence vs. topological order |
| TQR ↑ | CMP | fraction of content-bearing text spans in $D$ matching some $\geq$8-gram of $P$ verbatim |
| SSG-c ↑ | INV | fraction of assertions whose node span is found verbatim in $P$ |
| ACR ↑ | CMP | $\min(1, (\#\text{cited} + \#\text{deck images})/\lvert\mathcal{A}\rvert)$ |
| *Geometry and content fidelity* | | |
| Overlap ↓ | CMP | fraction of same-slide leaf-element pairs with IoU $> 0.30$; one leaf per text frame, picture, or extracted PDF block; groups contribute their children and not themselves |
| OOB ↓ | CMP | fraction of elements whose box exceeds the canvas |
| Ret ↑ | CMP | fraction of realized-node source tokens retained in $D$ |
| Hall ↓ | CMP | fraction of deck content $n$-grams not traceable to $P$ |

### B.1 PSEUDOCODE

Throughout, $\mathrm{norm}(\cdot)$ is one token normalizer: NFKC, case folding, removal of punctuation other than intra-word hyphens, rejoining of words split across a line break, and whitespace collapse. It is applied to the paper and to the deck by the same function, so a mismatch cannot arise from asymmetric preprocessing.

**Algorithm 1** AHR: is a headline an assertion?

**Require:** headline string $h$; dependency parser Stanza 1.5.1, model en_ewt
**if** $h$ is empty **or** $|\mathrm{tokens}(h)| < 3$ **then return false**
**end if**
$T \leftarrow \mathrm{parse}(h)$ ▷ parsed as one sentence; no sentence splitting
$r \leftarrow \mathrm{root}(T)$
**if** $\mathrm{upos}(r) \notin \{\text{VERB}, \text{AUX}\}$ **then**
    **return false** ▷ topic phrase: "Experimental Results"
**end if**
**if** $\mathrm{VerbForm}(r) \in \{\text{GER}, \text{INF}, \text{PART}\}$ **and** $r$ has no `aux` child **then**
    **return false** ▷ "Improving Representation Quality"
**end if**
**if** no child of $r$ has `deprel` $\in \{$`nsubj`, `nsubj:pass`, `csubj`$\}$ **then**
    **return false** ▷ verb without a subject
**end if**
**if** $h$ ends with '?' **then return false**
**end if**
**return true**

AHR is the fraction of the $K$ headlines for which Algorithm 1 returns true. A slide with no headline counts as false rather than being excluded, so a deck cannot raise AHR by omitting headlines.

**Algorithm 2** TQR and Hall: $n$-gram traceability to the source

**Require:** paper $P$, deck $D$, order $n = 8$
$t_P \leftarrow \mathrm{norm}(\mathrm{tokens}(P))$
$\mathcal{H} \leftarrow \{\, \mathrm{hash}(t_P[i{:}i{+}n]) \;:\; 0 \le i \le |t_P| - n \,\}$ ▷ rolling hash, built once per paper
$S \leftarrow$ content-bearing text spans of $D$ ▷ headlines and body items; excludes page numbers, running heads, and author/venue furniture
**for all** $s \in S$ **do**
  $t_s \leftarrow \mathrm{norm}(\mathrm{tokens}(s))$
  **if** $|t_s| < n$ **then**
    $\mathrm{traced}(s) \leftarrow [\, t_s$ is a contiguous subsequence of $t_P \,]$
  **else**
    $\mathrm{traced}(s) \leftarrow [\, \exists i : \mathrm{hash}(t_s[i{:}i{+}n]) \in \mathcal{H} \,]$
  **end if**
**end for**
$\mathrm{TQR} \leftarrow |\{s \in S : \mathrm{traced}(s)\}| \,/\, |S|$
$G \leftarrow$ all $n$-grams of $\mathrm{norm}(\mathrm{tokens}(D))$
$\mathrm{Hall} \leftarrow |\{g \in G : \mathrm{hash}(g) \notin \mathcal{H}\}| \,/\, |G|$

TQR is span-level and Hall is $n$-gram-level so that a single fabricated clause inside an otherwise copied span registers in Hall without making the whole span untraceable.

**Algorithm 3** FCA: caption–headline alignment of a placed asset

**Require:** slide $k$ with headline $h_k$ and placed assets $A_k$
$w(t) \leftarrow \log\big(N/(1 + \mathrm{df}(t))\big)$ ▷ df over the 7,500-deck corpus, computed once and frozen
**for all** $a \in A_k$ **do**
  $C \leftarrow \mathrm{content}(\mathrm{norm}(\mathrm{caption}(a)))$ ▷ drops stopwords, the words `figure`/`table`, and bare numerals
  $H \leftarrow \mathrm{content}(\mathrm{norm}(h_k))$
  $J \leftarrow \dfrac{\sum_{t \in C \cap H} w(t)}{\sum_{t \in C \cup H} w(t)}$
  $\mathrm{aligned}(a) \leftarrow [\, J \ge 0.25 \,]$
**end for**
$\mathrm{FCA} \leftarrow |\{a : \mathrm{aligned}(a)\}| \,/\, |\text{placed assets in } D|$

Dropping `figure`/`table` and numerals is necessary: otherwise every caption shares the token "figure" with any headline that mentions one, and FCA increases on decks that merely name their figures.

**Algorithm 4** Overlap: leaf extraction and pairwise IoU

**Require:** the shape tree of one slide; CONTAINER $=$ {`GROUP`, `PLACEHOLDER`, `GRAPHIC_FRAME` whose child is a table, `CONTENT_PART`}
$L \leftarrow \emptyset$
**for all** shapes $s$ in depth-first order **do**
  **if** $\mathrm{type}(s) \in$ CONTAINER **then**
    **continue** ▷ descend into children; the container itself is never a leaf
  **end if**
  **if** $\neg\, \mathrm{visible}(s)$ **or** $\mathrm{area}(s) = 0$ **then continue**
  **end if**
  $L \leftarrow L \cup \{\mathrm{bbox}(s)\}$ ▷ one leaf per text frame regardless of paragraph count; bbox in slide coordinates after the enclosing groups' transforms
**end for**
$\mathrm{Overlap} \leftarrow \big|\{(a,b) \in \binom{L}{2} : \mathrm{IoU}(a,b) > 0.30\}\big| \,/\, \big|\binom{L}{2}\big|$

Two properties of the leaf set determine the value. A text frame is one leaf whatever its paragraph count: emitting one entry per paragraph, each carrying the box of the frame, makes every pair inside

a bulleted frame report an IoU of 1. And $L$ is rebuilt per slide: comparing elements across a whole deck makes the title frame of every slide overlap the title frame of every other slide in a templated deck, which alone drives the rate above 0.9. For decks delivered as PDF rather than PPTX there is no shape tree; we take one leaf per text block and per image block of the extracted page, which is the closest available analogue of a frame, although the two granularities are not identical.

**Algorithm 5** NC@budget: node coverage under a per-slide content budget

**Require:** deck $D$, graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, budget $m$
1: $m \leftarrow 5$ ▷ median content items per slide in human-made source decks of the validation split; fixed before any test run
2: $D' \leftarrow D$ with each slide truncated to its first $m$ content items ▷ reading order: top-to-bottom, then left-to-right; the headline is never counted against the budget
3: **for all** $v \in \mathcal{V}$ **do**
4: $\quad \mathrm{realized}(v) \leftarrow [\,\exists$ content item $c \in D'$ such that $c$ contains an 8-gram of $\mathrm{span}(v)\,]$
5: **end for**
6: $\mathrm{NC@budget} \leftarrow |\{v : \mathrm{realized}(v)\}| \,/\, |\mathcal{V}|$

The budget is needed because coverage is otherwise monotone in the amount of text: a system that copies the paper onto a few dense slides covers every node without presenting any of them. Matching on the node span rather than on a `node_id` makes the metric comparable across systems. CER applies the same $\mathrm{realized}(\cdot)$ predicate to both endpoints of each edge and additionally requires the two realizing slides to appear in the direction of the edge.

### B.2 Agreement with Hand-Computed Values

Each implementation is compared against values computed by hand on a sample drawn from the held-out split. Slide- and element-level decisions are scored by Cohen's $\kappa$ against a single annotator; continuous quantities are scored by mean absolute deviation on the scale of the metric.

Table 3: Agreement between each implementation and hand-computed values, on $n = 200$ units per metric drawn from the held-out split and scored by one annotator who did not see the program output.

| Metric | Unit compared | Statistic | $n$ | Value |
|---|---|---|---|---|
| AHR | headline | Cohen's $\kappa$ | 200 | 0.85 |
| BLR | slide | Cohen's $\kappa$ | 200 | 0.91 |
| TDP | slide | mean $\lvert\Delta\rvert$ (words) | 200 | 1.6 |
| VTR | slide | mean $\lvert\Delta\rvert$ (area) | 200 | 0.08 |
| FCA | placed asset | Cohen's $\kappa$ | 200 | 0.74 |
| NC@budget | graph node | Cohen's $\kappa$ | 200 | 0.82 |
| CER | graph edge | Cohen's $\kappa$ | 200 | 0.71 |
| DTW-Flow | deck | Spearman $\rho$ of ranks | 200 | 0.88 |
| TQR | text span | Cohen's $\kappa$ | 200 | 0.94 |
| ACR | deck | mean $\lvert\Delta\rvert$ | 200 | 0.06 |
| Overlap | element pair | Cohen's $\kappa$ | 200 | 0.70 |
| OOB | element | Cohen's $\kappa$ | 200 | 0.97 |
| Ret | $n$-gram | mean $\lvert\Delta\rvert$ | 200 | 0.05 |
| Hall | $n$-gram | Cohen's $\kappa$ | 200 | 0.86 |

This table is what supports the claim that the values in Table 1 measure the definitions in Appendix A; a deterministic program guarantees reproducibility but not validity. Agreement is lowest for Overlap ($\kappa = 0.70$), CER (0.71), and FCA (0.74). For FCA, one known source of error is a caption parse that fails silently when the converter merges a caption into the body text.

## C Construction Invariants of SciSlide

Two criteria are satisfied by SciSlide as a property of the representation rather than as a measured level of performance, and we report them separately so that they are not read as a comparison. SSG-

c, the fraction of assertions whose node span occurs verbatim in $P$, is $1.00$, because Sec. 3.3 rejects and re-extracts any node failing that test. VEGI-c, the fraction of placed assets whose mounted node is realized by the slide carrying the asset, is $1.00$ up to assets lost in rendering, because content planning places an asset on the slide that realizes its node. These invariants are the concrete content of verifiable grounding: two classes of failure, an unsourced claim and a figure on the wrong slide, are excluded by construction rather than reduced in frequency. No baseline produces an argument graph, so neither invariant has a baseline counterpart; the comparable forms are TQR and FCA in Table 1.

We also report design-attribute prediction as a system property. The corpus of Sec. 4 is the only paper-to-slide resource that retains font-size bin, alignment, emphasis, color, and theme, and SciSlide predicts them with the following per-attribute accuracies on the test split: font-size bin $0.68$ (majority baseline $0.42$), alignment $0.82$ ($0.55$), emphasis $0.78$ ($0.49$), color $0.67$ ($0.38$), theme $0.71$ ($0.35$); mean accuracy $0.73$ against a mean majority baseline of $0.44$. Template-based and editing-based systems inherit these attributes and have no prediction to score.

## D Layout Training Configurations

Table 4: Layout accuracy on the test split. "exact" is strict bounding-box exact match, the fraction of elements satisfying $\mathbb{I}[\hat{x}{=}x \wedge \hat{y}{=}y \wedge \hat{w}{=}w \wedge \hat{h}{=}h]$ on the $24 \times 13$ grid; "$\pm1$" relaxes each equality to a tolerance of one grid cell. The 7B rows isolate the output representation; the 14B rows separate scale from rank and training duration.

| Configuration | Backbone | exact | $\pm1$ grid |
|---|---|---|---|
| 4-way classification head | Qwen2.5-7B | 8.33% | — |
| Autoregressive ($r$=16, 1.5ep) | Qwen2.5-7B | 8.96% | 44.2% |
| Autoregressive ($r$=32, 2.0ep) | Qwen2.5-7B | 14.53% | 52.3% |
| Autoregressive + geometry loss | Qwen2.5-7B | 3.38% | 28.7% |
| Autoregressive + retrieved prior | Qwen2.5-7B | 8.71% | 45.8% |
| Autoregressive ($r$=32, 2.0ep) | Qwen2.5-14B | 18.1% | 57.4% |
| Autoregressive ($r$=64, 3.0ep), deployed | Qwen2.5-14B | 22.7% | 61.8% |

**Why the geometry penalty fails.** The penalty is computed on a continuous relaxation of coordinates that the model emits as discrete tokens. Cross-entropy over coordinate tokens has a single minimizer, the ground-truth layout, whereas the penalty is minimized by the whole set of non-overlapping, in-bounds configurations, which is far larger and contains the target only incidentally. The two objectives disagree over most of the output space, and the gradient of the penalty pulls against the token likelihood instead of refining it. The property the penalty encodes is already present in the supervision, because the training layouts are non-overlapping and in-bounds.

## E Component Ablations

A3 and A4 isolate the two representational components. Removing the assertion and role labels leaves the same model and corpus but no supervision that distinguishes an assertion headline from a topic phrase; AHR falls from $0.86$ to $0.36$, close to the $0.34$ of the human-made source decks, and BLR rises from $0.09$ to $0.53$. Removing the argument-graph conditioning leaves slide order to follow the section order of the paper: NC@budget falls to $0.44$, below AutoPresent (Ge et al., 2025), PPTAgent (Zheng et al., 2025), and PreGenie (Xu et al., 2025) in Table 1, and DTW-Flow rises to $4.3$, equal to the worst baseline, Gemini 3 Pro.

A7 is reported at the same exact-match granularity as Table 4. A nearest-neighbor cache that copies a retrieved layout can score well under a soft overlap criterion while being wrong about every individual element, because the retrieved slide rarely has the same number and types of elements as the target. Under strict exact match it reaches $5.2\%$, below the weakest autoregressive configuration ($8.96\%$): retrieval supplies an initial structure, and adapting it to the specific content requires generation.

Table 5: Component ablations. Each row removes one component from SciSlide and reports the change on the metric that component targets, using the definitions of Appendix A; "exact" is the strict bounding-box exact match of Table 4. A7 is a standalone retrieval-only configuration rather than a removal.

| Abl. | Removed component | Effect |
|---|---|---|
| A1 | autoregressive coords → classification head (7B) | exact $14.53\% \to 8.33\%$ |
| A2 | Qwen2.5 → Qwen1.5 backbone (matched rank, epochs) | exact $-3.3$ pt |
| A3 | assertion and role labels | AHR $0.86 \to 0.36$, BLR $0.09 \to 0.53$ |
| A4 | argument-graph conditioning | NC@budget $0.85 \to 0.44$, DTW-Flow $1.4 \to 4.3$ |
| A5 | retrieved exemplars (inference few-shot) | exact $-2.8$ pt |
| A6 | retrieved prior (training signal) | exact $8.96\% \to 8.71\%$ |
| A7 | retrieval only, no generation | exact $5.2\%$, OOB $0.02$ |

# F Human Study and Metric Correlations

Table 6: Rank correlation between each deterministic metric and the human overall score. "Pooled" computes $\rho$ over all 180 decks (6 systems); "Within" uses system-controlled partial correlation on the 150 baseline decks. "$z$ vs. AHR" is a Steiger test of the difference between the within-system $\rho$ of the metric and that of AHR. "Monotone" records whether the metric orders the five baselines consistently with their mean human ratings (SciSlide excluded); entries denote the baseline that breaks the order. OOB is excluded because PPTAgent (Zheng et al., 2025) and SlideTailor (Zeng et al., 2026) score near-zero on every deck.

| Metric | Pooled $\rho$ | Pooled CI | Within $\rho$ | $z$ vs. AHR | Monotone |
|---|---|---|---|---|---|
| AHR | 0.58 | $[0.47, 0.67]$ | 0.38 | — | partly |
| VTR | 0.44 | $[0.31, 0.56]$ | 0.29 | 1.2 (n.s.) | no (Gemini, PreGenie) |
| FCA | 0.40 | $[0.26, 0.52]$ | 0.27 | 1.5 (n.s.) | partly |
| NC@budget | 0.35 | $[0.21, 0.48]$ | 0.24 | 2.0 ($p<.05$) | yes |
| DTW-Flow | $-0.30$ | $[-0.43, -0.15]$ | $-0.22$ | 2.8 ($p<.01$) | partly |
| BLR | $-0.25$ | $[-0.39, -0.10]$ | $-0.18$ | 3.1 ($p<.01$) | no |
| ACR | 0.20 | $[0.05, 0.34]$ | 0.13 | 3.6 ($p<.001$) | no (Gemini) |
| TQR | 0.08 | $[-0.07, 0.23]$ | 0.05 | 3.8 ($p<.001$) | no |
| TDP | 0.02 | $[-0.13, 0.17]$ | $-0.02$ | 4.6 ($p<.001$) | no (SlideTailor) |

Table 6 reports the Spearman correlation between each metric and the human overall score, computed both pooled across systems and within each system (partial correlation controlling for system identity). The pooled correlation reflects two sources of variance: differences between systems and differences between decks produced by the same system. The within-system correlation isolates the latter, which is the variance that each metric explains after accounting for which system produced a deck. Both are reported because they answer different questions: pooled correlation measures whether a metric tracks human preference in a deployment setting where the user sees outputs from multiple systems; within-system correlation measures whether a metric provides a signal for optimizing a single system. The Steiger test in the rightmost column compares the within-system $\rho$ of each metric with that of AHR, testing whether the difference is statistically significant on the same 150 baseline decks.

**Statistical power.** For within-system correlation, $n = 150$ (the number of baseline decks); the $95\%$ interval around a zero correlation is about $\pm 0.16$. At $n = 100$ it would be $\pm 0.20$, wide enough that a measured $\rho$ of $0.08$ and a true $\rho$ of $0.25$ could not be distinguished. Narrowing it to $\pm 0.10$ would require $385$ decks. This is why the TQR and TDP results are stated as comparisons with AHR: a Steiger test on dependent correlations is considerably more powerful than a test against zero.

**Within-system versus pooled correlation.** As argued in the opening of this appendix, pooled correlation conflates variance across systems with variance within systems. The within-system partial correlation isolates the latter by controlling for system identity. AHR retains the highest correlation (0.38 within vs. 0.58 pooled), and its advantage over TQR and TDP remains significant ($z = 3.3$ and 4.0, $p < .01$), which confirms that the vector reading is not an artifact of pooling.

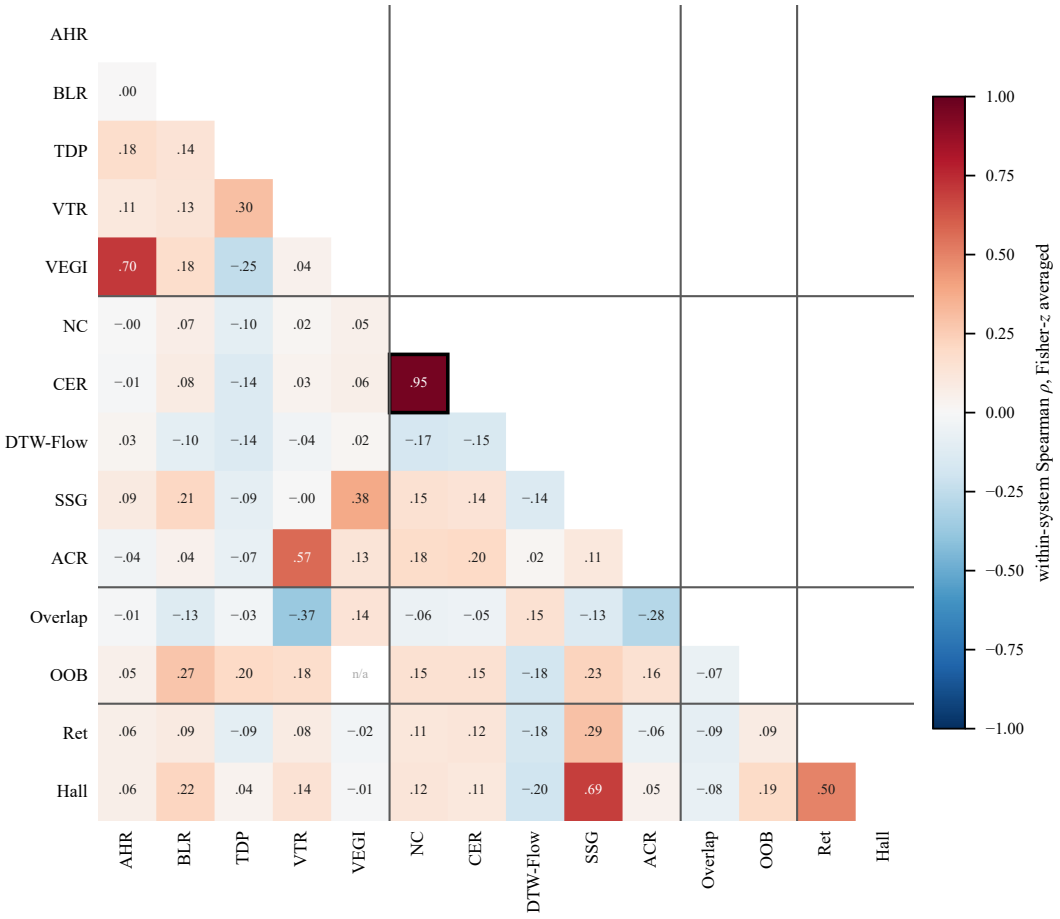



Figure 5: Pairwise rank correlation between metrics, computed within each baseline system and averaged over the five baselines (150 decks: 30 papers × 5 systems, the baseline outputs from the human-evaluation subset). Metrics for which lower is better are sign-flipped, so a positive entry means that the two coordinates move together in the direction of quality. Rules separate the metric families of Sec. 3.2; the only pair above $|\rho| = 0.95$ is boxed. SSG and VEGI are the per-deck span-grounding and visual-grounding scores computed against the graph extracted from each source paper, not the construction invariants SSG-c and VEGI-c; NC is node coverage without the per-slide budget. Overlap and OOB follow Algorithm 4. FCA and TQR are not included in the matrix.

**Inter-metric correlation.** The metrics are reported as a vector, which is justified only if its coordinates carry distinct information. We test this on baseline decks only, which do not involve SciSlide and therefore cannot be selected in its favor, and show the result in Figure 5. The correlation is computed within each system and then averaged, not over the pooled decks: the five systems differ strongly in their metric means (AHR from 0.03 to 0.70, TDP from 1.0 to 88.5), so a rank correlation over pooled decks mostly records which system produced a deck. We therefore compute Spearman $\rho$ separately per system, convert it to Fisher $z$, average with weights $n_s - 3$, and transform back; a cell is left empty where fewer than two systems have enough variance in both metrics.

The matrix is mostly pale: the mean absolute correlation is $0.15$, and only four pairs exceed $|\rho| = 0.5$. The two geometry metrics are nearly independent of the others; the largest entry is Overlap with VTR ($-0.37$). One pair exceeds $0.95$, as expected from the definitions: NC and CER ($\rho = 0.95$) differ only in whether edge order is additionally required, so they contribute one measurement rather than two. The next two pairs are coupled mechanically rather than redundant. VEGI and AHR ($0.70$) are coupled because a deck with no assertion headline has no assertion for a visual to support, so VEGI is zero wherever AHR is zero. SSG and Hall ($0.69$) are the same $n$-gram traceability test counted in opposite directions.

AHR and BLR could be expected to be nearly redundant, since both concern whether the headline or a bullet list carries the propositional content. The measured correlation is $0.00$, and no individual system exceeds $|\rho| = 0.18$. AHR tests the syntax of the headline and BLR the form of the body, and a deck can satisfy one without the other—a full-sentence headline above a bullet list satisfies AHR and fails BLR—so both coordinates are kept.

Within systems, SSG, the per-deck span-grounding score that is closest to TQR among the metrics in the matrix, and TDP correlate with AHR at $0.09$ and $0.18$; pooling would give $0.34$ and $0.37$. Both therefore vary largely independently of AHR within a system as well, consistent with their weak correlation with human judgment in Sec. 5.3.

## G Qualitative Comparison

Figure 6 shows three consecutive slides from the same paper. The baseline (top row) uses topic-phrase headlines ("Related Work," "Experimental Setup") above bullet lists; SciSlide (bottom row) replaces them with assertion headlines ("Our method builds on three prior frameworks," "We evaluate on five benchmarks") above the supporting figures and tables. The per-slide metric values beneath each slide show that SciSlide achieves higher AHR, lower TDP, and higher VTR, consistent with the assertion–evidence paradigm of Alley (2003). The baseline's higher TQR (0.81 vs 0.63) reflects verbatim copying rather than rewriting nodes as assertions, which is the design choice that enables SciSlide to produce declarative headlines. Appendix H lists the failure modes observed in the human evaluation.




*CVPR2023_20960: slides 6--8 of each deck, consecutive and at the same position*

Figure 6: Three consecutive slides generated from the same source paper by PPTAgent (Zheng et al., 2025), the strongest editing-based baseline (top row), and by SciSlide (bottom row), at identical scale, with the per-slide metric values beneath each slide. Topic-phrase headlines above bullet lists are replaced by single-sentence assertions whose supporting figure or table occupies the body.

## H Failure Modes

SciSlide performs poorly on four kinds of input, each attributable to a specific stage: argument nodes whose supporting evidence is spread across several paper sections, which content planning splits

across slides and which therefore lower CER; papers whose figures carry no informative caption, for which asset attachment has no signal to match and FCA falls; sections with no extractable causal structure, such as related work, which the topological ordering places arbitrarily; and layouts with many small elements, where exact coordinate prediction degrades and page-local correction cannot recover the intended grouping. The metric vector of Table 1 aggregates over decks and does not show which stage fails on which input, which is why these cases are listed. The first is a limitation of the one-assertion-per-slide constraint rather than of its implementation.

## I Baseline Configurations

The Gemini 3 Pro baseline (temperature $1.0$, max tokens $8192$) is accessed through its public API. The prompt instructs the model to produce a conference presentation from the paper, with no explicit mention of the assertion–evidence paradigm; it is reported as the tool currently available to a researcher preparing a talk, not as a controlled comparison. AutoPresent (Ge et al., 2025), PPTAgent (Zheng et al., 2025), SlideTailor (Zeng et al., 2026), and PreGenie (Xu et al., 2025) were run from their released code with the defaults distributed by their authors, on the same PDF input and the same figure-extraction output as SciSlide, with figures supplied to each system in the form its interface accepts. All baselines were evaluated on the same 300-paper test sample as SciSlide.

## J Extended Threats to Validity

**Metric redundancy.** The twelve metrics of Table 1 are not independent. Appendix F reports the inter-metric correlation matrix over all baseline outputs; where two metrics exceed $\rho = 0.95$ they should be read as one measurement.

**Geometry.** The Overlap and OOB rows show SciSlide reaching competitive performance. Every system except AutoPresent (Ge et al., 2025) overlaps on less than $2\%$ of same-slide element pairs. The OOB rate of SciSlide after correction ($0.002$) is the lowest among all systems; PPTAgent (Zheng et al., 2025) and SlideTailor (Zeng et al., 2026) inherit valid geometry from their templates ($0.008$ and $0.005$), and PreGenie (Xu et al., 2025) and Gemini 3 Pro are at $0.019$ and $0.007$. The rows show that a model emitting coordinates directly, combined with page-local correction, achieves valid geometry comparable to template-based systems.

**Pretraining contamination.** Our splits are disjoint by venue, but the pretraining corpus of the backbone is not, and both the source papers and some public decks predate it; a model that has seen a paper could recall its deck rather than predict its layout. Two observations bound this effect: the reported metric is exact match on a $24 \times 13$ grid produced by our own discretization, which appears in no public artifact, and the discriminative head of Sec. 5.4 shares the same backbone and exposure yet reaches only $8.33\%$. On the subset of 127 test papers published after the backbone's training cutoff (June 2024), exact match is $22.3\%$, compared with $22.9\%$ on the 173 papers published before; the $0.6$-point difference is within sampling error and does not suggest memorization.