

SlidesGen-Bench: Evaluating Slides Generation via Computational and Quantitative Metrics

Yunqiao Yang^{1*} Wenbo Li^{1*} Houxing Ren^{1*†} Zimu Lu¹ Ke Wang¹
Zhiyuan Huang² Zhuofan Zong¹ Mingjie Zhan^{2‡} Hongsheng Li^{1,3‡}

¹CUHK MMLab, ²SenseTime Research, ³CPII under InnoHK
yangyunqiao7@gmail.com zhanmingjie@sensetime.com hsli@ee.cuhk.edu.hk
<https://slidesgenbench.yqy314.top>

Abstract

The rapid evolution of Large Language Models (LLMs) has fostered diverse paradigms for automated slide generation, ranging from template-based file generation and code-driven layouts to image-centric synthesis. However, evaluating these heterogeneous systems remains challenging, as existing protocols often struggle to provide comparable scores across architectures or rely on uncalibrated judgments. In this paper, we introduce **SlidesGen-Bench**, a structured evaluation protocol that assesses slide generation along three practically important dimensions: *content fidelity*, *visual aesthetics*, and *editability*. To compare heterogeneous systems under one framework, all outputs are projected into the visual domain for content and aesthetic assessment, while editability is evaluated through format-conditional structural parsing of the generated files. Each dimension is quantified with structured, verifiable pipelines rather than monolithic LLM judgment. For the highly subjective aesthetic dimensions, we construct the Slides-Align1.5k human preference dataset and show that our computational metrics align with human judgment substantially better than standard LLM-as-a-Judge baselines. Our code and data are released to support further research.

1 Introduction

Driven by recent breakthroughs in Large Language Models (LLMs) (OpenAI, 2023; Touvron et al., 2023a,b; Dubey et al., 2024; Bai et al., 2023; Jiang et al., 2023; Anthropic, 2024; Yang et al., 2024) and the rapid evolution of code agents (Wang et al., 2024b; Dong et al., 2025; Wang et al., 2024c), researchers have increasingly explored the use of LLM-based agents for automated slide generation. This surge in interest has led to a wide array of

research initiatives (Ge et al., 2025; Zheng et al., 2025; Tang et al., 2025; Jung et al., 2025; Yang et al., 2025; Bandyopadhyay et al., 2024) and commercial applications. Current presentation generation approaches typically follow one of three paradigms: (1) Template-based Generation, which populates pre-defined schemas; (2) Code-driven Layouts, which utilize generated markup to render slide structures; and (3) Image-centric Generation, which synthesizes entire slide pages directly.

To evaluate automatically generated slides, prior work has proposed several pipelines that generally fall into two categories: reference-based comparisons (Ge et al., 2025) and LLM-as-a-Judge frameworks (Zheng et al., 2025). Reference-based metrics rely heavily on source files or target images, restricting their applicability in open-ended generation scenarios. Conversely, while LLM-based evaluation offers flexibility, it is susceptible to inherent model behaviors—such as verbosity bias (Saito et al., 2023; Wang et al., 2024a) or reasoning failures (Krumdick et al., 2025; Thakur et al., 2025). The significant diversity in system architectures and output formats introduces three core challenges for systematic assessment: (1) achieving comparable evaluation across diverse architectures; (2) conducting a comprehensive and quantitative assessment; and (3) verifying the reliability and rationality of the evaluation scheme.

To this end, we introduce **SlidesGen-Bench**, a benchmark designed to quantitatively evaluate slide generation by tackling these challenges. First, we leverage the observation that the final output of current slide generators is a visual rendering. By grounding our aesthetic and content evaluation in the image domain, our framework establishes a comparable baseline across paradigms, while carefully maintaining a format-conditional hierarchy to evaluate underlying structural editability. Second, we evaluate slides across Content, Aesthetics, and Editability using structured, verifiable

*Equal contribution.

†Project Lead.

‡Corresponding author.

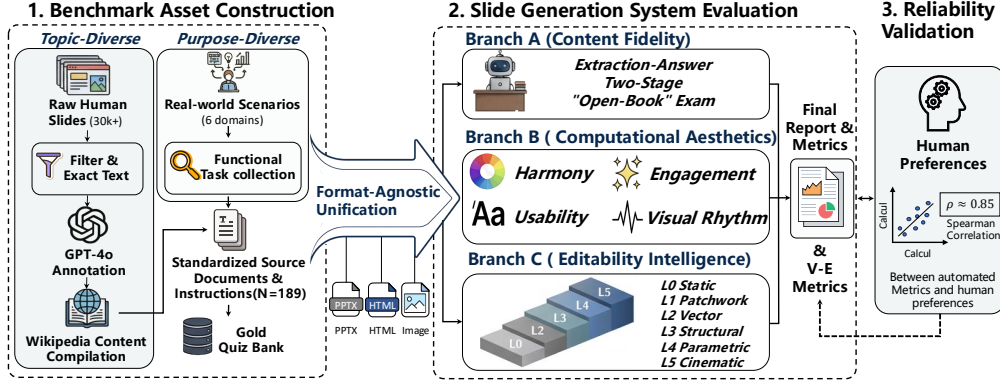


Figure 1: The main pipeline of SlidesGen-Bench.

pipelines. Rather than relying on a single monolithic LLM prompt (Zheng et al., 2025), we decouple content evaluation into a structured Vision-Language Model (VLM) extraction and question-answering pipeline, and ground aesthetic evaluation in cognitive and visual ergonomics formulas. Finally, to validate the reliability of our scheme, we apply a divide-and-conquer strategy. While the Content and Editability branches are grounded in a structured pipeline and verifiable structural metadata—yielding systematic and quantifiable metrics that enable consistent comparisons across disparate models—we address the inherent subjectivity of aesthetics through human alignment. Specifically, we incorporate human alignment studies on Slides-Align1.5k. This human preference-aligned dataset covers slides from nine mainstream generation systems across seven scenarios. Our experiments demonstrate a high consistency between our automated aesthetic metrics and human ratings. The whole pipeline is shown in Figure 1.

In summary, our contributions are as follows:

- We propose SlidesGen-Bench, a structured benchmark covering content fidelity, visual aesthetics, and editability. By grounding content and aesthetic evaluation in the visual domain and evaluating editability through format-conditional structural parsing, the protocol yields comparable scores across template-based, code-driven, and image-centric systems under one framework.
- We construct Slides-Align1.5k, a human preference-aligned dataset covering slides from nine mainstream generation systems, bridging the gap between automated metrics and perceptual quality.
- Comprehensive experiments demonstrate that our structured evaluation pipeline aligns with

human preferences substantially better than standard LLM-as-a-Judge protocols, with robustness analyses (bootstrap confidence intervals, oracle-context QA, and style-transfer validation) supporting the reliability of the reported scores.

2 SlidesGen-Bench

To achieve comparable evaluation across diverse architectures (Challenge 1) and quantitative assessment (Challenge 2), SlidesGen-Bench projects all outputs into consistent visual renderings and structural schemas before evaluation. Concretely, content and aesthetic evaluation operate on rendered visual outputs, while editability is measured through format-conditional structural parsing of the generated files, allowing template-based PPTX, HTML/code-generation, and image-centric systems to be compared under one framework. The pipeline comprises data curation (Sec. 2.1) and three evaluation branches (Sec. 2.2–2.4). Furthermore, to ensure the reliability of these metrics (Challenge 3), we systematically validate them against human preferences (Sec. 2.5).

2.1 Data collection

We curate our benchmark based on two primary dimensions: *topics* and *purposes*. Topic-based instructions assess the model’s ability to process diverse content types and generate relevant multimodal elements, such as images and illustrations. Purpose-based instructions evaluate the capacity to control stylistic coherence and produce visually appealing presentations tailored to specific scenarios and audiences. Constructing the dataset through this dual lens enables a systematic examination of a generative tool’s comprehension of requirements and its proficiency in slide creation.

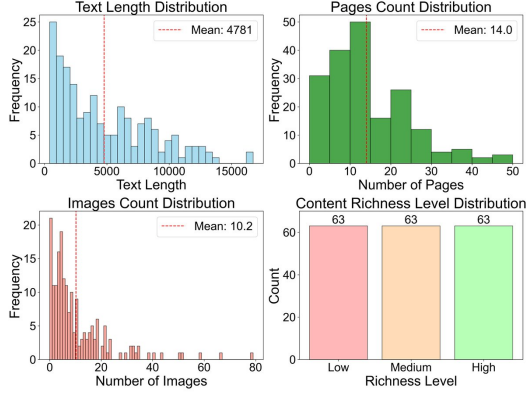


Figure 2: Statistical distribution of content metrics.

Determining the Data Format. We choose document-based inputs (comprehensive outlines with required text and images) as our primary data format to test the models’ long-context summarization and formatting stability.

Topic-Diverse Instruction Collection. To ensure instructional diversity, we initially collected approximately 30k human-authored slides and templates from various sources. We applied a length filter to exclude decks shorter than 5 pages or longer than 40 pages. We employed GPT-4o to annotate the slides regarding their topics and purposes. The detailed topic distribution analysis is provided in Appendix A. Subsequently, we compiled detailed source documents for each topic to serve as input. We leveraged Wikipedia to gather core content, including detailed text, hierarchical subtitles, and image descriptions, while preserving the original structural integrity. Filtering invalid or sparse entries yielded a final set of 94 instructions.

Purpose-Diverse Instruction Collection. The practical utility of automated slide generation depends on its ability to adapt to different contexts. To evaluate this, we expanded our benchmark to encompass functional tasks, not just different topics. We collected another 95 instructions across six real-world scenarios: brand promotion, business plans, course preparation, personal statements, product launches, and work reports. These categories cover a wide range of goals, from the narrative style of personal statements to the logical structure of business plans. While the final benchmark consists of 189 filtered instructions, each is a highly complex, document-level synthesis task (mean: 4,525 chars, requiring multi-modal image integration; Figure 2).

Statistics of the Instructions. As illustrated in Figure 2, the curated dataset exhibits significant

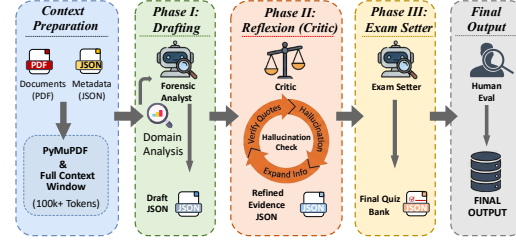


Figure 3: The QuizBank construction pipeline.

diversity in text length (mean: 4,525 chars), page count, and image count. To quantitatively categorize difficulty, we introduce a Content Richness metric, stratifying the dataset into three balanced levels (Low, Medium, High). Detailed formulations and distributions are provided in Appendix B. While the final benchmark consists of $N = 189$ filtered instructions, each input represents a highly complex, document-level synthesis task. Although this scale might appear modest compared to standard text-only NLP benchmarks, it is a deliberate design choice driven by the high operational cost and complexity of end-to-end multimodal slide generation. A detailed justification is provided in Appendix C.

2.2 Content Evaluation

To assess salient information retention, we propose a structured VLM-based QuizBank evaluation framework. Given that presentation slides represent a condensed abstraction of source documents, ensuring critical information preservation is a primary challenge. Rather than relying on a single black-box LLM prompt to judge “content quality” (Zheng et al., 2025), we construct a reference QuizBank derived directly from the source text. We then employ a decoupled Vision-Language Model (VLM) pipeline to extract text from generated slides and answer the questions. While this approach fundamentally relies on the capabilities of the underlying VLM, decoupling extraction from reasoning minimizes hallucination artifacts inherent in standard LLM-as-a-Judge approaches.

QuizBank Construction. The construction of the QuizBank is a multi-agent pipeline designed to extract ground-truth knowledge with high precision. As illustrated in Figure 3, the process moves through three distinct phases:

Phase I: Drafting. The process uses the Forensic Analyst agent, scanning the full context window (100k+ tokens) of the raw PDF documents. Its objective is to identify high-value information seg-

ments to produce a draft JSON of key points.

Phase II: Critic. This phase runs a cyclic feedback loop of hallucination checks, quote verification, and gap expansion, producing a Refined Evidence JSON.

Phase III: Exam Setting. Finally, the Exam Setter agent converts the refined evidence into a structured assessment. To ensure an equitable and length-agnostic evaluation across all baseline models, the system dynamically samples a standardized probe of exactly 10 Multiple-Choice Questions (MCQs) per document: 5 Concept Questions focus on high-level understanding, while 5 Data Questions focus on specific details. This standardized length effectively prevents evaluation bias toward overly verbose generations and strictly tests core summarization abilities. In the end, all questions were manually verified by authors to ensure they cannot be answered using general world knowledge alone.

QuizBank Evaluation. To quantify information preservation, we employ a VLM-based “open-book” exam protocol. Instead of directly feeding slide images to the VLM to generate answers, we adopt a two-stage pipeline to decouple visual content extraction from semantic comprehension. This design enables us to accurately isolate whether an incorrect answer stems from a visual extraction failure or a comprehension shortfall. We first parse all slide images into structured Markdown via a VLM. Then, the VLM is constrained by the prompt to answer QuizBank questions relying solely on this slide-derived context. Accuracy rates serve as a proxy for content quality. Our Error Analysis in Appendix Q shows that only 1.2% of errors stem from VLM extraction failures, ensuring the robustness of our evaluation against VLM bottlenecks. More details are in Appendix D.

2.3 Aesthetics Evaluation

To quantitatively evaluate the aesthetic quality of presentation slides against globally recognized professional standards, we introduce a four-dimensional evaluation framework grounded in cognitive load theory and visual ergonomics. Unlike traditional image quality assessment (IQA) metrics that treat images in isolation, our framework operates directly on the rendered visual domain, remaining architecture-agnostic and modeling the presentation as a temporal sequence that assesses both the *spatial quality* (single slide) and the *temporal coherence* (deck pacing). We focus

on four core dimensions: Harmony, Engagement, Usability, and Visual Rhythm. The parameter decision process is shown in Appendix F. Detailed diagnostic patterns and examples are provided in Appendix I.

Harmony Score. We compute a normalized per-slide harmony score $S_{slide}^{(i)}$ by optimizing its fit to hue templates in HSV space (Cohen-Or et al., 2006). To ensure consistency between our main formulation and the empirical deck-level calibration (detailed in Appendix F.5), the overall Harmony Score aggregates the mean harmony μ_{deck} and penalizes cross-slide inconsistency (standard deviation σ_{deck}):

$$S_{harmony} = w_m \times \mu_{deck} - w_d \times \sigma_{deck} \quad (1)$$

Engagement Score. We adapt Hasler and Suesstrunk’s colorfulness metric (Hasler and Suesstrunk, 2003) using opponent channels $rg = R - G$ and $yb = \frac{1}{2}(R + G) - B$:

$$M_{slide} = \sqrt{\sigma_{rg}^2 + \sigma_{yb}^2} + 0.3\sqrt{\mu_{rg}^2 + \mu_{yb}^2} \quad (2)$$

We further define a deck-level pacing score from the standard deviation of slide colorfulness σ_{pacing} :

$$\text{Score}_{pacing} = e^{-\frac{(\sigma_{pacing} - \mu_{target})^2}{2w^2}} \quad (3)$$

Usability Score. We evaluate figure-ground contrast within detected text regions using a layout analysis model (Cui et al., 2025). Relative luminance L is computed with sRGB linearization. With contrast ratio $c = (L_{max} + 0.05)/(L_{min} + 0.05)$, we map contrast into $S_{contrast} \in [0, 1]$ via logarithmic normalization (mapping the 21 : 1 ratio to 1.0):

$$S_{contrast} = \frac{\ln(c)}{\ln(21)} \quad (4)$$

Visual Rhythm Score. We introduce Visual Heart Rate Variability (VisualHRV), combining per-slide clutter (Subband Entropy) and temporal variability (RMSSD). Subband Entropy is computed as: $E_{SE} = \frac{E_L + w_c(E_a + E_b)}{1 + 2w_c}$. For a raw entropy value x , calculate the normalized Score S (0 to 1): $S(x) = \frac{1}{1 + e^{-k \cdot (x - \mu)}}$. For the valid sequence of entropy-derived scores ($N \geq 5$, the length-filter criterion in Sec. 2.1), we compute the length-normalized RMSSD:

$$\text{RMSSD} = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N-1} (S_{i+1} - S_i)^2} \quad (5)$$

We then combine:

$$\text{Score}_{VHRV} = \lambda_1 \cdot \overline{S_{entropy}} + \lambda_2 \cdot \text{RMSSD} \quad (6)$$

Overall Aesthetics Score. The overall Aesthetics score is computed as the sum of its weighted component scores, integrating spatial and temporal evaluations into a metric.

$$\text{Score}_{total} = H_{score} + E_{score} + U_{score} + V_{score} \quad (7)$$

2.4 Editability Evaluation

Recent generative models produce visually convincing slides, yet their structural properties vary significantly depending on their design objectives. To quantify this, we introduce the Presentation Editability Intelligence (PEI) framework, a hierarchical scale from static visual mimicry to fully editable, narrative presentations. Levels 1–2 capture surface fidelity without true semantic structure, Levels 3–4 add consistent document organization and native, data-driven objects (e.g., editable charts), and Level 5 models temporal and multimedia dynamics so the deck functions as a directed experience. Systems are scored with a dependency-based “knockout” rule: failing a lower level precludes credit at higher ones. It is important to note that while PEI levels describe user-facing operational capabilities, our benchmark evaluates them fully automatically by directly parsing the underlying structural files (e.g., Office Open XML schemas via python-pptx and DOM trees for HTML), requiring zero human intervention. To make clustered systems easier to compare, we additionally report parser-derived subcriteria (text object editability, grouping, native chart/table editability, narrative structure, and animation/multimedia support; Appendix L) while keeping the main PEI levels unchanged. Note that PEI is format-conditional: it measures the editability features exposed by the generated file format and parser, so it captures operational editability rather than overall visual or semantic quality. Detailed definitions are provided in Appendix K.

2.5 Reliability Verification via Human

To systematically verify the rationality and reliability of our evaluation scheme (Challenge 3), we integrate a human-preference validation stage into our benchmark. We construct the **Slides-Align1.5k** dataset, a ground-truth testbed annotated by a diverse, standardized pool including CS Ph.D. stu-

dents, art undergraduates, and office clerks (Annotation details in Appendix P). We quantify the reliability of automated evaluation methods against this human ground truth using three core indicators: Average Spearman correlation, its Standard Deviation, and the Average Identical ratio.

Scope of the Metrics. The three branches provide different levels of evidence. QuizBank is a practical proxy for content retention (source-grounded questions, human audit of failures, oracle-context QA, backend-sensitivity checks; Sec. 3.3.2, Appendix R); the aesthetic metrics are human-aligned computational proxies validated on Slides-Align1.5k within the tested style distribution, not universal measures of visual quality (Appendix J); PEI is closer to a direct structural measurement because it is parsed from the generated file structures. This separation lets readers assign the right scope to each score rather than treating all components as equally direct measures of quality.

3 Experiments

3.1 Experimental Setup

Slides Generation Frameworks. We evaluate several prominent slides-generation frameworks, encompassing both commercial platforms and recent open-source academic pipelines. To ensure the evaluation reflects the most up-to-date capabilities, all commercial systems were tested between November and December 2025. Crucially, all generated artifacts from these evaluations have been open-sourced to facilitate reproducible research. We present the results as a dated SlidesGenBench v1.0 snapshot; system access details and the rerun/update protocol are documented in Appendix N.

The evaluated frameworks can be categorized into three paradigms. **(1) Source File Generation** directly outputs standard .pptx files. Commercial systems such as Gamma.com.ai, Quark-PPT (Quark, 2025) and Kimi-PPT (Moonshot AI, 2025) adopt a template-filling approach. In contrast, recent open-source frameworks employ code-centric manipulation strategies. To ensure these frameworks could be evaluated fairly under our zero-shot, end-to-end benchmark setting, we implemented necessary input adaptations. For instance, AutoPresent* (Ge et al., 2025) was augmented with an LLM pre-processing step to convert our raw PDF inputs into its required bullet-point format,

and its image backend was standardized. Similarly, PPTAgent* (Zheng et al., 2025) requires a reference template; we adapted it by providing a curated local template library and an LLM routing mechanism to dynamically select appropriate styles (detailed configurations in Appendix M). We denote these adapted pipelines with an asterisk (*) to distinguish them from their raw original versions.

(2) HTML Code Generation leverages LLMs to synthesize structural code (HTML/CSS/JS). Platforms like Zhipu PPT (Zhipu AI, 2025) and Skywork (Skywork AI, 2025) utilize this method to achieve superior typographic precision and layout flexibility; however, they face limitations regarding fidelity loss when converting to standard .pptx formats.

(3) Image Generation Models represent a shift toward visual synthesis, employing multi-modal models to generate slides directly as images. Exemplified by NotebookLM, Skywork-Banana Mode (Skywork AI, 2025) and Kimi-Banana Mode (Moonshot AI, 2025), this method achieves studio-grade design aesthetics often unattainable by standard templates, but treating slides as pixel data compromises text editability and necessitates additional OCR layers for accessibility.

3.2 Experimental Results

Content Results. The content quality of the slides is reflected by the accuracy of the QuizBank test. Table 1 presents the QuizBank accuracy results. Zhipu emerges as the top performer with an overall average of 88.29%, demonstrating broad competency across topics. Furthermore, it proves to be the most robust model for complex content, achieving the highest accuracy on both High and Medium difficulty questions. In terms of domain-specific strengths, Skywork-Banana leads in ‘Brand’ and ‘Report’ generation, while Kimi-Standard achieves near-perfect performance (96.47%) in ‘Personal’ topics. The data highlights a trend where ‘Business’ topics remain the primary bottleneck for current models, yielding the lowest average accuracy of 61.61% among commercial systems.

To gauge the stability of these comparisons, we computed 95% bootstrap confidence intervals (CIs) and the evaluated n for each system (Appendix R). The intervals are tight (typically $\pm 2-5\%$), and the top-ranked system’s lower bound remains at or above the runner-up’s upper bound, so the headline conclusions are not driven by small samples.

The evaluation is also robust to the VLM backend: an all-question comparison of DeepSeek-v4-Pro and MiniMax-M2.7 yields nearly identical accuracies (60.8% vs. 61.1%), and an oracle-context QA test, where the evaluator answers the same MCQs from ground-truth source evidence instead of slide context, reaches 100%. Observed failures thus primarily reflect generated-slide content loss or mismatch rather than evaluator error (a full-bank coverage and bias audit is in Appendix S).

Aesthetics Results. As shown in Table 3, Skywork-Banana outperforms all baselines in the aggregate evaluation. Notably, Skywork-Banana achieves the highest scores in the overall Aesthetics Score (27.28 vs. 26.58 for the runner-up) and Engagement (8.30 vs. 6.41), suggesting that Skywork-Banana is particularly effective at generating high-fidelity, visually engaging slides. Models with lower aesthetic scores, such as Quark and Auto-Present, have a relatively flat narrative pace in the generated slides, resulting in significantly lower Rhythm scores. The analysis of the aesthetics results is in Appendix O.

To examine style transferability, we ran a style-transfer validation over the 1,435 Slides-Align1.5k aesthetics records, covering six style buckets—standard corporate/mixed (66.8%), minimalist-light (14.8%), image-rich marketing (8.6%), diagrammatic/scientific (6.3%), data-heavy report (2.6%), and dark-mode minimalist (0.8%)—with full diagnostics in Appendix J. We find no systematic dark-mode penalty (dark decks receive slightly better human mean ranks than light decks, 3.84 vs. 4.27), and removing any single style bucket changes the composite human-alignment score by less than 0.05. Boundary cases (e.g., image-rich marketing decks and creative palettes) show larger rank errors, so we present these metrics as human-aligned proxies within the tested style distribution rather than universal measures of visual quality.

Human Alignment Results. Table 2 presents the evaluation on the Slides-Align1.5k dataset. This protocol achieved a robust inter-annotator agreement (Average Spearman $\rho = 0.85$, Std = 0.12), serving as a reliable gold standard. For baselines, we compare against PPT-Eval from PPTAgent (Zheng et al., 2025), as well as standard LLM-as-a-Judge paradigms (Rating and Arena) (Zheng et al., 2023). To ensure a strong upper bound, the LLM judges are powered by a state-of-the-art multimodal model (gemini-3-flash-preview), taking the

Product	Purpose Categories						Topic Intro.	Difficulty Level		
	Brand Promote	Business Plan	Personal Statement	Product Launch	Course Preparation	Work Report		High	Low	Med
Gamma	63.57	52.50	74.00	54.44	85.00	48.46	76.34	67.26	75.56	68.45
Kimi-Banana	84.29	77.50	96.00	90.43	84.00	85.38	86.22	83.33	91.58	86.67
Kimi-Smart	N/A	N/A	95.00	N/A	N/A	84.29	79.68	78.78	88.50	78.57
Kimi-Standard	N/A	N/A	96.47	N/A	N/A	80.00	76.88	75.71	88.05	75.43
NotebookLM	83.00	45.00	N/A	84.21	93.00	86.15	68.33	69.81	86.00	71.16
Quark	84.29	51.25	94.00	79.05	N/A	68.00	82.13	69.78	90.52	82.12
Skywork	80.00	68.75	78.50	76.52	84.67	86.15	83.12	80.00	82.58	81.31
Skywork-Banana	92.14	67.50	91.50	89.57	87.69	90.00	79.33	79.67	88.00	84.31
Zhipu	80.00	68.75	96.00	86.08	90.00	84.17	90.44	84.07	92.00	88.97
PPTAgent*	62.25	47.37	69.15	61.80	66.86	61.27	61.94	58.94	66.76	61.68
AutoPresent*	56.13	42.72	62.35	55.73	60.29	55.25	55.85	53.15	60.20	55.62
Average	76.19	57.93	85.30	75.31	81.44	75.37	76.39	72.77	82.70	75.84

Table 1: The QuizBank accuracy (%) for different products by topic and difficulty level. The *Purpose Categories* group contains six functional-task scenarios, while the *Topic Intro.* column reports topic-introduction instructions; N/A denotes task types not supported by the corresponding system (a true constraint rather than a missing evaluation).

Method	Spearman		Identical Avg(↑)
	Avg(↑)	Std (↓)	
SlidesGen-Bench	0.71	0.16	32.6
LLM-as-Judge Rating	0.57	0.23	20.7
LLM-as-Judge Arena	0.52	0.27	17.3
PPTAgent	0.53	0.26	17.8
Humans	0.85	0.12	45.3

Table 2: Human alignment results on SlidesGen-Bench.

rendered slide images as direct inputs with prompts mirroring our aesthetic criteria (Appendix V).

As shown in the table, SlidesGen-Bench demonstrates superior reliability across all metrics. It achieves the highest Average Spearman correlation of 0.71, significantly surpassing the runner-up LLM-as-Judge Rating (0.57) and PPTAgent (0.53). Furthermore, our method exhibits the most stable performance with the lowest Standard Deviation (0.16). Notably, in terms of the Average Identical ratio, our approach reaches 32.6%, delivering a substantial improvement over existing LLM-based evaluators (which struggle around 20%). These results verify that our computational methodology aligns more closely and robustly with human perceptual standards than existing protocols.

Editability Results. Table 4 reveals a distinct “Structural Barrier” in the landscape of presentation generation. While the majority of evaluated systems (e.g., Gamma, Skywork, Kimi-Smart) have mastered visual fidelity, achieving Level 2 (Vector Visual), they suffer from Structural Amnesia. These systems generate slides as isolated artistic canvases without global logic (e.g., $\langle p : sldMaster \rangle$), meaning layout changes cannot be propagated system-wide. Quark stands out as the sole system to breach the “Toy-to-Tool” threshold

(L3), demonstrating the capability to generate cohesive, grouped, and master-based hierarchies that support professional editing workflows. Although the Parametric Gap at Level 4 remains universally unsolved by commercial models, Open-source academic frameworks uniquely bridge these gaps: AutoPresent achieves L4 by natively generating programmable chart objects ($\langle c : chart \rangle$), while PPTAgent reaches L5 by utilizing raw templates as structural priors to inherit temporal dynamics (animations and slide transitions). Notably, PPTAgent’s L5 capability comes with a caveat: these inherited dynamic behaviors act as rigid priors and cannot be parametrically modified.

We also note that the current PEI levels cluster most commercial systems at L1–L2, reflecting limited discriminative granularity, with the substantive implication that many visually strong systems still expose limited structural editability under parseable file formats. To support finer-grained comparison of clustered systems, we report parser-derived subcriteria in Appendix L.

V-E Matrix Results. To quantify the interplay between aesthetics and functional utility, we map the evaluated systems onto the Visual-Editability (V-E) Matrix (Figure 4). Commercial systems dominate the left hemisphere (L0–L2), prioritizing visual polish over structural depth, while the lower-right quadrant (Q4) houses Quark (L3) and academic frameworks like AutoPresent (L4) and PPTAgent (L5), which remain confined there due to aesthetic limitations or reliance on unmodifiable rigid templates. Crucially, the upper-right quadrant (Q1) remains vacant: no system achieves peak performance in both dimensions simultaneously. This gap highlights the trade-off between visual expres-

Product	Usability (↑)	Engagement (↑)	Harmony (↑)	Rhythm (↑)	Aesthetics (↑)
Skywork-Banana	5.62	8.30	-0.47	13.84	27.28
Kimi-Banana	5.72	6.41	-0.55	15.01	26.58
NotebookLM	4.13	7.32	-0.35	11.72	22.82
Zhipu	4.87	7.52	-1.60	11.27	22.06
Skywork	4.83	7.60	-1.18	9.44	20.69
Kimi-Standard	4.61	6.28	-1.75	10.12	19.25
Kimi-Smart	4.13	7.99	-1.88	8.06	18.30
Gamma	5.31	6.31	-1.51	6.99	17.09
Quark	5.03	7.41	-1.91	6.33	16.86
PPTAgent*	5.01	6.12	-1.43	5.62	15.32
AutoPresent*	4.28	5.93	-2.15	4.57	12.63

Table 3: Slides Aesthetics evaluation results. We report detailed component scores (↑) including Usability, Engagement, Harmony, and Rhythm. The table is sorted by the Aesthetics score (sum of these components).

System	T1	T2	T3	T4	T5	PEI
PPTAgent*	✓	✓	✓	✓	✓	L5
AutoPresent*	✓	✓	✓	✓	✗	L4
Quark	✓	✓	✓	✗	✗	L3
Gamma	✓	✓	✗	✗	✗	L2
Skywork	✓	✓	✗	✗	✗	L2
Kimi (Standard)	✓	✓	✗	✗	✗	L2
Kimi (Smart)	✓	✓	✗	✗	✗	L2
Zhipu	✓	✓	✗	✗	✗	L2
Kimi (Banana)	✓	✗	✗	✗	✗	L1
Skywork (Banana)	✓	✗	✗	✗	✗	L1
NotebookLM	✗	✗	✗	✗	✗	L0

Table 4: **Evaluation Results against the PEI Framework.** Columns T1–T5 represent the hierarchical pass/fail criteria (✓/✗).

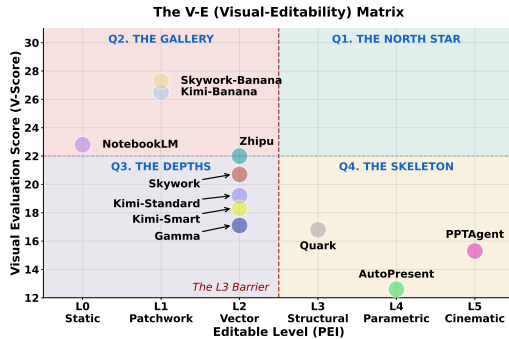


Figure 4: **The V-E (Visual-Editability) Matrix.**

siveness and structural depth in current methodologies, suggesting future opportunities: bridging static visual proxies and functionally rigid schemas to enable fully controllable, dynamic visual storytelling.

3.3 Ablation Studies

3.3.1 Analysis of the Aesthetics Metrics

To assess the individual contributions of our proposed aesthetic metrics, we conducted a comprehensive ablation study, as presented in Table 5. The parameters for our computational metrics were optimized on a completely independent verification set

of 1,000 slide pairs from human-generated slides (Detailed results are shown in Appendix F).

Metric Configuration	Spearman Avg(↑)	Std(↓)	Identical Avg(↑)
Only Engagement	0.224	0.349	14.8
Only Harmony	0.312	0.414	15.6
Only Usability	0.574	0.207	21.5
Only Visual HRV	0.618	0.198	24.4
Harmony + Engagement	0.371	0.440	20.7
Usability + Engagement	0.612	0.212	23.7
Usability + Harmony + Engagement	0.667	0.206	24.4
Full Method (All)	0.710	0.160	32.6

Table 5: Ablation study of aesthetic metrics.

Specifically, we observe that Usability serves as a strong baseline among the traditional heuristic metrics (Avg. Spearman 0.574), significantly outperforming Engagement (0.224) and Harmony (0.312) when used in isolation. This suggests that readability and visual clarity are the primary drivers of user preference in slide design.

Furthermore, the synergy between metrics is evident. Combining Harmony, Engagement, and Usability yields a correlation of 0.667. Yet, the Full Method, which includes all metrics, achieves the highest performance across all dimensions: an Average Spearman correlation of 0.710, the lowest standard deviation of 0.160, and an Average Identical ratio of 32.6%. This confirms that our multi-dimensional approach captures the complex interplay of perceptual ease and visual interest better than any single metric.

3.3.2 Analysis of Errors in QuizBank

We conduct a failure analysis on the 2,499 incorrect instances (19.1% of 13,023 total samples) to understand generation limitations. All incorrect instances were categorized by human annotators, who inspected the rendered slide images against the ground truth and assigned each case to the error taxonomy. A full taxonomy of error types

and model breakdowns is provided in Appendix Q. As shown in Table 20, *Missing Content* is the primary bottleneck, accounting for 66.3% of errors. This indicates that current models prioritize high-level visual narratives over the retrieval of granular facts (e.g., specific dates or metrics) required by the ground truth. *Content Value Mismatch* (31.7%) constitutes the secondary error source. Comparing model performance (Table 21, Appendix Q) reveals a distinct trade-off: top-performing systems (e.g., Zhipu, Kimi-Banana) significantly reduce content omission but exhibit higher rates of value mismatch (49–52%). This suggests that as models improve at retrieving detailed content, the challenge shifts from data recall to maintaining factual consistency.

4 Related work

Slides Evaluation. Existing methodologies predominantly fall into reference-based metrics and LLM-based assessments. Early approaches, such as SlideCoder (Tang et al., 2025) and parts of AutoPresent (Ge et al., 2025), frame evaluation as a reconstruction problem, calculating fidelity against ground-truth references. While quantifiable, this reliance on paired data precludes their use in open-ended generation where canonical references are absent. To mitigate this, recent frameworks like PPTAgent (Zheng et al., 2025) employ LLMs to judge content and aesthetics. However, these LLM-based evaluators introduce stochasticity, exhibiting susceptibility to bias and reasoning failures (Thakur et al., 2025; Krumdick et al., 2025). Critically, existing paradigms rarely incorporate a human alignment stage to validate metric reliability. This absence of correlation with human preference highlights the need for a consistent, reference-free evaluation standard that is both quantitatively robust and perceptually aligned.

Computational Aesthetics. Computational aesthetics has evolved from heuristic features to comprehensive frameworks like AIM (Oulasvirta et al., 2018) and deep learning models such as NIMA (Talebi and Milanfar, 2018). However, presentation slides constitute a unique multimodal domain requiring a distinct balance between high-contrast symbology and aesthetic cohesion, often rendering generic photography metrics insufficient. To address this, we distill evaluation into four dimensions that prioritize perceptual fluency. For visual appeal, we refine standard Color Harmony (Cohen-Or et al., 2006) via saturation weighting

and adapt Hasler’s metric (Hasler and Suesstrunk, 2003) into an Engagement Score to quantify vibrancy. Functionally, we adopt a Usability Score based on WCAG luminance (Caldwell et al., 2008) to ensure deterministic legibility, and utilize Sub-band Entropy (Rosenholtz et al., 2007) to model Visual HRV, providing a robust proxy for audience cognitive load.

5 Conclusion

In this work, we introduce SlidesGen-Bench, a standardized benchmark designed to systematically evaluate heterogeneous slide generation systems. By projecting diverse architectural outputs into a unified visual and structural space, we enable comparable, quantitative assessments across Content, Aesthetics, and Editability. Furthermore, validation on our novel Slides-Align1.5k dataset demonstrates that our structured computational metrics achieve significantly higher alignment with human preferences than existing evaluation pipelines. Ultimately, SlidesGen-Bench provides a reliable and reproducible framework to accelerate future research in automated presentation synthesis.

6 Limitations

Despite the contributions of SlidesGen-Bench, there are limitations of our current framework. First, our evaluation focuses exclusively on static visual content, overlooking temporal dynamics such as animations and slide transitions; this limitation is prevalent across existing evaluation protocols, highlighting a broader community challenge for dynamic presentation flows. Second, content evaluation depends on the chosen VLM extraction and question-answering configuration: QuizBank is a structured content-retention evaluation under a specified evaluator setup rather than evaluator-independent ground truth, and although our human audit, oracle-context QA, and backend-sensitivity checks bound this risk (Sec. 3.3.2, Appendix R), results may shift with different evaluator stacks. Third, PEI is format-conditional—it measures editability features exposed by the generated file format, so image- or web-output systems face a natural architectural ceiling unrelated to visual quality. Fourth, the aesthetic metrics carry cultural and domain assumptions: derived from visual-ergonomics principles and validated on our slide-preference distribution, they may need recalibration for styles outside it (Appendix J). Finally, Slides-Align1.5k

remains primarily English-centric and optimized for globalized professional communication (e.g., corporate and academic standards); future iterations will explore how varying cultural tolerances for visual density affect human preference (e.g., medical or legal reports).

Acknowledgement

This study was supported in part by the Centre for Perceptual and Interactive Intelligence (CPII) Ltd., a CUHK-led InnoCentre under the InnoHK initiative of the Innovation and Technology Commission of the Hong Kong SAR Government, and in part by NSFC-RGC Project N_CUHK498/24. Hongsheng Li is a PI of CPII under the InnoHK.

References

- Anthropic. 2024. [The claude 3 model family: Opus, sonnet, haiku](#).
- Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. [Qwen technical report](#). *CoRR*, abs/2309.16609.
- Sambaran Bandyopadhyay, Himanshu Maheshwari, Anandhavelu Natarajan, and Apoorv Saxena. 2024. Enhancing presentation slide generation by llms with a multi-staged end-to-end approach. *arXiv preprint arXiv:2406.06556*.
- Ben Caldwell, Michael Cooper, Loretta Guarino Reid, Gregg Vanderheiden, Wendy Chisholm, John Slatin, and Jason White. 2008. Web content accessibility guidelines (wcag) 2.0. *WWW Consortium (W3C)*, 290(1-34):5–12.
- Daniel Cohen-Or, Olga Sorkine, Ran Gal, Tommer Leyvand, and Ying-Qing Xu. 2006. Color harmonization. In *ACM SIGGRAPH 2006 Papers*, pages 624–630.
- Cheng Cui, Ting Sun, Manhui Lin, Tingquan Gao, Yubo Zhang, Jiaxuan Liu, Xueqing Wang, Zelun Zhang, Changda Zhou, Hongen Liu, et al. 2025. Paddleocr 3.0 technical report. *arXiv preprint arXiv:2507.05595*.
- Yihong Dong, Xue Jiang, Jiaru Qian, Tian Wang, Kechi Zhang, Zhi Jin, and Ge Li. 2025. A survey on code generation with llm-based agents. *arXiv preprint arXiv:2508.00083*.
- Nancy Duarte. 2010. *Resonate: Present visual stories that transform audiences*. John Wiley & Sons.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Jiaxin Ge, Zora Zhiruo Wang, Xuhui Zhou, Yi-Hao Peng, Sanjay Subramanian, Qinyue Tan, Maarten Sap, Alane Suhr, Daniel Fried, Graham Neubig, et al. 2025. Autopresent: Designing structured visuals from scratch. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2902–2911.
- David Hasler and Sabine E Suesstrunk. 2003. Measuring colorfulness in natural images. In *Human vision and electronic imaging VIII*, volume 5007, pages 87–95. SPIE.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. 2023. [Mistral 7b](#). *CoRR*, abs/2310.06825.
- Kyudan Jung, Hojun Cho, Jooyeol Yun, Soyoung Yang, Jaehyeok Jang, and Jaegul Choo. 2025. Talk to your slides: Language-driven agents for efficient slide editing. *arXiv preprint arXiv:2505.11604*.
- Wolfgang K  hler. 1967. Gestalt psychology. *Psychologische forschung*, 31(1):XVIII–XXX.
- Michael Krumdick, Charles Lovering, Varshini Reddy, Seth Ebner, and Chris Tanner. 2025. No free labels: Limitations of llm-as-a-judge without human grounding. *arXiv preprint arXiv:2503.05061*.
- Yiwen Luo and Xiaoou Tang. 2008. Photo and video quality evaluation: Focusing on the subject. In *European conference on computer vision*, pages 386–399. Springer.
- Moonshot AI. 2025. [Kimi slides: Ai-powered presentation creator](#).
- Peter O’Donovan, Aseem Agarwala, and Aaron Hertzmann. 2011. Color compatibility from large datasets. In *ACM SIGGRAPH 2011 papers*, pages 1–12.
- OpenAI. 2023. [GPT-4 technical report](#). *CoRR*, abs/2303.08774.
- Antti Oulasvirta, Samuli De Pascale, Janin Koch, Thomas Langerak, Jussi Jokinen, Kashyap Todi, Markku Laine, Manoj Krishnombuge, Yuxi Zhu, Aliaksei Miniukovich, et al. 2018. Aalto interface metrics (aim) a service and codebase for computational gui evaluation. In *Adjunct Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*, pages 16–19.

- Ruben Post, Tran Nguyen, and Paul Hekkert. 2017. Unity in variety in website aesthetics: A systematic inquiry. *International Journal of Human-Computer Studies*, 103:48–62.
- Quark. 2025. [Quark ai ppt: One-click intelligent presentation generator](#).
- Ruth Rosenholtz, Yuanzhen Li, and Lisa Nakano. 2007. Measuring visual clutter. *Journal of vision*, 7(2):17–17.
- Keita Saito, Akifumi Wachi, Koki Wataoka, and Youhei Akimoto. 2023. Verbosity bias in preference labeling by large language models. *arXiv preprint arXiv:2310.10076*.
- Skywork AI. 2025. [Skywork ai: Ai-powered workspace and presentation generator](#).
- John Sweller. 1988. Cognitive load during problem solving: Effects on learning. *Cognitive science*, 12(2):257–285.
- Hossein Talebi and Peyman Milanfar. 2018. Nima: Neural image assessment. *IEEE transactions on image processing*, 27(8):3998–4011.
- Wenxin Tang, Jingyu Xiao, Wenxuan Jiang, Xi Xiao, Yuhang Wang, Xuxin Tang, Qing Li, Yuehe Ma, Junliang Liu, Shisong Tang, et al. 2025. Slidecoder: Layout-aware rag-enhanced hierarchical slide generation from design. *arXiv preprint arXiv:2506.07964*.
- Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hupkes. 2025. Judging the judges: Evaluating alignment and vulnerabilities in llms-as-judges. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*, pages 404–430.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. [Llama: Open and efficient foundation language models](#). *CoRR*, abs/2302.13971.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. [Llama 2: Open foundation and fine-tuned chat models](#). *CoRR*, abs/2307.09288.
- Conny MA van Ravenswaaij-Arts, Louis AA Kollee, Jeroen CW Hopman, Gerard BA Stoeltinga, and Herman P van Geijn. 1993. Heart rate variability. *Annals of internal medicine*, 118(6):436–447.
- Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, et al. 2024a. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450.
- Xingyao Wang, Yangyi Chen, Lifan Yuan, Yizhe Zhang, Yunzhu Li, Hao Peng, and Heng Ji. 2024b. Executable code actions elicit better llm agents. In *Forty-first International Conference on Machine Learning*.
- Xingyao Wang, Boxuan Li, Yufan Song, Frank F Xu, Xiangru Tang, Mingchen Zhuge, Jiayi Pan, Yueqi Song, Bowen Li, Jaskirat Singh, et al. 2024c. Opendevin: An open platform for ai software developers as generalist agents. *arXiv preprint arXiv:2407.16741*, 3.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. [Qwen2.5 technical report](#). *CoRR*, abs/2412.15115.
- Yuheng Yang, Wenjia Jiang, Yang Wang, Yiwei Wang, and Chi Zhang. 2025. Auto-slides: An interactive multi-agent system for creating and customizing research presentations. *arXiv preprint arXiv:2509.11062*.
- Hao Zheng, Xinyan Guan, Hao Kong, Wenkai Zhang, Jia Zheng, Weixiang Zhou, Hongyu Lin, Yaojie Lu, Xianpei Han, and Le Sun. 2025. Pptagent: Generating and evaluating presentations beyond text-to-slides. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 14413–14429.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot

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

Zhipu AI. 2025. [Ai slides | chatglm intelligent presentation tool](#). Part of the ChatGLM/Zhipu Qingyan suite.

Appendix

A Dataset Distribution Details

Figure 5 illustrates the topic distribution of the 30k collected human-authored slides used to guide our instruction generation process.

Figure 6 depicts the topic distribution of the curated instructions.

Table 6 shows the domain distribution of the instructions.

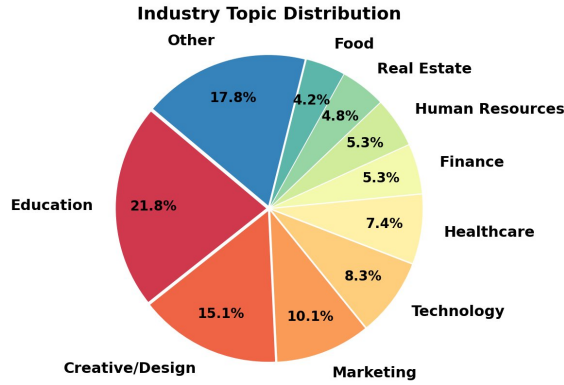


Figure 5: The topic distribution of 30k human-generated slides, highlighting prevalent categories such as Education and Technology.

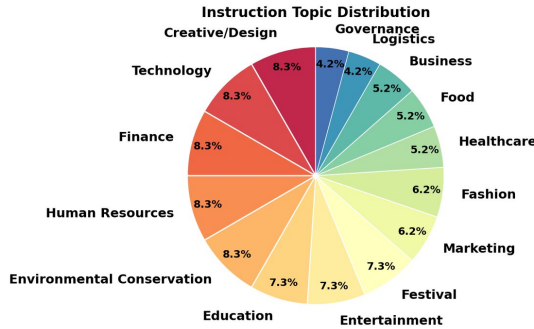


Figure 6: The distribution of instructions with different topics.

Domain	Count
Brand Promote	15
Business Plan	8
Knowledge Teaching	15
Personal Statement	20
Product Launch	24
Work Report	13
Topic Introduction	94
Total	189

Table 6: Domain Distribution

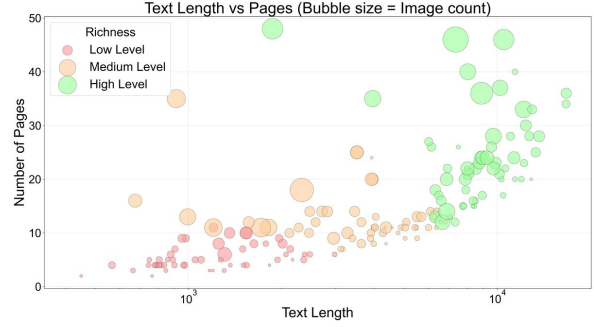


Figure 7: Analysis of Content Richness.
B Dataset Richness Category Details

The richness score S is defined as follows, where T and I denote the text length and image count:

$$S = w_t \cdot \frac{T - \min(T)}{\max(T) - \min(T)} + w_i \cdot \frac{I - \min(I)}{\max(I) - \min(I)} \quad (8)$$

where we set $w_t = 0.7$ and $w_i = 0.3$ to prioritize textual semantic density while accounting for visual requirements.

Based on the calculated scores, we stratify the dataset into three balanced levels—Low, Medium, and High—each containing exactly 63 instructions (33.3% of the total). As shown in the bubble plots in Figure 7, these levels correlate strongly with structural complexity: “Low” richness instructions typically require synthesizing fewer than 2,500 characters into brief decks, whereas “High” richness instructions demand the organization of over 7,500 characters and 20+ images into extensive presentations, effectively testing the model’s stability and long-context summarization ability.

C Justification for the Benchmark Size ($N = 189$)

The dataset size of $N = 189$ was a deliberate choice, necessitated by the high operational cost and technical complexity of evaluating professional slide generation. We offer the following clarifications regarding this design:

1. Practicality and Time Cost. High-quality presentation generation is significantly more time-intensive than standard text generation. For instance, tools like Zhipu PPT require approximately 20 minutes to produce a multi-slide deck in our settings. Evaluating nine mainstream systems across 189 instructions already demands nearly 1,300 generation hours. Thus, $N = 189$ represents a rigorous balance between evaluation depth and practical feasibility.

2. Representativeness and Scenario Richness.

We prioritized the quality, complexity, and diversity of each sample over sheer volume. Unlike benchmarks utilizing short prompts, SlidesGen-Bench employs comprehensive source documents (averaging 4,525 characters) to generate full-length decks (mean: 13.3 pages). The dataset covers seven distinct domains: 49.7% focuses on general knowledge synthesis (*Topic Introduction*), while the remaining 50.3% comprises specialized *Functional Tasks* (e.g., *Business Plans*, *Personal Statements*) to evaluate domain-specific adaptability and logical rigidity.

3. Depth via Difficulty Stratification. To ensure the benchmark remains sufficiently challenging, we implemented a Content Richness stratification (Low, Medium, and High). The “High” richness category specifically stress-tests a model’s formatting stability and long-context summarization abilities by demanding the organization of over 7,500 characters and 20+ images into cohesive, extensive presentations.

D QuizBank Evaluation Details

D.1 Multimodal Content Extraction Rather than directly feeding slide images into the evaluator to generate answers, we adopt a two-stage pipeline that begins with explicit content extraction. The primary motivation for this design is to decouple visual content extraction from semantic comprehension. By isolating these two processes, we can accurately diagnose the root causes of evaluation failures, explicitly distinguishing whether an incorrect answer stems from a visual extraction error or a reasoning/understanding shortfall. To transform slides into machine-readable context, we employ a state-of-the-art multimodal model (gemini-3-flash-preview) to extract the contents from slide images (the prompt used is in Appendix U). The extraction targets three specific modalities:

- **Textual Claims:** Verbatim headlines, core bullet points, and callout text.
- **Quantitative Data:** Specific numerical values extracted from charts, tables, and in-text metrics.
- **Visual Interpretation:** Semantic descriptions of non-textual elements (e.g., “green up-arrow indicating growth”) to capture visual information flow.

Benefiting from this decoupled design, our Error Analysis in Appendix Q shows that only 1.2% of errors stem from VLM extraction failures, ensuring the robustness of our evaluation against VLM bottlenecks.

D.2 The “Open-Book” Exam The evaluator VLM (gemini-3-flash-preview) is provided with the structured Markdown derived from the slides and the ground-truth QuizBank questions. Then the VLM is constrained by the prompt to “*Answer strictly based on the provided slide context and cite the source*” before attempting to answer QuizBank questions relying solely on this slide-derived context (the prompt used is in Appendix U).

D.3 Scoring and Analysis The evaluation metric relies on the binary success of the VLM examinee:

- **Correct Answer:** Confirms the specific concept or data point was successfully transferred from the source document to the slides.
- **Incorrect/Insufficient Information:** Indicates information loss. This distinction allows us to measure quality granularly, differentiating between the preservation of high-level concepts and specific data points.

E Mathematical Formulation of Color Harmony

For a given slide image, we utilize the HSV color space and a set of seven geometric templates $\mathcal{T} = \{i, I, V, L, T, Y, X\}$ on the hue wheel (Cohen-Or et al., 2006). An optimization problem is solved to find the optimal template T and rotation angle α that minimizes the weighted angular distance of the slide’s pixels to the template sectors. The harmonic deviation D is defined as:

$$D(H, S) = \frac{\sum_p S_p \cdot \text{dist}(H_p, T_\alpha)}{\sum_p S_p} \quad (9)$$

where H_p and S_p represent the hue and saturation of pixel p , and $\text{dist}(\cdot)$ calculates the shortest angular distance (in degrees) to the valid sector of the rotated template T_α (O’Donovan et al., 2011). We map this raw deviation D to the normalized quality probability score $S_{slide}^{(i)} \in [0, 1]$ used in the main paper via a Gaussian decay function.

F Aesthetic Metrics Parameter Tuning

This appendix details the parameter decision process for four key aesthetic metrics: **SubbandEn-**

Table 7: Raw Score Distributions

Metric	μ	σ^2
<i>Color Harmony</i>		
Mean Distance	0.202	0.131
Deck Mean Score	0.869	0.039
Deck Consistency	0.147	0.018
<i>Colorfulness</i>		
Mean	51.09	767.72
Std	11.28	72.89
<i>Subband Entropy</i>		
Mean	3.878	0.533
Std	0.457	0.036

trophyMetric, **ColorfulnessMetric**, **ColorHarmonyMetric**, and **VisualHRVMetric**. Our calibration follows a three-stage methodology: (1) empirical distribution analysis on human-generated slides, (2) initial parameter selection based on distribution characteristics, and (3) optimization against human preference rankings.

F.1 Methodology Overview

F.1.1 Stage 1: Empirical Distribution Analysis

We collected and analyzed over 3,000 human-generated PowerPoint slides from diverse professional contexts. For each metric, we computed raw scores across the corpus to establish baseline distributions (Table 7).

F.1.2 Stage 2: Initial Parameter Selection

Based on the empirical distributions (Figures 8–10), we established initial parameter values capturing international corporate design standards (e.g., Duarte’s principles) to minimize localized stylistic bias.

F.1.3 Stage 3: Human Preference Optimization

We constructed a verification set of 1,000 slide pairs sampled from the initial 3,000 slides with human preference rankings. Using Bayesian optimization, we tuned parameters to maximize Spearman correlation (ρ) between metric scores and human rankings.

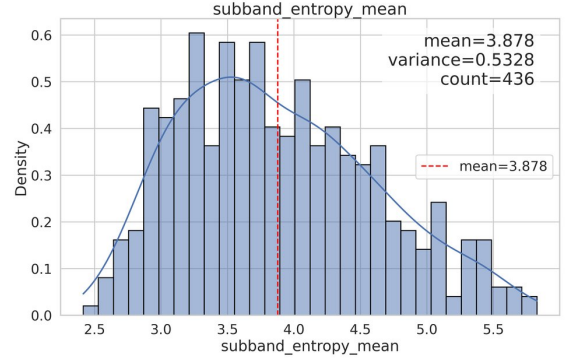
F.2 Subband Entropy Metric Parameters

The Subband Entropy metric measures visual clutter through steerable pyramid decomposition (Rosenholtz et al., 2007). The metric decomposes images in the CIE-LAB color space and computes Shannon entropy across subbands:

$$SE = w_L \bar{H}_L + w_{ab} \bar{H}_a + w_{ab} \bar{H}_b \quad (10)$$

Table 8: SubbandEntropyMetric Parameters

Param	Sym	Val	Rationale
Lum. Weight	w_L	0.84	From (Rosenholtz et al., 2007); luminance dominates clutter perception
Chrom. Weight	w_{ab}	0.08	Equal weight per channel; $\sum w = 1.0$
Pyramid Levels	L	3	Multi-scale without excess computation
Orientations	K	4	Standard config (0°, 45°, 90°, 135°)
Zero Thresh.	τ_0	0.008	Low-variation channels ignored

Figure 8: Distribution of Subband Entropy scores ($\mu = 3.878$).

where $\bar{H}_c = \frac{1}{N_c} \sum_{i=1}^{N_c} H(S_c^i)$ is the average entropy across subbands for channel c , with adaptive binning (bins = $\sqrt{|S|}$).

Figure 8 shows the distribution with a mean of 3.878 bits, aligning with structured presentation content.

F.3 Figure-Ground Contrast (Usability) Metric Parameters

The Usability score is computed from the mean figure-ground contrast of detected text regions, using a layout analysis model (Cui et al., 2025):

$$\text{Usability} = w_u \cdot \bar{C} \quad (11)$$

where $\bar{C}$ is the deck-level mean figure-ground contrast ratio and $w_u = 7$ is the calibrated linear weight.

F.4 Colorfulness Metric Parameters

The Colorfulness metric quantifies chromatic variety using the Hasler & Süssstrunk method (Hasler and Suesstrunk, 2003):

$$C = \sqrt{\sigma_{rg}^2 + \sigma_{yb}^2} + 0.3 \sqrt{\mu_{rg}^2 + \mu_{yb}^2} \quad (12)$$

where $rg = R - G$ and $yb = 0.5(R + G) - B$.

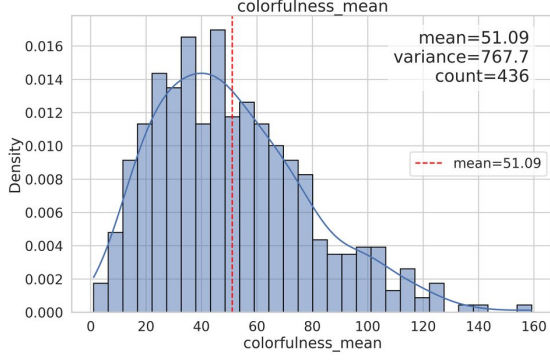


Figure 9: Colorfulness distribution ($\mu \approx 51$, $\sigma = 27.7$).

Table 9: Harmonic Templates (center, width as fractions of 360°)

Type	Sectors	Desc.
i	(0, 0.05)	Monochromatic
V	(0, 0.26)	Analogous
L	(0, 0.05), (0.25, 0.22)	Split-comp.
I	(0, 0.05), (0.50, 0.05)	Complementary
T	(0.25, 0.50)	Triadic
Y	(0, 0.26), (0.50, 0.05)	Split-comp. var.
X	(0, 0.26), (0.50, 0.26)	Double comp.

The coefficient $\alpha = 0.3$ was empirically derived through psychophysical experiments, balancing chromatic spread (σ) and bias (μ). Figure 9 shows professional slides exhibit moderate colorfulness ($\mu = 51.09$).

Deck-Level Engagement Scoring: The deck-level mean and standard deviation of slide colorfulness are mapped through two Gaussian kernels:

$$\text{Engagement} = w_{cf} w_{cf}^m e^{-(\mu - \mu_{cf})^2 / 2\sigma_{cf}^2} + w_{cf} w_{cf}^p e^{-(\sigma - \mu_{pace})^2 / 2\sigma_{pace}^2} \quad (13)$$

with $w_{cf} = 10$, $w_{cf}^m = 0.55$, $\mu_{cf} = 26$, $\sigma_{cf} = 40$, $w_{cf}^p = 0.5$, $\mu_{pace} = 15$, $\sigma_{pace} = 6$.

F.5 Color Harmony Metric Parameters

The Color Harmony metric evaluates adherence to harmonic color templates (Cohen-Or et al., 2006). For each pixel p with hue H_p and saturation S_p :

$$D_{\text{template}}(\alpha) = \frac{\sum_p S_p \cdot d(H_p, T_\alpha)}{\sum_p S_p} \quad (14)$$

where $d(H_p, T_\alpha)$ is the angular distance to template T rotated by α . The slide score uses Gaussian decay:

$$S_{\text{slide}} = \exp(-\bar{D}^2 / 2\sigma^2) \quad (15)$$

Deck-Level Scoring:

$$\text{Score} = w_h \cdot (w_m \mu_{\text{deck}} - w_d \sigma_{\text{deck}}) \quad (16)$$

Table 10: ColorHarmonyMetric Parameters

Param	Sym	Val	Rationale
Gaussian σ	σ	0.0005	Strictness; optimized via human pref.
Deck Weight	w_h	0.4	Deck-level scaling
Mean Weight	w_m	5	Reward high average harmony
Var. Penalty	w_d	20	Penalize cross-slide variance
Angular Res.	N	360	Full-degree granularity
Sat. Thresh.	$S_{\min}$	0.1	Exclude achromatic pixels

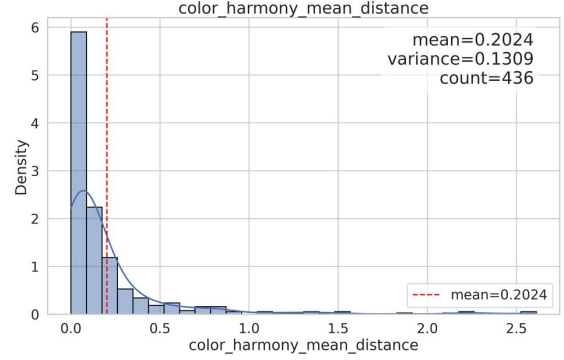


Figure 10: Color Harmony distance distribution ($\bar{D} = 0.202$).

rewarding high average harmony while penalizing inconsistency, with $w_h = 0.4$, $w_m = 5$, and $w_d = 20$.

F.6 VisualHRVMetric Parameters

The Visual HRV metric assesses presentation pacing by measuring variability across consecutive slides, inspired by physiological HRV analysis (van Ravenswaaij-Arts et al., 1993). For a raw entropy value x , calculate the normalized Score S (0 to 1):

$$S(x) = \frac{1}{1 + e^{-k \cdot (x - \mu)}}$$

where μ (Midpoint): The value of a “standard” slide, k (Slope): Sensitivity, x : The raw Subband Entropy of the current slide. In our configuration, $\mu = 2.1$ and $k = 1.5$.

Given normalized scores $\{S_1, \dots, S_n\} \in [0, 1]$:

$$\text{RMSSD} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n-1} |S_{i+1} - S_i|^2} \quad (17)$$

The final VisualHRV score combines the deck-level entropy term with the RMSSD pacing term:

$$\text{Score}_{\text{VHRV}} = w_{se} w_{se}^{\text{score}} e^{-(\bar{E}_{SE} - \mu_{se})^2 / 2\sigma_{se}^2} + 100 w_{se} (1 - \min(1, |\text{RMSSD} - \tau| / \tau_w)) \quad (18)$$

Table 11: VisualHRVMetric Parameters

Param	Sym	Val	Rationale
Target RMSSD	τ	0.03	Ideal variability center
Half-width	τ_w	0.21	Range [0, 0.5]
Sigmoid k	k	1.5	Per-slide normalization slope
Sigmoid μ	μ	2.1	Entropy of a “standard” slide
Entropy Mean	μ_{se}	5.4	Optimal complexity mean
Entropy Std	σ_{se}	1.6	Entropy Gaussian width
Entropy Weight	w_{se}	0.08	Entropy term weight
Entropy Scale	w_{se}^{score}	100	Entropy score scale

Table 12: VisualHRV Interpretation Bands

Category	RMSSD	Interpretation
Flatline	< 0.01	Monotonous; disengaging
Healthy	$[0.01, 0.1]$	Optimal visual rhythm
Transitional	other	Borderline pacing
Strobe Light	> 0.30	Jarring transitions

with $w_{se} = 0.08$, $w_{se}^{score} = 100$, $\mu_{se} = 5.4$, $\sigma_{se} = 1.6$, $\tau = 0.03$, and $\tau_w = 0.21$.

F.7 Parameter Optimization Results

Using 1,000 human-ranked slide pairs, we performed grid search to maximize Spearman ρ (Table 13).

SubbandEntropy and Colorfulness parameters retained literature values ($p > 0.05$). ColorHarmony σ showed the largest sensitivity, with a stricter threshold better distinguishing professional designs.

F.8 Summary

Our tuning combines principled estimates from empirical distributions with human judgment refinement:

- **Usability:** Linear weight $w_u = 7$ on the mean figure-ground contrast
- **SubbandEntropy:** Literature weights ($w_L = 0.84$, $w_{ab} = 0.08$) retained; entropy score Gaussian ($\mu_{se} = 5.4$, $\sigma_{se} = 1.6$, $w_{se} = 0.08$)
- **Colorfulness:** Hasler-Süsstrunk $\alpha = 0.3$ generalizes well; engagement score $w_{cf} = 10$, $\mu_{cf} = 26$, $\sigma_{cf} = 40$, $\mu_{pace} = 15$, $\sigma_{pace} = 6$
- **ColorHarmony:** Stricter $\sigma = 0.0005$ captures professional standards; deck score $w_h = 0.4$, $w_m = 5$, $w_d = 20$
- **VisualHRV:** $\tau = 0.03$, $\tau_w = 0.21$, sigmoid $k = 1.5$, $\mu = 2.1$ balances variety vs. cognitive load

Table 13: Selected Parameters after Human-Preference Tuning

Metric	Param	Value	Status
ColorHarmony	σ	0.0005	Human-pref. tuned
VisualHRV	τ	0.03	Human-pref. tuned
SubbandEntropy	w_L	0.84	Retained
Colorfulness	α	0.30	Retained

G Layout Detection Examples

We employ the PP-DocLayout_plus-L model (Cui et al., 2025) to detect and localize textual elements within presentation slides. This deep learning-based layout analysis model provides precise bounding box coordinates for various element types, including document titles, body text, images, and footers. Each detected element is assigned a confidence score indicating the model’s certainty in the detection.

G.1 Detection Output Format

The model outputs a structured JSON format containing the following information for each detected element:

- **label:** The semantic category of the detected region (e.g., text, image, doc_title, footer)
- **score:** Confidence score ranging from 0 to 1, representing the model’s certainty
- **coordinate:** Bounding box coordinates in the format $[x_{\min}, y_{\min}, x_{\max}, y_{\max}]$

G.2 Example for OCR Detections

Figure 11 illustrates the layout detection results on representative slides from the *Modern Architecture* presentation. The detected bounding boxes are visualized with different colors corresponding to element types: titles (green), body text (blue), images (red), and footers (gray).

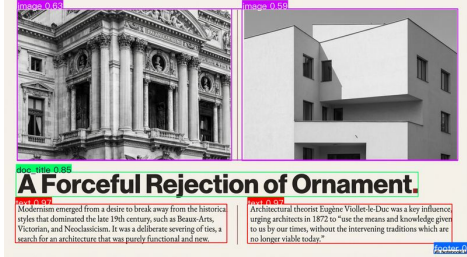
G.3 Detection Statistics

Table 14 summarizes the detection results across the example presentation slides. The model demonstrates high accuracy in identifying textual elements, with average confidence scores of 0.95 for text regions and 0.78 for document titles.

By leveraging these precise bounding boxes, we can accurately extract text content from presentation slides while avoiding interference from complex background images, decorative elements, and other non-textual components. This approach ensures that the text extraction process focuses exclusively on genuine textual content, improving the quality of downstream processing tasks.



(a) Title slide with a full-width background image, document title, subtitle text, and footer.



(b) Content slide with side-by-side images, title, dual text blocks, and footer.



(c) Two-column layout with title, multiple text regions, and a large image on the right.

Figure 11: Layout detection results using PP-DocLayout_plus-L on presentation slides. Bounding boxes are color-coded by element type: images, document titles, text, and footers. Confidence scores are displayed alongside each detection.

H Visual Rhythm Score Formulation Details

This appendix provides the detailed mathematical steps for calculating the Visual Rhythm Score described in the main text.

H.1 Subband Entropy (E_{SE})

The calculation of Subband Entropy for a single slide follows the methodology of Rosenholtz et al. (Rosenholtz et al., 2007).

1. Color Space Normalization. To decouple luminance from chromaticity, we operate in the CIELAB color space. The channels are first nor-

Table 14: Summary of layout detection results on example slides.

Element Type	Count	Avg. Score	Min. Score
Document Title	15	0.78	0.69
Text	28	0.95	0.87
Image	22	0.82	0.59
Footer	15	0.79	0.78

malized to a unit range:

$$L' = L/100 \quad (19)$$

$$\{a', b'\} = (\{a, b\} + 128)/255 \quad (20)$$

2. Channel Decomposition and Entropy.

We decompose each normalized channel $C \in \{L', a', b'\}$ using a Steerable Pyramid with 3 scales and 4 orientations. This simulates the receptive fields of the primary visual cortex (V1). For the resulting set of subbands $\{S_1, \dots, S_n\}$, we calculate the mean Shannon entropy for each channel:

$$E_C = \frac{1}{n} \sum_{i=1}^n \left(- \sum_k p_k \log_2(p_k) \right)_{S_i} \quad (21)$$

where p_k is the probability of a given intensity level k within a subband S_i .

3. Composite Score. The composite Subband Entropy E_{SE} is a weighted sum emphasizing luminance, as the human visual system is most sensitive to structural contrast. Following the channel weights of Eq. 10 ($w_L = 0.84$, $w_{ab} = 0.08$), the deck-level mean entropy $\bar{E}_{SE}$ is mapped to a scalar score $S_{entropy}$ using a Gaussian function centered on an optimal complexity mean:

$$S_{entropy} = w_{se} w_{se}^{score} \exp \left(- \frac{(\bar{E}_{SE} - \mu_{se})^2}{2\sigma_{se}^2} \right) \quad (22)$$

with $w_{se} = 0.08$, $w_{se}^{score} = 100$, $\mu_{se} = 5.4$, and $\sigma_{se} = 1.6$.

H.2 Temporal Pacing (RMSSD)

The RMSSD calculation quantifies the variability in visual complexity across the slide sequence.

1. Calculating Successive Differences. First, we compute the flux Δ_i as the absolute change in the complexity score $S_{entropy}$ between adjacent slides i and $i + 1$:

$$\Delta_i = |S_{entropy}^{(i+1)} - S_{entropy}^{(i)}| \quad (23)$$

This yields a sequence of differences $[\Delta_1, \dots, \Delta_{N-1}]$.

2. Root Mean Square. The RMSSD is then calculated as the root mean square of this sequence of successive differences, as shown in the main text.

I Aesthetic Metrics: Interpretation and Insights

I.1 Color Harmony: Interpretation and Insights

This metric is designed to detect “color clutter” by measuring how well a slide’s hue distribution fits established harmonic templates (Cohen-Or et al., 2006). Intuitively, professional decks maintain a coherent hue structure across slides (low deviation) even when layouts and content types change, which preserves brand identity and reduces perceptual noise.

Good Pattern (Low Deviation / Stable Profile).

A well-designed deck typically stays within a small deviation band (often near 0° to 10° in practice (Luo and Tang, 2008)), consistent with a chosen template (e.g., split-complementary). Even as the deck transitions from title to data or diagrams, hues remain constrained to the same template sectors, creating a unified palette.

Bad Pattern (Spikes / Palette Violations). Sudden spikes in harmonic deviation often arise from inserted assets (stock photos, copied charts, screenshots) whose palettes conflict with the established template. These deviations are perceived as clutter and can increase the viewer’s cognitive load by breaking the deck’s visual “contract.”

I.2 Engagement: Interpretation and Insights

Harmony alone can yield visually correct but emotionally flat slides. The adapted colorfulness metric M captures perceived vibrancy (Hasler and Suesstrunk, 2003), while the pacing term targets temporal coherence in that vibrancy across slides, aligning with the idea that presentations should not feel static or erratic (Post et al., 2017).

Volatile Pacing (“Jumpscare” Effect). Large slide-to-slide swings in M (high σ_{pacing}) create visual strobing: the audience repeatedly re-adapts to different intensity regimes, which feels jarring and amateurish.

Near-Zero Pacing (“Ghostly” Effect). If M is consistently low and $\sigma_{\text{pacing}} \approx 0$, the deck may be formally tidy yet perceptually monotonous. This pattern often corresponds to over-reliance on

grayscale text-on-white with minimal accent colors, which fails to sustain attention.

Ideal State (Controlled Band). High-quality decks maintain a stable “plateau” of vibrancy with moderate, purposeful variation—enough to mark emphasis and section boundaries without causing volatility.

I.3 Usability (Figure-Ground Contrast): Interpretation and Insights

Aesthetic choices must preserve legibility (Köhler, 1967). By computing contrast within detected text regions, the metric targets the local perceptual condition that determines readability, rather than relying on global image statistics.

The “Fog” Pattern (Low Local Contrast). Low contrast ratios in text regions (e.g., light text on light background) make information inaccessible, especially for visually impaired users or when projected in bright rooms. Such failures may remain hidden if only global contrast is considered.

Why Layout-Aware Contrast Matters (Local vs. Global). A slide can have high global contrast (e.g., dark background with a bright shape) yet still place text in a locally low-contrast region (e.g., dark gray on black). Region-level evaluation correctly flags the usability failure that histogram-based or global metrics can miss.

I.4 Visual Rhythm (VisualHRV): Interpretation and Insights

Presentations are temporal media; beyond per-slide complexity, the sequence structure matters for sustained comprehension (Duarte, 2010). VisualHRV connects computational cues (entropy-derived clutter and its temporal change) with pacing principles in cognitive psychology.

Static vs. Dynamic Cost. Subband Entropy proxies “feature congestion” and the cost of visual search (Rosenholtz et al., 2007). However, engagement and comprehension depend on how this cost evolves over time, not only on a single slide’s burden.

Effective Rhythm (Non-Flat, Intentional Variation). Motivated by narrative “sparkline” framing (Duarte, 2010) and cognitive load theory (Sweller, 1988), effective decks often alternate between dense and sparse slides, enabling consolidation and preventing working-memory overload.

This manifests as meaningful temporal variability (higher RMSSD) rather than a uniform sequence.

Failure Mode (High Complexity + Low Variation). A persistently demanding sequence with low fluctuation (low RMSSD at high mean complexity) can induce sustained cognitive strain and audience disengagement over time (a “flatline” difficulty profile), consistent with attention fatigue under continuous high perceptual load.

J Style-Transfer Validation of Aesthetic Metrics

To examine whether the aesthetic metrics transfer across visual styles, we ran a style-transfer validation over 1,435 Slides-Align1.5k aesthetics records spanning six style buckets. Table 15 summarizes the key diagnostics.

These results are reassuring but also help us state the limitations precisely. We present the aesthetic metrics as human-aligned proxies within the tested style distribution, disclose the style composition above, and note that domain- or style-specific recalibration may still be needed for styles outside this distribution.

K Detailed Specifications of the PEI Taxonomy

This appendix provides the technical criteria and definitions for the Presentation Editability Intelligence (PEI) levels introduced in Section 2.4. Table 16 shows the Technical Hallmark and Critical Failure Condition for each level.

K.1 Executive Summary

The **Presentation Editability Intelligence (PEI)** framework is a hierarchical standard for evaluating the structural integrity, semantic logic, and editability of AI-generated presentations. Unlike traditional metrics that measure visual similarity (e.g., FID scores), PEI measures **Editability Depth**. It asserts that a professional presentation is not merely a static image but a complex database of relationships defined by the Office Open XML standard.

K.1.1 The Knockout Rule (Dependency Logic)

The framework operates on a strict **dependency-based knockout mechanism**.

- Higher levels (e.g., L4 Data) rely on the existence of lower levels (e.g., L3 Structure).

- **Evaluation Protocol:** If a file fails a specific criterion, evaluation ceases immediately. The file is assigned the highest level it successfully completes.

- *Example:* A file with perfect animations (L5) but broken charts (L4 failure) is classified as Level 3.

K.2 Input Triage & Routing

The evaluation process begins with Input Format Analysis, which determines the evaluation pipeline and the Maximum Attainable Level (MAL).

K.2.1 Scenario A: The Static Input

- Supported Formats: .pdf, .png, .jpg, .jpeg
- Protocol: Immediate Termination.
- Maximum Attainable Level: L0
- Technical Rationale: These formats are flattened raster containers. They do not support object separation, text reflow, or XML data binding.

K.2.2 Scenario B: The Web Input

- Supported Formats: URL (Web Viewers, HTML5 Decks, Online Canvas links)
- Protocol: Visual & Interactive Inspection.
- Maximum Attainable Level: L2 (Vector)
- Technical Rationale: Web viewers render the Document Object Model (DOM) visually but obscure the underlying file structure. It is technically impossible to verify deep editability features—such as Master Slide inheritance (p:sldMaster) or embedded Excel binary binding (c:chart)—through a standard web interface. Therefore, the score is capped at the limit of visual verification (L2).

K.2.3 Scenario C: The Native Input

- Supported Formats: .pptx, .potx (Office Open XML formats)
- Protocol: Full Deep-Scan Evaluation.
- Maximum Attainable Level: L5 (Cinematic)
- Technical Rationale: These files provide full access to the XML schema, allowing verification of Master Slides, Data Relationships, and Animation Timings.

K.3 The PEI Hierarchy: Detailed Definitions

This section defines the technical criteria for each level.

Table 15: Style-transfer diagnostics of the aesthetic metrics over 1,435 Slides-Align1.5k records.

Diagnostic	Result	Interpretation
Style coverage	1,435 records across six buckets: standard corporate/mixed 959 (66.8%), minimalist-light 212 (14.8%), image-rich marketing 124 (8.6%), diagrammatic/scientific 91 (6.3%), data-heavy report 37 (2.6%), dark-mode minimalist 12 (0.8%)	The validation covers multiple visual styles, while standard corporate/mixed remains the dominant bucket and is disclosed.
Dark-mode check	Dark decks have slightly better human mean rank than light decks (3.84 vs. 4.27), higher figure-ground contrast (0.764 vs. 0.734), and similar corrected colorfulness (0.570 vs. 0.585)	We do not find a systematic dark-mode penalty.
Leave-one-style-out	Removing any single style bucket changes the composite alignment score by less than 0.05	The robustness result is not driven by one style bucket alone.
Boundary cases	Image-rich marketing decks have larger rank errors; harmony can penalize creative/non-standard palettes; visual rhythm is lower on dark images because of the entropy computation	The metrics should be interpreted as style-tested proxies, not universal aesthetics measures.

Table 16: The PEI Hierarchical Framework. This table utilizes vertical color coding to contrast capabilities with limitations. The **Technical Hallmark** column identifies the core value proposition, while the **Critical Failure Condition** column highlights the “knockout” factors that disqualify a system.

Level	Class Name	Operational Status	Technical Hallmark	Critical Failure Condition
L5	Cinematic	Dynamic Experience	Animation Logic, Media Embeds	Static slides only; Videos treated as static images; No temporal transitions.
L4	Parametric	Enterprise Tool	Native Data (<c:chart>), SmartArt	Charts drawn as vector shapes; Data uneditable (broken Excel link).
L3	Structural	Functional Tool	Global Masters (<p:sldMaster>)	Hardcoded layouts (no master inheritance); Logical groups ungrouped.
L2	Vector	Visual Toy	SVG Paths, Scalable Primitives	Icons or diagrams become blurry when zoomed in; Text fragmentation.
L1	Patchwork	Text-Editable Toy	Raster Backgrounds + OCR Text	Rasterized Text (OCR failed); Non-selectable background elements.
L0	Static	Flat Image	Text is rasterized pixels	

K.3.1 Phase 0: The Flat Phase

Level 0: Static (The Flat Image)

- Definition: The content is indistinguishable from a static bitmap.
- Technical Hallmark: Content is flattened. Text is rasterized pixels, not character strings.
- Critical Failure Condition (Knockout): N/A (Bottom of the hierarchy)

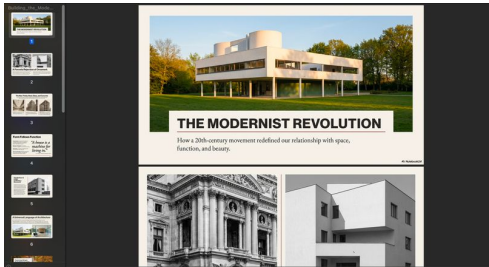


Figure 12: Level 0: Static (Uneditable PDF)

K.3.2 Phase 1: The Visual Phase (Surface Fidelity)

Level 1: Patchwork (The Text-Editable Toy)

- Definition: The file contains editable elements, but they are fragmented and structurally broken.
- Technical Hallmark: “OCR-style” reconstruction. Paragraphs are split into multiple single-line text boxes. Layouts rely on absolute positioning coordinates rather than flow.
- Critical Failure Condition (Knockout):
 - **Rasterized Text:** The “text” looks like letters but is actually an image (bitmap). Users cannot select, copy, or edit the characters.
 - **No Selectability:** Clicking on content selects the entire slide background instead of individual elements.

Level 2: Vector (The Visual Toy)

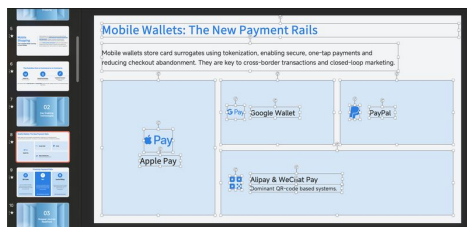


(a) Fragmented OCR text blocks

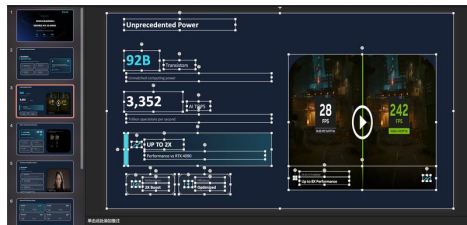


(b) Broken background elements

Figure 13: Level 1 Examples



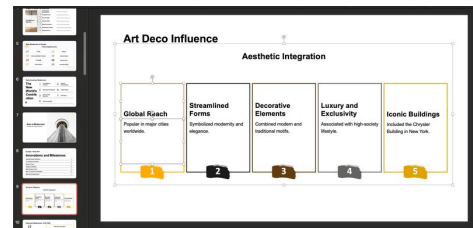
(a) Scalable primitives



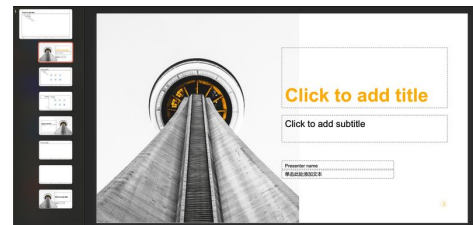
(b) SVG icons and paths

Figure 14: Level 2 Examples

- Definition: Visual clarity is achieved via vector graphics, but elements lack logical grouping.
- Technical Hallmark: Usage of SVG paths and Scalable Primitives. Graphics remain sharp at 400% zoom.
- Critical Failure Condition (Knockout):
 - **Pixelated Graphics:** Icons or diagrams become blurry when zoomed in (indicating they are Screenshots, not Vectors).
 - **Text Fragmentation:** Paragraphs are broken into separate text boxes for each line (failing the “Reflow” requirement of a true Vector container).



(a) The use of Grouping



(b) Master of the original slides

Figure 15: Level 3 Examples

K.3.3 Phase 2: The Structural Phase (Logic & Data)

Level 3: Structural (The Functional Tool)

- Definition: The system adheres to presentation software logic (Masters and Grouping).
- Technical Hallmark:
 - Logical Grouping: Related vector elements are bound using the Group function.
 - Master Inheritance: The file utilizes the <p:sldMaster> schema. Layout changes in the Master View propagate globally to all slides.
- Critical Failure Condition (Knockout):
 - **Atomic Isolation:** Complex graphics consist of hundreds of loose shapes requiring individual selection (No Grouping).
 - **Hardcoded Backgrounds:** Background elements are pasted onto every individual slide rather than inherited from the Master Slide.

Level 4: Parametric (The Enterprise Tool)

- Definition: Visuals are driven by native data parameters.
- Technical Hallmark:
 - Native Data Binding: Charts are instantiated as <c:chart> objects linked to an embedded Excel binary (.xlsx).
 - SmartArt/Diagrams: Process flows use semantic connectors, not just distinct lines.

- Critical Failure Condition (Knockout):

- **“Dead” Vector Charts:** A chart looks perfect but is constructed from static rectangles and text boxes. Right-clicking shows “Un-group” instead of “Edit Data”.
- **Broken Data Link:** The chart object exists, but the underlying Excel data is missing or corrupt.

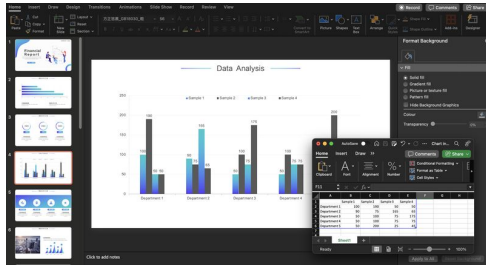


Figure 16: Level 4: Parametric (Native data in the chart) (Human-made)

K.3.4 Phase 3: The Experience Phase (Time & Narrative)

Level 5: Cinematic (The Dynamic Experience)

- Definition: The presentation functions as a directed, temporal narrative.
- Technical Hallmark:
 - Animation Logic: Elements utilize Build-In/Build-Out effects sequences that match the reading order.
 - Native Media: Video/Audio is embedded in the DOM with playback controls.
- Critical Failure Condition (Knockout):
 - **Static State:** The presentation is functionally perfect (Data & Structure are correct), but lacks time-dimension attributes (No animations, no transitions).
 - **External Dependency:** Media files are linked to a local path on the creator’s machine rather than embedded, causing playback failure.

K.4 Evaluation Protocols

Select the protocol below matching your input type.

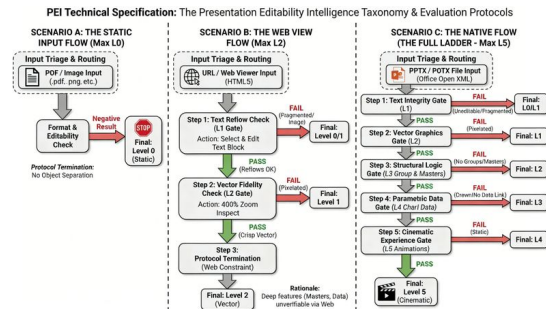


Figure 17: Evaluation Protocols Flow

K.4.1 Protocol A: The Static Flow

Input: PDF / Image

Procedure:

1. Format Check: Identify file extension (.pdf, .png, etc.).
2. Editability Check: Attempt to select text or move an object.
 - Result: Negative.
3. Final Classification: Level 0 (Static).

K.4.2 Protocol B: The Web Flow

Input: URL / Web Viewer

Constraint: Max Rating = L2.

Step 1: Text Reflow Validation (L1 Check)

- Action: Click on a text block in the web view.
- Check: Is it selectable text? If you delete words, does the text box resize or reflow naturally?
- Decision:
 - If Text is Image or Unselectable: Classify as L0.
 - If Text is fragmented/does not reflow: Classify as L1.
 - If Text behaves correctly: Proceed to Step 2.

Step 2: Vector Fidelity Validation (L2 Check)

- Action: Zoom the browser to 400%. Inspect icons and diagrams.
- Check: Are edges crisp (Vector/SVG) or pixelated (Raster)?
- Decision:
 - If Pixelated: Classify as L1.
 - If Crisp/Vector: Classify as L2.

Step 3: Protocol Termination

- *Reasoning:* Web views cannot reliably prove the existence of Master Slides or editable Excel data.
- *Final Classification:* Level 2 (Vector).

K.4.3 Protocol C: The PPTx Flow

Input: PPTX File

Procedure: Perform checks sequentially. Stop immediately upon failure.

Step 1: The Text Integrity Check (L1 Gate)

- *Action:* Select a paragraph. Edit the text to double its length.
- *Criteria:* The text must stay within its container and wrap automatically. The paragraph must be a single object, not multiple lines.
- *Result:*
 - Fail: Content is uneditable (L0) or fragmented (L1). STOP.
 - Pass: Proceed to Step 2.

Step 2: The Vector Graphics Check (L2 Gate)

- *Action:* Zoom to 400%. Inspect non-text elements (icons, shapes).
- *Criteria:* Elements must be vector shapes (Shapes/SVG), not raster screenshots.
- *Result:*
 - Fail (Pixelated): Downgrade to Level 1. STOP.
 - Pass: Proceed to Step 3.

Step 3: The Structural Logic Check (L3 Gate)

- *Action A (Grouping):* Click a complex icon. Does it move as one unit (Group) or scatter into pieces?
- *Action B (Masters):* View → Slide Master. Add a distinct shape to the layout. Close Master View. Does the shape appear on the slides?
- *Criteria:* Complex elements must be grouped; Layouts must inherit from Master.
- *Result:*
 - Fail: Classify as Level 2. STOP.
 - Pass: Proceed to Step 4.

Step 4: The Data Native Check (L4 Gate)

- *Action:* Identify a chart. Right-click the chart area. Look for “Edit Data.”
- *Criteria:* The “Edit Data” option must exist and successfully open an embedded Excel sheet. Changing a value in Excel must instantly update the chart visual.
- *Result:*
 - Fail (No option/Broken link): Classify as Level 3. STOP.
 - Pass: Proceed to Step 5.

Step 5: The Cinematic Check (L5 Gate)

- *Action:* Run “Slide Show” mode from the beginning.
- *Criteria:* Slides must transition automatically or smoothly. Elements should animate in (Build-ins). Embedded video must play natively.
- *Result:*
 - Fail (Static Show): Classify as Level 4. STOP.
 - Pass: Classify as Level 5 (Cinematic).

Note: The protocols described in Section K.4 serve as the conceptual and human-facing operational criteria to illustrate the rationale behind each PEI level. In the actual SlidesGen-Bench pipeline, these checks are executed fully automatically.

K.5 Automated Parsing Pipeline for PEI

For native .pptx outputs, our evaluation script programmatically unzips the archive and scans the underlying Office Open XML tree. For instance: Level 3 (Structural): The script checks for the existence of customized layout mapping in slideMaster.xml and validates <p:grpSp> (Group Shape) tags. Level 4 (Parametric): The script parses /chart-s/chart*.xml to verify if <c:externalData> nodes are valid and linked to an intact embedded Excel binary (.xlsx). Level 5 (Cinematic): The script inspects the <p:timing> node within the slide XML to confirm the presence of temporal animation tags (e.g., <p:bldLst>). For Web/Image outputs, the triage script automatically limits the max attainable level to L2 based on file extensions and DOM node extraction (e.g., counting <svg> vs <canvas> tags).

K.6 Disclaimer:

It is crucial to note that PEI is inherently *format-conditional*. Because it measures secondary structural editing capacity within the traditional Office Open XML paradigm, systems outputting flattened images or web-based HTML canvases face a natural architectural ceiling (e.g., typically capped at Level 2). This matrix does not penalize image-centric models for visual quality, but rather objectively delineates their current limitations in enterprise-level structural workflows.

L PEI Parser-Derived Subcriteria

To support finer-grained comparison of systems that cluster at the same PEI level, we additionally extract parser-derived subcriteria directly from the generated file structures. Table 17 lists the six subcriteria and how each is checked automatically; the main PEI levels remain unchanged.

M Evaluation Setup for Open-Source Frameworks

To ensure a fair and rigorous comparison between commercial platforms and open-source frameworks under our unified benchmark setting, we made necessary adaptations to the input pipelines of AutoPresent (Ge et al., 2025) and PPTAgent (Zheng et al., 2025). The detailed configurations are as follows:

Adaptations for AutoPresent. AutoPresent is designed to synthesize slides from scratch but does not natively support the direct ingestion of raw PDF documents or ultra-long texts. To circumvent this limitation, we introduced a text pre-processing step: we utilized an LLM to extract the core information from the source PDFs, summarize the content, and format it into structured bullet points that perfectly match AutoPresent’s required input format. Furthermore, while AutoPresent defaults to DALL-E for its image generation module, we replaced this backend with Nano-Banana-Pro. This modification was made to ensure optimal visual synthesis performance and to align the model’s visual capabilities with the state-of-the-art multi-modal standards evaluated in our benchmark.

Adaptations for PPTAgent. Evaluating PPTAgent presented a unique challenge, as its core methodology relies on extracting and modifying layouts from a user-provided reference PPT template, alongside the source document. Because our

benchmark evaluates “from-scratch” slide generation, supplying a bespoke reference template for each test case would introduce an unfair advantage. To establish an equitable comparison, we simulated the template-routing mechanisms utilized by commercial systems (e.g., Gamma.com.ai and Kimi standard mode). Specifically, we curated a localized template library comprising 50 high-quality, human-designed PPTs encompassing a wide range of usage scenarios and visual themes. Each template was meticulously annotated with semantic tags. During inference, we deployed an LLM to analyze the input document and dynamically retrieve the most contextually appropriate template from this library. The selected template and the document were then seamlessly fed into PPTAgent to render the final slides.

N System Snapshot and Versioning

Commercial slide-generation systems evolve quickly, so we present the reported results as a dated SlidesGen-Bench v1.0 snapshot: all commercial systems were accessed between November and December 2025, and the exact model snapshots are archived with the open-source release to support re-runs under the same protocol. Table 18 summarizes the evaluated systems, their generation paradigm, and the access window.

The released dataset, evaluation code, prompts, scoring scripts, and generated artifacts support independent verification and versioned benchmark snapshots as systems change. The fixed v1.0 results serve as a reproducible reference point, while the open-source release provides the mechanism for continued updates.

O Aesthetics Results by Presentation Purpose

We report the aesthetics results breakdown by purpose in Table 19. The results reveal distinct visual signatures for different presentation types. Product Launch slides demonstrate exceptional performance, achieving the highest Usability (5.53), Engagement (7.75), and Harmony (-1.01), reflecting a comprehensive design priority on capturing audience attention, ensuring clarity, and maintaining visual balance for marketing impact. Meanwhile, Work Report slides exhibit the highest Rhythm (11.31), suggesting that this domain often utilizes structured, repeating visual patterns to effectively convey professional updates and data.

Table 17: PEI parser-derived subcriteria.

Subcriterion	Parser check
Text object editability	Text is stored as selectable characters rather than rasterized pixels (e.g., <a: t> nodes in PPTX, text nodes in HTML DOM).
Object separability	Elements are individually addressable objects (shape IDs / DOM nodes) rather than a single flattened canvas.
Grouping	Related elements are bound with <p: grpSp> or equivalent group constructs.
Native chart/table editability	Charts/tables are instantiated as native objects (<c: chart> with embedded data) rather than drawn vector approximations.
Narrative structure	A document-wide hierarchy exists (e.g., <p: sldMaster> inheritance and layout mapping).
Animation/multimedia support	Temporal attributes exist (<p: timing>, build lists) and media is embedded natively in the DOM.

Table 18: Evaluated systems and access window.

System	Paradigm	Access window / notes
Gamma.com.ai	Source file (template)	Nov–Dec 2025
Quark PPT	Source file (template)	Nov–Dec 2025
Kimi-PPT (Standard / Smart)	Source file (template)	Nov–Dec 2025
Zhipu PPT	HTML code generation	Nov–Dec 2025
Skywork	HTML code generation	Nov–Dec 2025
NotebookLM	Image generation	Nov–Dec 2025
Skywork-Banana Mode	Image generation	Nov–Dec 2025
Kimi-Banana Mode	Image generation	Nov–Dec 2025
AutoPresent*	Source file (code)	Academic framework; adapted pipeline
PPTAgent*	Source file (template)	Academic framework; adapted pipeline

P Details of Human Annotation

To rigorously construct the Slides-Align1.5k dataset and evaluate human alignment, we implemented a strict annotation protocol. The details regarding annotator selection, training, and agreement are outlined below.

Annotator Qualifications and Diversity. Our annotator pool was carefully curated to represent a balanced mix of technical, aesthetic, and professional perspectives, ensuring that the ground truth reflects a holistic view of presentation quality. The group consisted of:

- **Ph.D. students in Computer Science**, who excel at evaluating the structural logic, content fidelity, and information hierarchy of the slides.
- **Undergraduate students from the College of Art**, who provided expert judgment on visual harmony, typography layout, and color palettes.
- **Office clerks**, representing the “end-user” perspective to assess the practical utility, readability, and professional suitability of the generated decks in real-world scenarios.

Standardized Training and Requirements. To minimize subjective variance across diverse backgrounds, all annotators were required to participate in unified training sessions before starting the ranking tasks. During these sessions, we provided

explicit definitions of the four targeted aesthetic dimensions: Harmony, Engagement, Usability, and Visual Rhythm. We systematically showcased representative “good” and “bad” patterns for each dimension and established clear operational guidelines for the ranking interface. A set of canonical examples was also provided as reference. This standardization ensured that all annotators shared a common, rigorous understanding of what constitutes a “professional” presentation.

Inter-annotator Agreement and Compensation. The reliability of human judgments is demonstrated by the inter-annotator agreement results (see the “Humans” entry in Table 2). The annotation consensus achieved an Average Spearman correlation of 0.85 with a remarkably low Standard Deviation of 0.12. This high correlation verifies that our protocol successfully elicited stable and consistent preferences, serving as a robust “gold standard” for calibrating our computational metrics. Furthermore, all participants were fairly compensated according to the local institutional standards for professional data labeling.

Annotation Platforms and Process. To obtain reliable human judgments on slide aesthetics, we developed a web-based annotation interface for ranking (Figure 18). Given the same input prompt and source documents, multiple AI systems generated slide presentations.

Purpose	Count	Usability	Engagement	Harmony	Rhythm
Brand Promote	94	5.17	7.39	-1.08	11.03
Business Plan	56	5.08	6.61	-1.24	10.86
Personal Statement	160	5.26	6.94	-1.35	10.13
Product Launch	150	5.53	7.75	-1.01	10.53
Topic Introduction	789	4.74	7.22	-1.28	10.13
Work Report	97	5.32	7.38	-1.14	11.31

Table 19: Aesthetics results breakdown across different presentation domains. The table reports the number of samples (Count) and average aesthetic scores for each category. Best values for each metric are highlighted in **bold**.

Human annotators were presented with high-resolution rendered slide images from the competing systems simultaneously. Rather than assigning absolute scores—which are highly susceptible to subjective calibration drift—we adopted a *relative ranking granularity*. For each prompt-matched set, annotators were required to rank the generated decks from best to worst.

Q QuizBank Error Analysis Details

Q.1 Error Taxonomy and Statistics

We conducted a human audit on all 2,499 instances. Human annotators directly inspected the rendered slide images against the ground truth. We categorize errors into four types. Table 20 presents the overall distribution, showing that missing information is the most prevalent issue.

Decoupling Generation Errors from Extraction Bottlenecks: A natural concern is whether “Missing Content” errors stem from the generation model’s omission or the extraction pipeline’s VLM failure to parse dense layouts. Table 20 shows that only 1.2% of the errors were falsely penalized due to VLM extraction failures on extremely dense or overlapping layouts. This confirms that our automated QuizBank evaluation faithfully reflects the generation models’ true capabilities with a highly acceptable error margin.

Q.2 Model-Specific Breakdown

Table 21 details error distributions by product. We observe that data-heavy topics (e.g., *internet_of_things*, *cryptocurrency*) suffer the highest error rates due to the demand for precise numerical reasoning.

Q.3 Qualitative Examples

Table 22 provides real-world examples of the three primary error categories.

R QuizBank Evaluation Robustness

Oracle-Context QA. To isolate VLM answering noise from generated-slide content loss, we ran

an oracle-context QA sanity check: the evaluator answers the same QuizBank MCQs from ground-truth source evidence instead of the slide-derived context. Across all questions, oracle-context QA achieves **100%** accuracy, showing that the questions are answerable from their source documents and that the VLM answer stage succeeds whenever the required evidence is present.

Backend Sensitivity. We ran an all-question backend-sensitivity comparison between DeepSeek-v4-Pro and MiniMax-M2.7 under otherwise identical settings (Table 23).

The nearly identical accuracies and symmetric disagreement cases suggest no clear systematic backend bias in the all-question comparison. The $\sim 60\%$ slide-context accuracy reflects generated-slide content fidelity; under oracle source evidence across all questions, the same QA stage reaches 100%.

Bootstrap Confidence Intervals. To quantify the stability of the main quantitative results, we computed 95% bootstrap confidence intervals by resampling the evaluated decks with replacement (1,000 iterations) and taking the 2.5th and 97.5th percentiles. Tables 24 and 25 report the per-system QuizBank accuracy and the composite total score produced by our scoring pipeline, together with the evaluated n (number of decks) per system. The intervals are tight, typically within $\pm 2\text{--}5\%$, and the top-ranked system’s lower confidence bound remains at or above the runner-up’s upper bound, indicating that the main benchmark conclusions are reasonably stable. For example, Kimi-Banana’s QuizBank lower confidence bound (85.4%) is at the same level as Skywork-Banana’s upper bound, and the composite total-score ranking is clearly separated at the top: Kimi-Banana’s interval [83.1, 84.0] lies above Skywork’s [80.6, 81.9].

S QuizBank Coverage and Bias Audit

To characterize the coverage and potential generation bias of the QuizBank, we conducted a full-



Figure 18: The web-based annotation interface for ranking.

Error Type	Count	Perc.	Definition
Type 1: Missing Content	1658	66.3%	Slide omits re-
Type 2: Value Mismatch	791	31.7%	quired facts/figures. Content exists but values differ from GT.
Type 3: VLM Extract Failure	31	1.2%	VLM fails to extract visible info.
Type 4: Other	19	0.8%	Unclassified/Formatting issues.

Table 20: **Distribution of Error Types.** Based on 2,499 incorrect answers.

Product	Total Errors	Missing Content	Value Mismatch	Other
Gamma	490	66%	32%	2%
NotebookLM	363	58%	41%	1%
Skywork	310	61%	36%	3%
Quark	278	70%	28%	2%
Skywork-Banana	262	67%	31%	2%
Kimi-Standard	222	69%	29%	2%
Kimi-Smart	197	63%	35%	2%
Kimi-Banana	191	44%	52%	4%
Zhipu	186	49%	49%	2%

Table 21: **Error Distribution by Product.** Sorted by total error count. Top-performing models (bottom rows) show a shift from “Missing Content” to “Value Mismatch.”

Type	Topic	Description & Root Cause
Missing Content	5G Deployment	Q: Which country achieved the first full commercial deployment? Issue: The slide discussed 5G concepts but omitted country-specific milestones.
Value Mismatch	Digital Marketing	Q: Cost per lead comparison? Issue: Slide stated “44% less expensive,” but Ground Truth expected a different statistic.
VLM Failure	Film Industry	Q: Nollywood’s global ranking? Issue: Slide text contained “fourth,” but VLM failed to extract it due to dense layout.

Table 22: Qualitative Examples of Common Errors.

Table 23: Backend-sensitivity comparison on the full QuizBank question set.

Backend	Accuracy	Questions answered
DeepSeek-v4-Pro	60.8%	All questions
MiniMax-M2.7	61.1%	All questions

bank audit over all available MCQs. Table 26 summarizes the main results.

The audit clarifies both the strength and the boundary of QuizBank. It is strongest for checking whether generated slides preserve concrete facts, quantities, entities, and relations from the source document. It is less direct for long-form narrative synthesis, rhetorical framing, and argument coherence, which we treat as complementary dimensions rather than claims made by the QuizBank score alone. The released QuizBank uses the shuffled option order.

T QuizBank Construction

We adopt a three-step LLM-based pipeline for quiz generation: (1) extracting quantitative and qualitative anchor points with verbatim citations (Figure 19), (2) verifying and refining extracted points against the source document (Figure 20), and (3) generating multiple-choice questions with source-grounded explanations (Figure 21). This multi-stage approach ensures high factual accuracy and traceability in the generated quiz bank.

U QuizBank Evaluation Prompt

Figure 22 shows the prompt we used to extract single slide contents for content evaluation. We extracted the exact text and the important information from the charts or images. Figure 23 shows the prompt to do the “open-book” QuizBank test.

V Prompt For LLM-as-Judge Aesthetics Evaluation

While we follow the standard nomenclature of “LLM-as-a-Judge”, our implementation is fully multimodal. To establish a rigorous baseline, we utilized the gemini-3-flash-preview API. Rather than relying on textual intermediate representations (e.g., Markdown or code), we fed the actual rendered slide images (in high resolution) directly to the model alongside the evaluation prompts. For the LLM-as-Judge Rating method, the model receives all slide images of a presentation and scores them based on the definitions of Harmony, Engagement, Usability, and Rhythm. For the LLM-as-Judge Arena, it takes two sets of slide images (Model A vs. Model B) and predicts human preference.

Figure 24 shows the head-to-head comparison prompt between slides. The criteria are the same as the proposed aesthetics metrics. Figure 25 shows the VLM rating prompt. The criteria are the same as the proposed aesthetics metrics.

W Generated Slides Examples

Figures 26–34 present qualitative comparisons of slides generated by Gamma, Kimi-Banana, Kimi-Smart, Kimi-Standard, NotebookLM, Quark, Skywork, Skywork-Banana, and Zhipu across three topics: 5G Technology (high complexity), Art Therapy (medium complexity), and Time Management (low complexity).

Step 1: Source of Truth Extraction Prompt:

System Role: Forensic Analyst.

Context:

- Domain: {domain} ({focus})
- Purpose: {one_sentence}
- Text: {document_text}

Task: Extract "Source of Truth" points.

CRITICAL: Do not use keywords. Write "Detailed Contextual Statements".

Instructions:

1. **Quantitative (Data):** Find 4-6 hard data points (Numbers, Dates, Specs).
 - *Bad:* "Revenue grew."
 - *Good:* "Q3 Revenue grew by 25% YoY to \$10M."
2. **Qualitative (Concepts):** Find 6-8 core concepts.
 - *Bad:* "AI Strategy."
 - *Good:* "Pivoting to 'AI-First' to reduce costs by 40%."
3. **Citation:** Quote the *original verbatim text* and Cite the Page Number using "[=== PAGE X START ===]".

Output JSON:

```
{
  "quantitative_anchors": [{"statement": "...",
    "source_quote": "...", "location": "Page X"}],
  "qualitative_key_points": [{"statement": "...",
    "source_quote": "...", "location": "Page X"}]
}
```

Figure 19: Step 1: Source of Truth Extraction Prompt for extracting quantitative and qualitative anchor points from source documents with verbatim citations.

Step 2: Verification and Refinement Prompt:

System Role: Strict Editor.

Context:

- Text: {document_text}
- Draft JSON: {draft_json}

Task: Audit, Refine, and Expand.

1. **Verify:** Check 'source_quote' exists in Text.
2. **Expand:** If a massive point is missing, ADD IT.
3. **Filter:** Remove trivial points.

Output: Return Polished JSON.

Figure 20: Step 2: Verification Prompt for auditing extracted anchor points against source documents, expanding coverage, and filtering trivial information.

Step 3: Quiz Generation Prompt:

System Role: Professional Exam Setter.

Input Context:

- Verified Data: {quantitative_anchors}
- Verified Concepts: {qualitative_key_points}

Task: Generate exactly {target_count} Multiple Choice Questions (MCQs).

STRICT FORMATTING RULES:

1. **Options:** Must be a list of 4 strings, explicitly starting with "A. ", "B. ", "C. ", "D. ".
2. **Correct Answer:** Must be a SINGLE LETTER ("A", "B", "C", or "D"). Do not write the full text.
3. **Explanation:** Must quote the source page.

EXAMPLE JSON OUTPUT (Follow this format exactly):

```
{
  "quiz_bank": [{
    "id": 1,
    "type": "Data",
    "question": "What is the active duration
                  for the 2025 campaign?",
    "options": [
      "A. January 1 through January 31, 2025",
      "B. January 6 through February 10, 2025",
      "C. February 1 through February 28, 2025",
      "D. March 1 through March 30, 2025"
    ],
    "correct_answer": "B",
    "explanation": "Based on Page 3: 'From January 6
                  through February 10, 2025...'"
  }]
}
```

Figure 21: Step 3: Quiz Generation Prompt for creating multiple-choice questions from verified anchor points with strict formatting constraints and one-shot example.

Single Slide Extraction Prompt:

Analyze the attached slide image for the specific purpose of a Content Quality & Accuracy Audit. I need to compare this slide against source documentation, so precision is paramount.

Please extract the content into the following structured concise Markdown file:

1. Textual Claims (Verbatim):

- **Headlines:** Extract the exact Title and Subtitle.
- **Core Statements:** List every distinct claim or bullet point found in the body text exactly as written. Do not summarize.
- **Callouts:** Extract text from any bubbles, arrows, or highlight boxes.

2. Quantitative Data Extraction:

- **Chart/Table Data:** For every chart or table, list the specific data points visible. (e.g., 'Q1 Revenue: \$10m', 'Year-over-Year growth: 15%').
- **In-Text metrics:** List any standalone numbers found in the text (e.g., '300+ employees', '50% reduction').

3. Visual Interpretation:

Describe important images or icons that contain information and explain if they convey a specific sentiment or data point (e.g., 'A green up-arrow indicating positive trend'). Ignore all decorative elements or irrelevant information.

Return the extracted content in a structured concise Markdown file.

Figure 22: Single Slide Extraction Prompt for content extraction from individual slides. This prompt is used to extract verbatim textual claims, quantitative data, and visual interpretations for downstream content evaluation.

Quiz Evaluation Prompt:

You are an expert quiz evaluator. Answer the following multiple-choice questions strictly based on the provided slide context and cite the source.

Presentation Topic: {topic}

Extracted Slide Contents: {slide_contents}

Quiz Questions: {quiz_questions}

INSTRUCTIONS

1. Read all extracted slide contents carefully.
2. For each question, select the best answer based ONLY on what's presented in the slides.
3. If the information is not covered in the slides, make your best inference or select "insufficient information".
4. Provide brief reasoning for each answer.

OUTPUT FORMAT (JSON)

```
{
  "answers": [
    {
      "question_id": <number>,
      "selected_answer": "<A|B|C|D>",
      "reasoning": "<brief explanation>"
    },
    ...
  ]
}
```

Figure 23: Quiz Evaluation Prompt for assessing content coverage through multiple-choice questions. The {topic}, {slide_contents}, and {quiz_questions} are replaced with the corresponding presentation data and generated quiz.

Arena Comparison Prompt:

You are an expert presentation judge. Compare two presentations (PPT A and PPT B) on the same topic based on their slide images.

Topic: {topic}

PPT A: {num_slides_a} slides **PPT B:** {num_slides_b} slides

CRITERIA

1. **VISUAL DESIGN:** Color scheme, typography, consistency, image quality, theme
2. **LAYOUT:** Spatial balance, alignment, no overlapping, professional structure

OUTPUT (JSON only)

```
{
  "Visual_Design": {
    "winner": "A"|"B"|"Tie",
    "score_difference": 1-5,
    "reason": "<brief reason>"
  },
  "Layout": {
    "winner": "A"|"B"|"Tie",
    "score_difference": 1-5,
    "reason": "<brief reason>"
  },
  "Overall_Winner": "A"|"B"|"Tie",
  "Overall_Reason": "<brief overall comparison>",
  "Confidence": 1-5
}
```

Figure 24: Arena Comparison Prompt for head-to-head evaluation of two presentations. The {topic}, {num_slides_a}, and {num_slides_b} are replaced with the corresponding presentation metadata.

Visual Evaluation Prompt:

You are an expert presentation designer evaluating the VISUAL DESIGN, LAYOUT, and COMPLEXITY of a Power-Point presentation. Evaluate based ONLY on the provided slide images.

Presentation Topic: {topic} **Number of Slides:** {num_slides}

VISUAL DESIGN CRITERIA (Weight: 40%)

- **Color_Scheme** (20%): Harmonic balance, contrast ratios, unified aesthetic.
- **Typography** (20%): Readable fonts, consistent sizes, clear hierarchy.
- **Visual_Consistency** (20%): Color coherence, recurring motifs, layout stability.
- **Image_Quality** (20%): High quality, relevant, properly integrated images.
- **Theme_Appropriateness** (20%): Visual theme matches content and audience.

LAYOUT CRITERIA (Weight: 40%)

- **Spatial_Balance** (40%): Effective whitespace, balanced elements.
- **Element_Alignment** (30%): Proper alignment of text, images, elements.
- **No_Overlapping** (30%): No obscured or cut-off elements.

COMPLEXITY CRITERIA (Weight: 20%)

- **Charts_and_Data** (25%): Charts, graphs, data visualizations where appropriate.
- **Visual_Elements** (25%): Icons, illustrations, diagrams, infographics.
- **Advanced_Design** (25%): Gradients, shadows, animations, depth effects.
- **Layout_Variety** (25%): Varied layouts appropriate for content types.

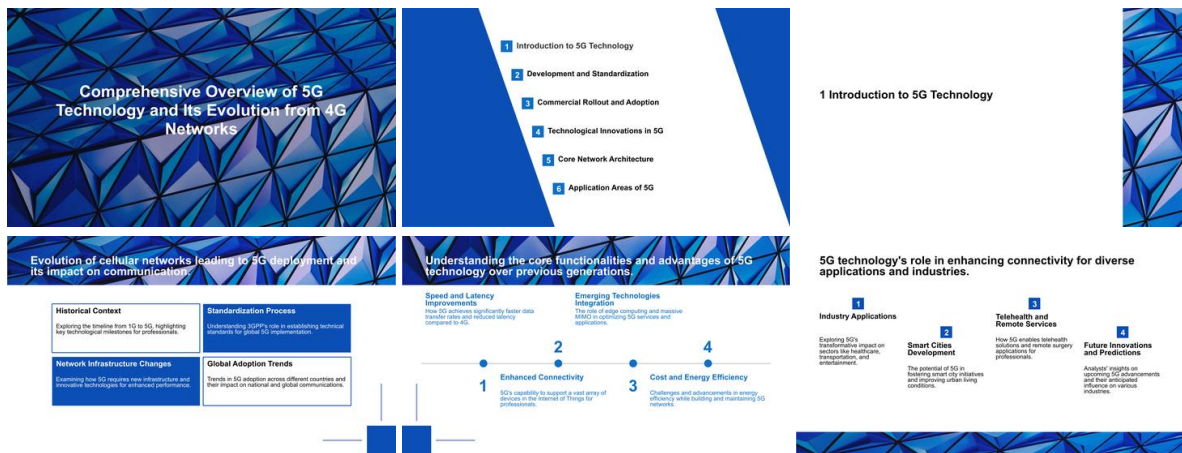
INSTRUCTIONS

1. Examine ALL slide images carefully in sequence.
2. Assign an EXACT INTEGER score from 0-10 for each sub-criterion (no decimals).
3. Provide constructive feedback on strengths and areas for improvement.

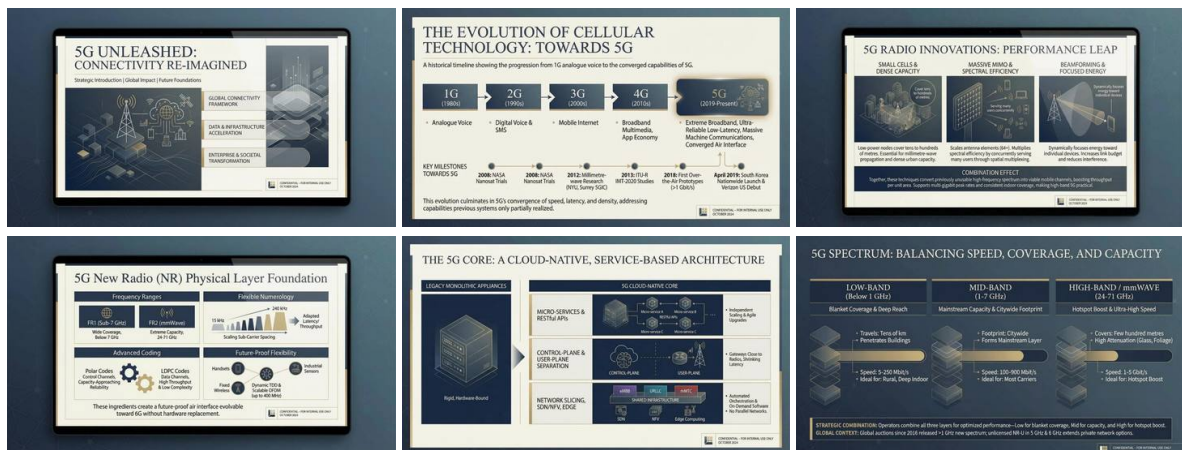
OUTPUT FORMAT (JSON)

```
{
  "Visual_Design": {
    "sub_scores": {
      "Color_Scheme": <0-10>, "Typography": <0-10>,
      "Visual_Consistency": <0-10>, "Image_Quality": <0-10>,
      "Theme_Appropriateness": <0-10>
    },
    "reason": "<detailed reasoning>"
  },
  "Layout": {
    "sub_scores": {
      "Spatial_Balance": <0-10>, "Element_Alignment": <0-10>,
      "No_Overlapping": <0-10>
    },
    "reason": "<detailed reasoning>"
  },
  "Complexity": {
    "sub_scores": {
      "Charts_and_Data": <0-10>, "Visual_Elements": <0-10>,
      "Advanced_Design": <0-10>, "Layout_Variety": <0-10>
    },
    "reason": "<detailed reasoning>"
  },
  "Overall_Feedback": "<brief summary>",
  "Top_Strengths": ["<strength 1>", "<strength 2>"],
  "Areas_for_Improvement": ["<area 1>", "<area 2>"]
}
```

Figure 25: Visual Evaluation Prompt for assessing presentation aesthetics. The {topic} and {num_slides} are replaced with the corresponding presentation metadata. Each sub-criterion is scored on a 0-10 integer scale.



(a) Gamma



(b) Kimi-Banana

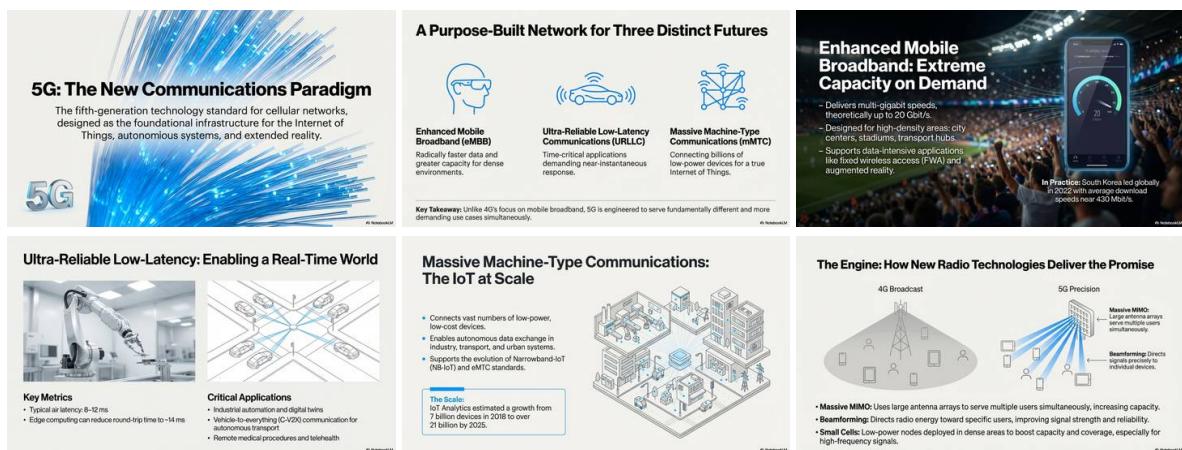


(c) Kimi-Smart

Figure 26: Comparison of slides generated on “5G Technology” (Part 1 of 3). Each subfigure shows 6 consecutive slides (2 rows $\times$ 3 columns) from a single product.



(d) Kimi-Standard

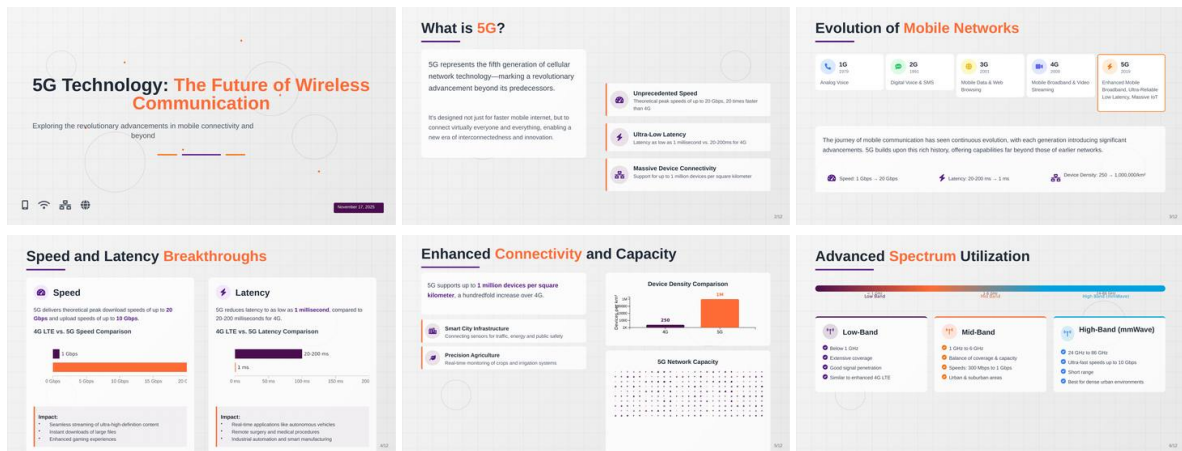


(e) NotebookLM



(f) Quark

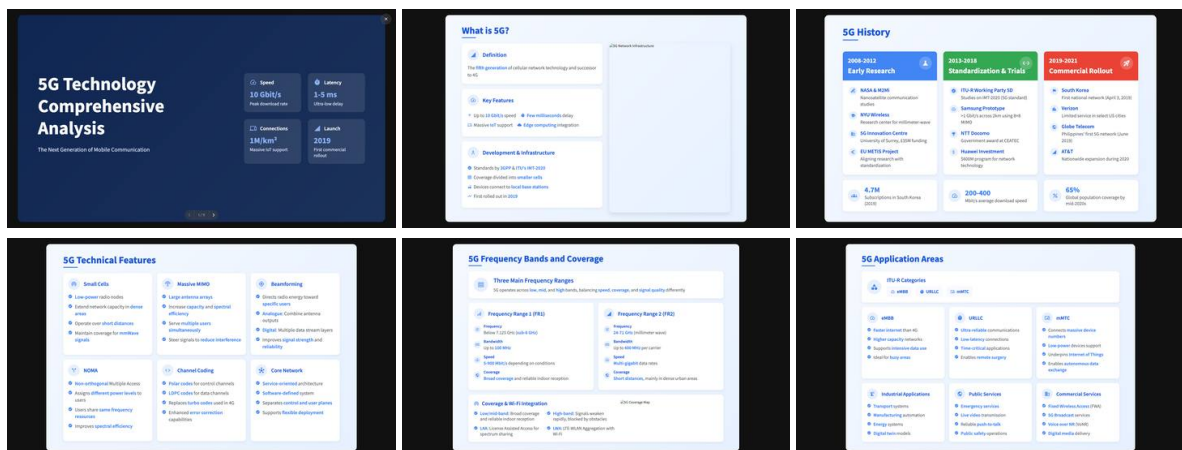
Figure 27: Comparison of slides generated on “5G Technology” (Part 2 of 3). Each subfigure shows 6 consecutive slides (2 rows × 3 columns) from a single product.



(g) Skywork



(h) Skywork-Banana

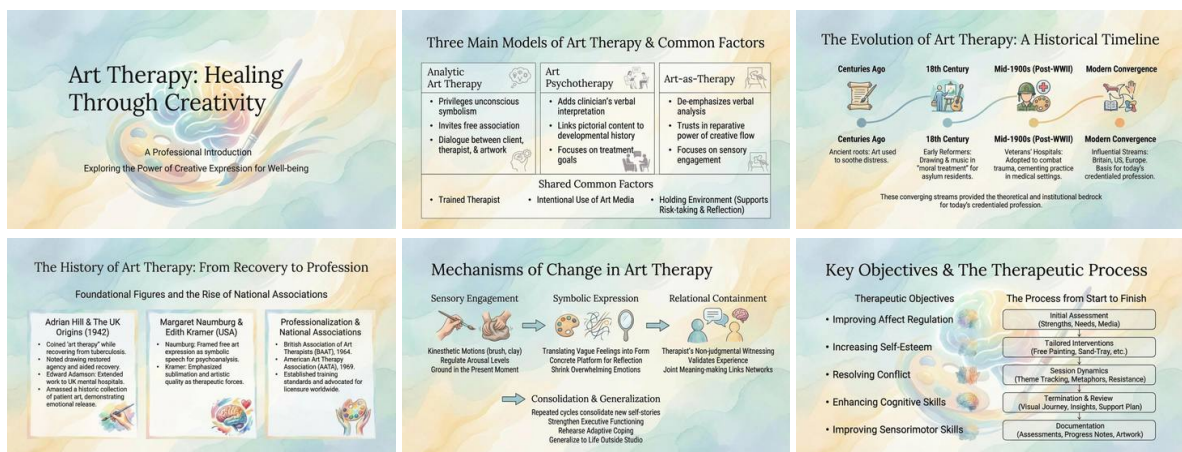


(i) Zhipu

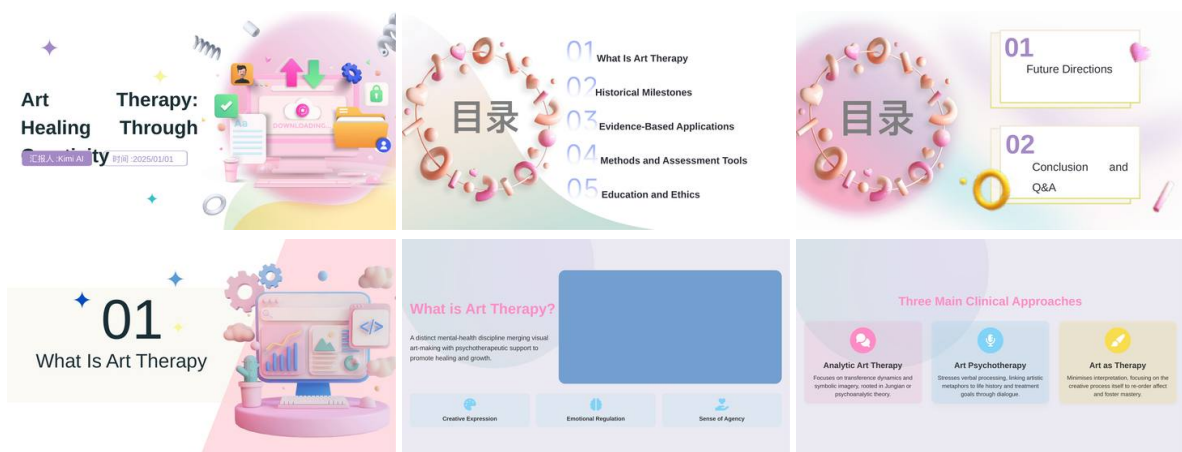
Figure 28: Comparison of slides generated on “5G Technology” (Part 3 of 3). Each subfigure shows 6 consecutive slides (2 rows × 3 columns) from a single product.



(a) Gamma

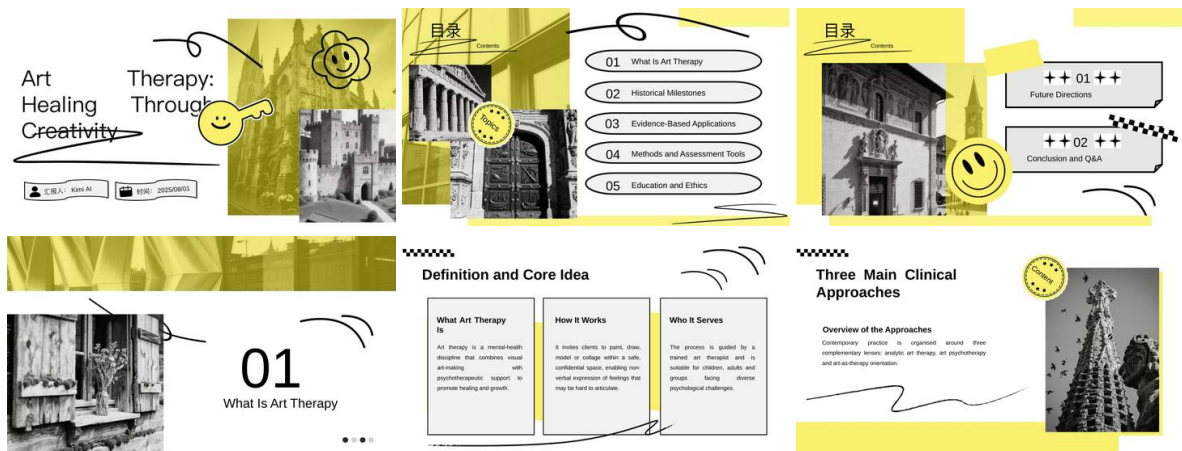


(b) Kimi-Banana

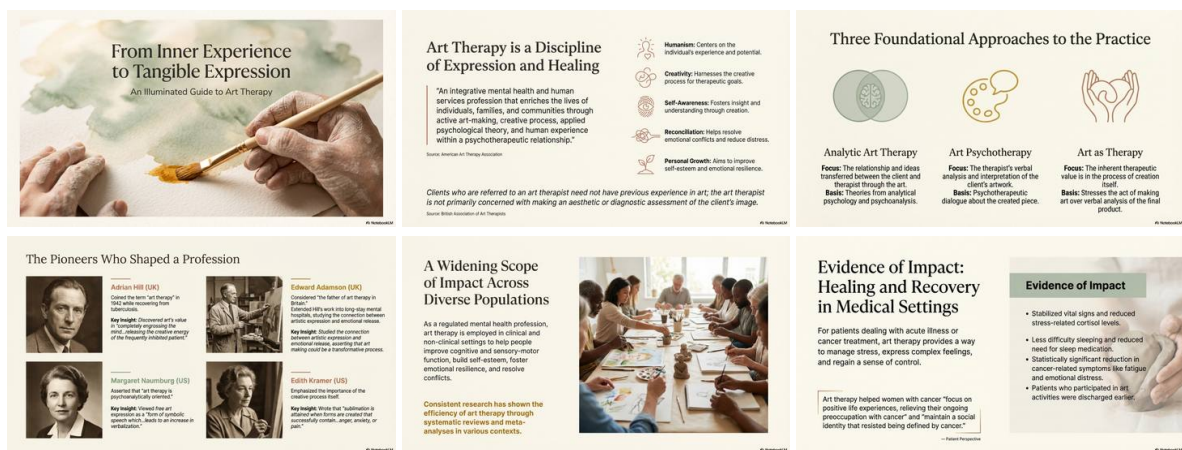


(c) Kimi-Smart

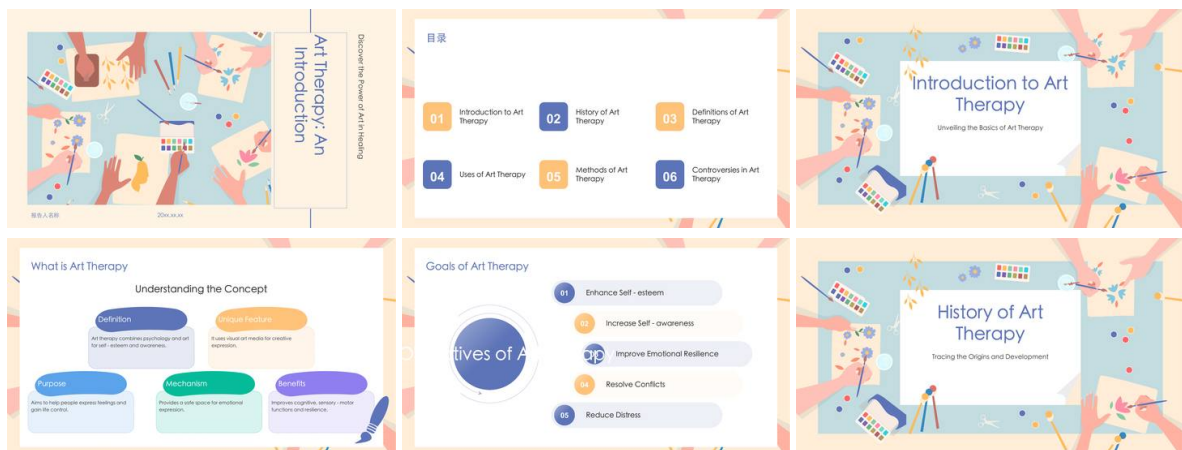
Figure 29: Comparison of slides generated on “Art Therapy” (Part 1 of 3). Each subfigure shows 6 consecutive slides (2 rows × 3 columns) from a single product.



(d) Kimi-Standard

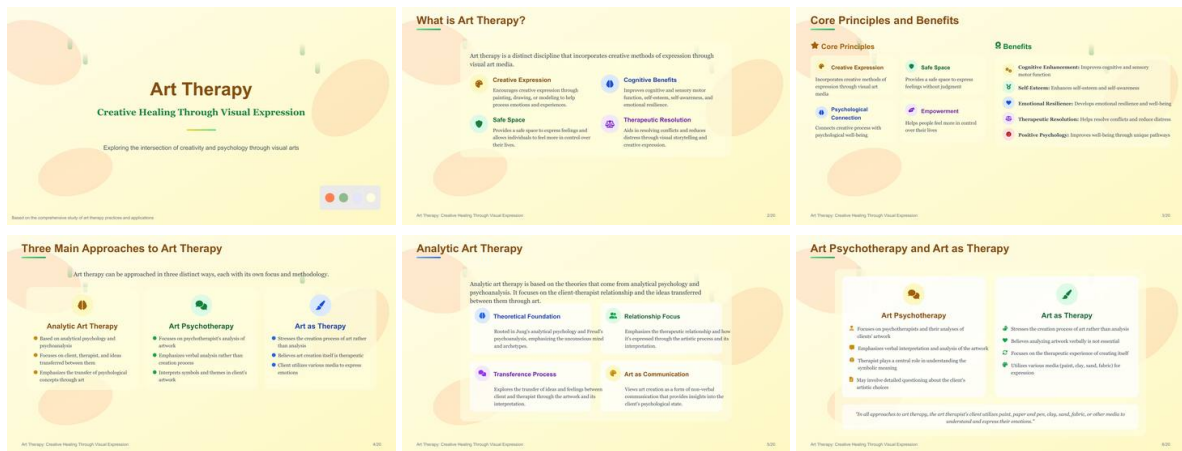


(e) NotebookLM

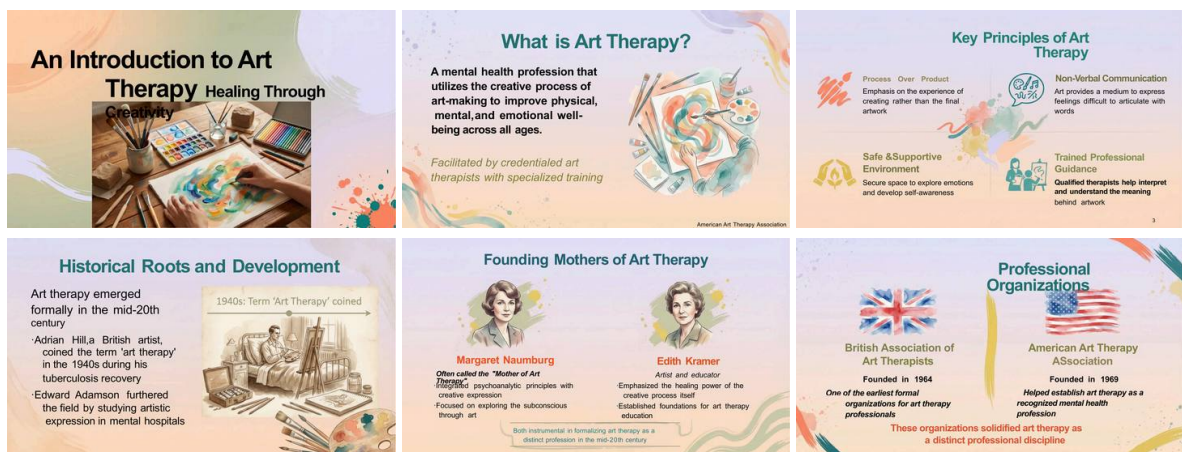


(f) Quark

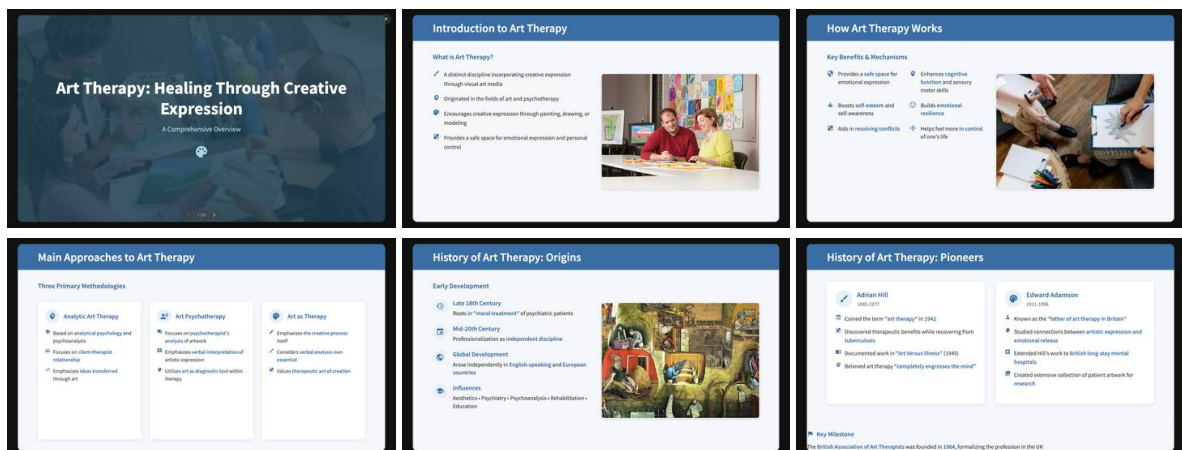
Figure 30: Comparison of slides generated on “Art Therapy” (Part 2 of 3). Each subfigure shows 6 consecutive slides (2 rows × 3 columns) from a single product.



(g) Skywork

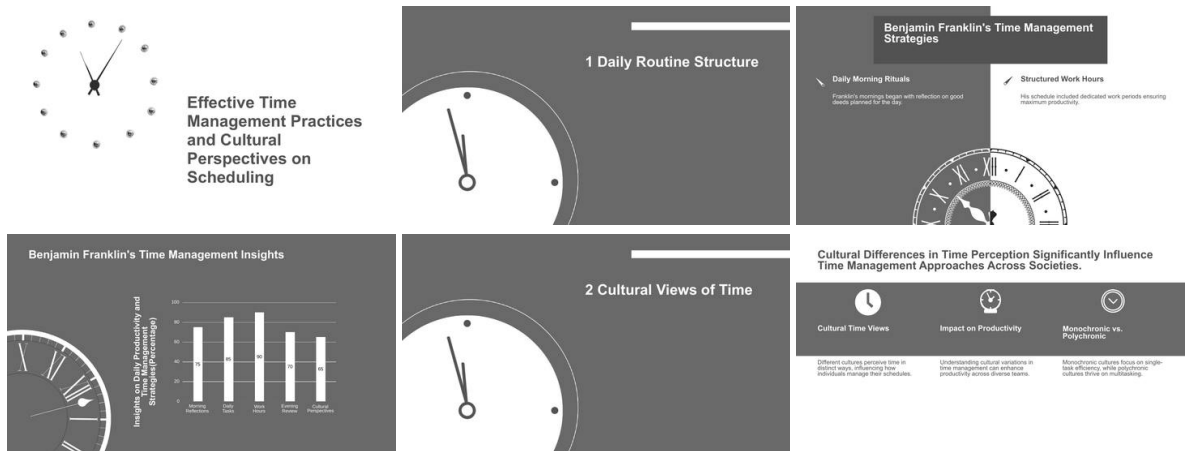


(h) Skywork-Banana

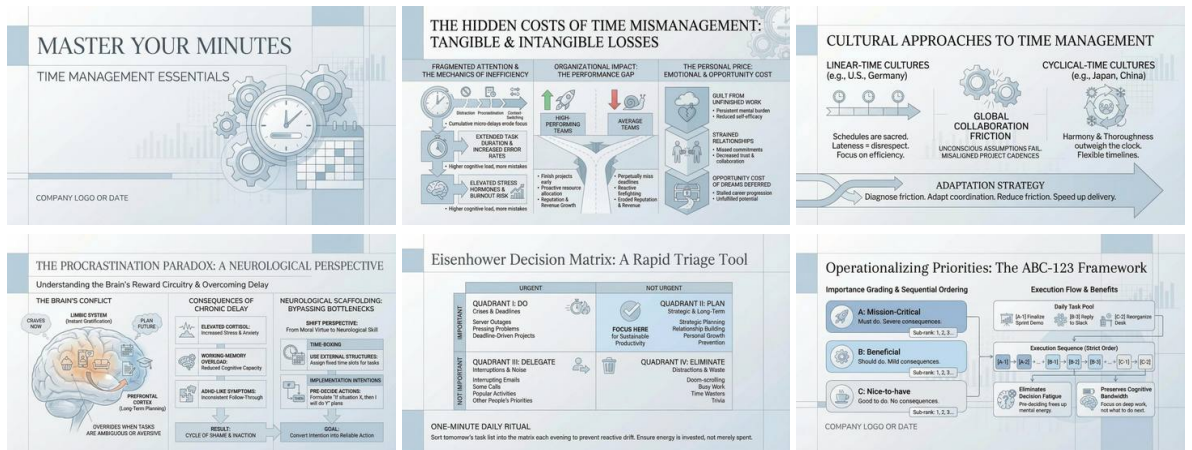


(i) Zhipu

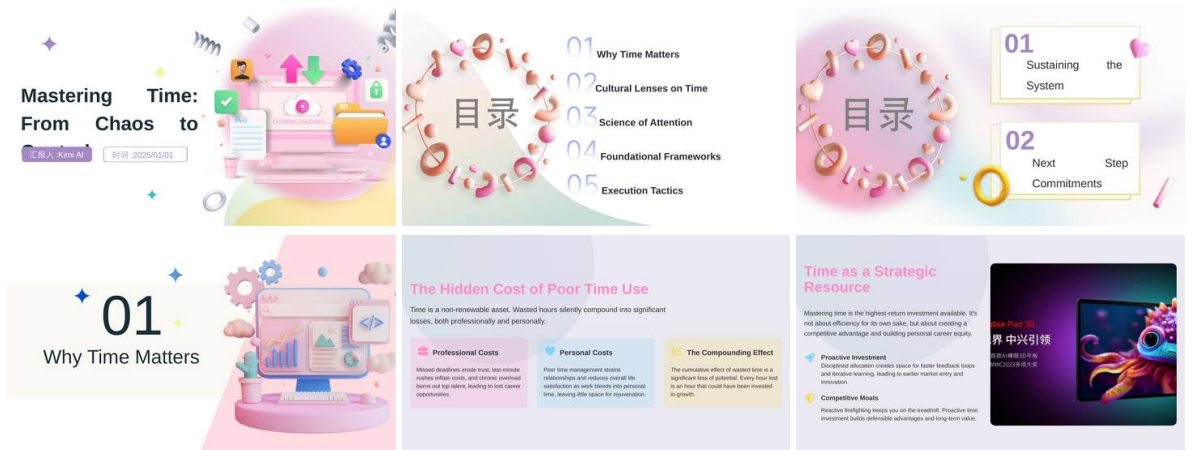
Figure 31: Comparison of slides generated on “Art Therapy” (Part 3 of 3). Each subfigure shows 6 consecutive slides (2 rows $\times$ 3 columns) from a single product.



(a) Gamma

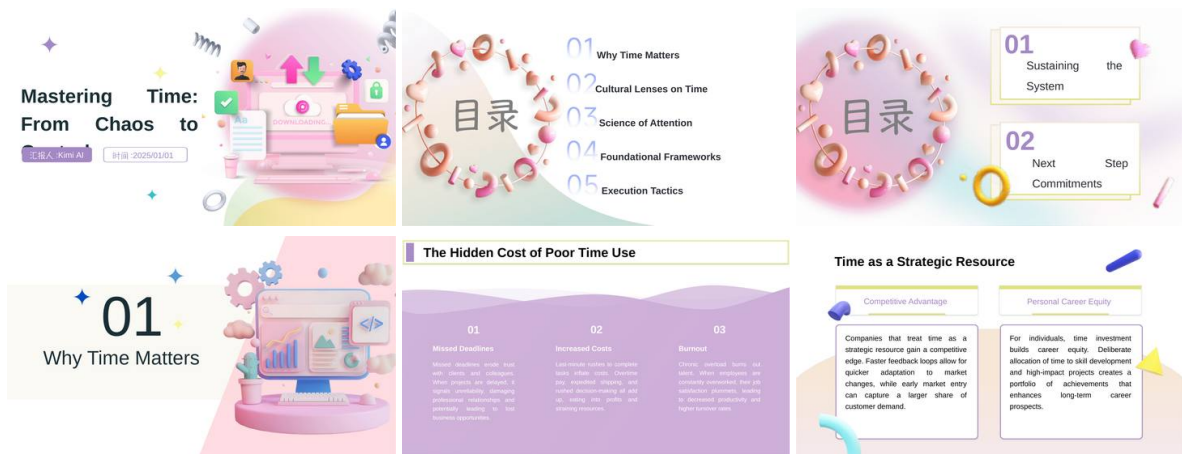


(b) Kimi-Banana

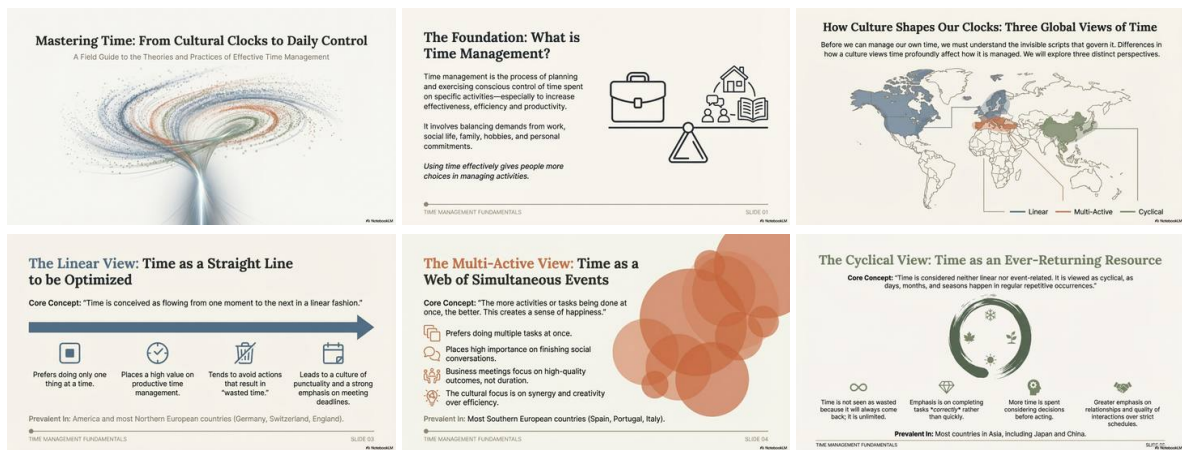


(c) Kimi-Smart

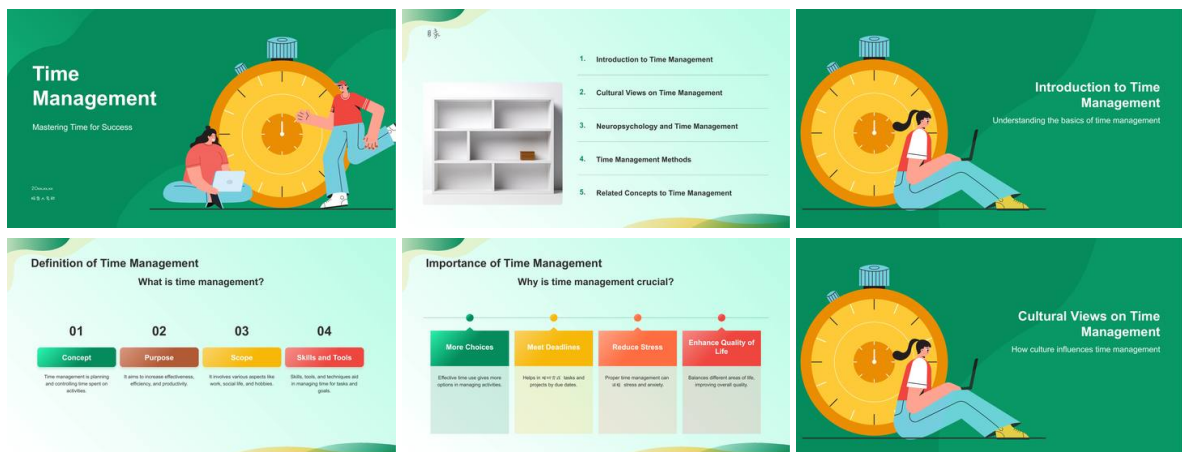
Figure 32: Comparison of slides generated on “Time Management” (Part 1 of 3). Each subfigure shows 6 consecutive slides (2 rows × 3 columns) from a single product.



(d) Kimi-Standard

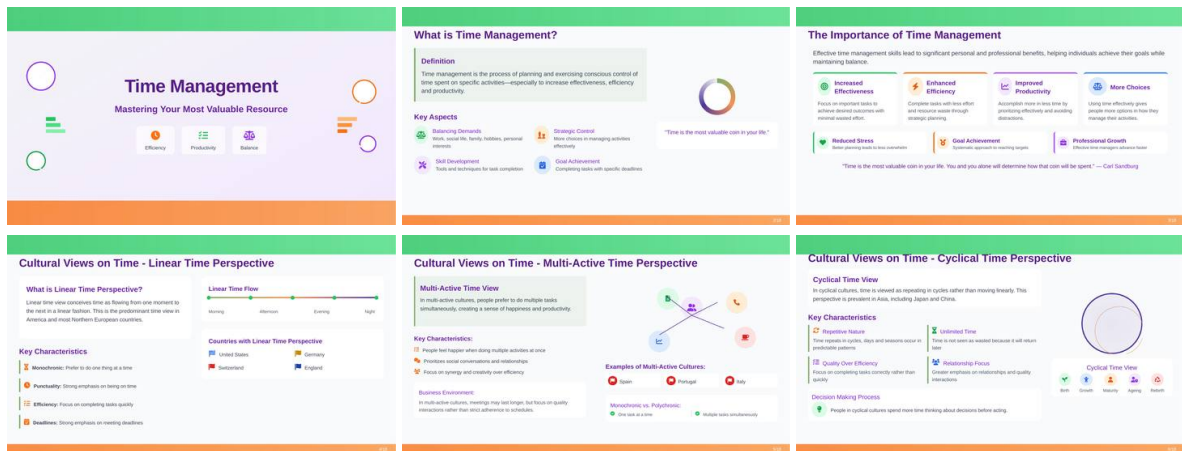


(e) NotebookLM



(f) Quark

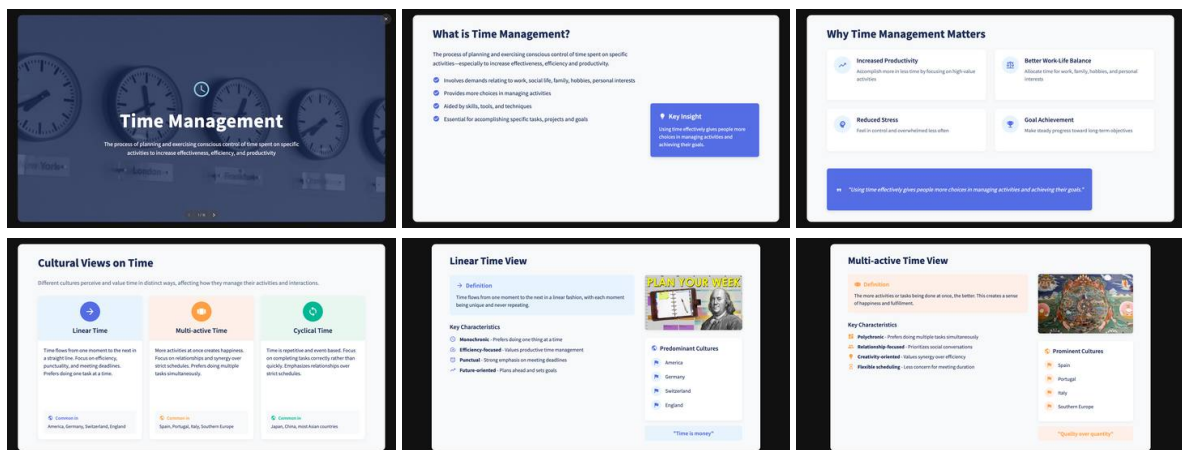
Figure 33: Comparison of slides generated on “Time Management” (Part 2 of 3). Each subfigure shows 6 consecutive slides (2 rows × 3 columns) from a single product.



(g) Skywork



(h) Skywork-Banana



(i) Zhipu

Figure 34: Comparison of slides generated on “Time Management” (Part 3 of 3). Each subfigure shows 6 consecutive slides (2 rows $\times$ 3 columns) from a single product.

Table 24: QuizBank accuracy by product with 95% bootstrap confidence intervals and evaluated n .

Product	Mean Accuracy	95% CI	n
Kimi-Banana	87.1%	[85.4%, 88.7%]	175
Zhipu	86.8%	[83.8%, 89.5%]	180
Skywork-Banana	82.9%	[80.2%, 85.4%]	182
Kimi-Smart	82.1%	[79.1%, 84.8%]	116
Skywork	81.3%	[78.8%, 83.7%]	186
Quark	80.9%	[77.5%, 83.9%]	158
Kimi-Standard	79.9%	[76.4%, 83.1%]	118
Gamma	69.9%	[66.3%, 73.5%]	183
NotebookLM	69.0%	[62.6%, 75.2%]	137

Table 25: Composite total score by product with 95% bootstrap confidence intervals.

Product	Score	95% CI
Kimi-Banana	83.5	[83.1, 84.0]
Skywork	81.2	[80.6, 81.9]
Skywork-Banana	79.2	[78.4, 80.0]
Zhipu	78.4	[77.0, 79.6]
Kimi-Smart	70.0	[69.3, 70.7]
Quark	68.9	[67.9, 69.7]
NotebookLM	67.6	[63.8, 71.4]
Kimi-Standard	67.6	[66.9, 68.2]
Gamma	65.1	[63.9, 66.2]

Table 26: Full-bank QuizBank coverage and bias audit.

Audit item	Result	Interpretation
Available MCQs	1,879 over 188 document records	Full-bank audit rather than a small manual-only subset
Declared type balance	Data 49.7%, Concept 50.3%	The intended 5-data/5-concept design is preserved
Data share by domain	47.0%–57.7%	No single domain is dominated by one declared type
Data share by difficulty	Low 47.7%, Medium 50.0%, High 51.3%	Type balance is stable across difficulty levels
Evidence anchors before sampling	14.24 per document on average	The final 10-question probe is sampled from a broader verified pool
Question-demand tendency	Numeric/date/statistic 70.3%; entity/framework 9.2%; relation/comparison 7.5%; procedure/action 5.1%; definition/concept 3.4%	QuizBank primarily measures source-grounded factual retention
Answer-position prior (after fixed-seed shuffling)	A 25.3%, B 25.0%, C 24.3%, D 25.4%	The answer-position prior is removed without changing question semantics
Wrong selected labels in stored evaluator outputs	A 36.7%, B 15.7%, C 14.5%, D 33.2%	Observed errors are not driven by exploiting the original B/C prior