Any2Poster: Any-Source Poster Generation Across Modalities and Domains

Amogh Vinaykumar [1] **Aiden Li** [2] **Suozhi Huang** [3] **Shilong Liu** [3]

Abstract

Visual posters are a compact medium for communicating dense information, yet progress on automatic poster generation remains difficult to measure because existing evaluations are often restricted to paper-only inputs, narrow domains, or surface-level visual similarity. We introduce Any2Poster Bench, a benchmark for any-source poster generation that evaluates systems across eight input modalities—PDFs, URLs, PPTX, DOCX, Markdown, LaTeX, notebooks, and videos—and five content domains. Any2Poster Bench pairs each source with quiz-based probes of verbatim factual retention and interpretive understanding, together with VLM-based judgments of visual quality, layout, readability, content completeness, and logical flow, enabling reproducible assessment of both information fidelity and visual communication. To instantiate and validate this benchmark, we further present Any2Poster Agent, an end-to-end reference agent that parses heterogeneous sources, organizes salient content, plans poster layouts, renders posters, and iteratively refines them using visual feedback. On Any2Poster Bench, Any2Poster Agent achieves 87.25% average accuracy across input modalities and 87.28% across content domains. On PaperQuiz-style evaluation, where prior paper-to-poster agents are directly comparable, Any2Poster Agent improves over PosterAgent-4o from 51.06–51.33% to 72.58% overall accuracy and from 116–121 to 145.16 in density-augmented score. Together, Any2Poster Bench and Any2Poster Agent provide a reusable evaluation resource and a competitive baseline for studying multimodal, domain-general poster generation.

[1]Flower Mound High School, Flower Mound, TX, USA [2]University College London, London, UK [3]Princeton University, Princeton, NJ, USA. Correspondence to: Amogh Vinaykumar <amogh.vinaykumar@gmail.com>.

Accepted to the ICML 2026 Workshop on Efficient Multimodal Question Answering (EMM-QA), Seoul, South Korea.

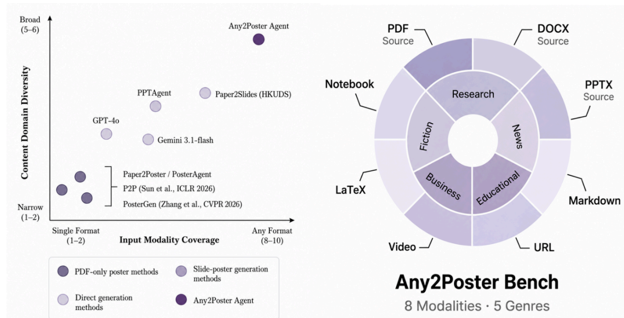


Figure 1. Positioning of Any2Poster Bench and representative generation systems along input modality coverage and content-domain diversity. Existing poster-generation methods mainly focus on single-format academic-paper inputs, while Any2Poster Bench targets broader source modalities and more diverse generation scenarios.

1. Introduction

Visual posters are widely used for scientific communication, education, business reporting, and public information sharing. A high-quality poster compresses rich source content into a single visually coherent page, requiring content selection, visual hierarchy, spatial planning, and cross-modal alignment. Prior work shows that even the paper-to-poster setting requires long-context understanding, multimodal asset extraction, layout planning, and visual-in-the-loop refinement (Xu & Wan, 2022; Pang et al., 2025; Sun et al., 2025; Zhang et al., 2026).

Existing poster-generation evaluations remain limited in scope. Most prior benchmarks and systems focus on a single input format, especially academic PDFs. In real-world poster creation, however, users may start from papers, webpages, slide decks, documents, notebooks, videos, Markdown files, or LaTeX projects. These sources differ in modality, structure, length, domain, and communication goal, so a system that works well on papers may fail on weakly structured prose, business reports, instructional materials, or web-based content. As shown in Figure 1, this leaves a gap between paper-only poster generation and general-purpose poster generation across modalities and domains.

To address this gap, we introduce **Any2Poster Bench**, a benchmark for *any-source poster generation*. Any2Poster Bench covers eight input modalities—PDFs, URLs, PPTX files, DOCX files, Markdown documents, LaTeX projects, notebooks, and videos—and five content domains: research, news, education, business, and fiction. It evaluates both cross-modality and cross-domain generalization, testing whether systems can transform diverse real-world sources into informative, readable, and visually coherent single-page posters. To measure information fidelity, Any2Poster Bench uses **BenchQuiz**, a quiz-based reading-comprehension protocol inspired by PaperQuiz (Pang et al., 2025), and complements it with VLM-as-judge assessment of visual and communicative quality.

We also provide **Any2Poster Agent** as a strong reference system for heterogeneous poster generation. Any2Poster Agent uses unified parsing to convert heterogeneous inputs into a shared structured representation, performs content-adaptive poster planning to assign panel roles and visual treatments, and renders posters through editable HTML/CSS with a lightweight VLM-guided repair loop for localized visual refinement.

We evaluate Any2Poster Agent on both Any2Poster Bench and the existing PaperQuiz-style paper-to-poster setting. On Any2Poster Bench, Any2Poster Agent achieves 87.25% average accuracy across eight input modalities and 87.28% average accuracy across five content domains, improving over GPT-5 by 3.35 points in cross-modality average accuracy and over GPT-4o by 11.78 points in cross-domain average accuracy. In the paper-to-poster setting, Any2Poster Agent achieves 72.58% overall accuracy and a 145.16 density-augmented score, improving over PosterAgent-4o from 51.06–51.33% to 72.58% overall accuracy and from 116–121 to 145.16 in density-augmented score.

Our contributions are:

- We introduce **Any2Poster Bench**, a benchmark for any-source poster generation across eight input modalities and five content domains.
- We propose **BenchQuiz**, a quiz-based evaluation protocol for verbatim factual retention and interpretive understanding, complemented by VLM-as-judge visual assessment.
- We provide **Any2Poster Agent**, a reference system combining unified input parsing, content-adaptive planning, HTML/CSS rendering, and VLM-guided repair.
- We show that Any2Poster Agent generalizes across modalities and domains on Any2Poster Bench and improves over prior paper-to-poster agents under PaperQuiz-style evaluation.

2. Related Work

Table 1 compares Any2Poster Agent with representative poster and presentation generation systems. Existing systems are typically designed for paper-only poster generation or slide-deck generation, whereas Any2Poster Agent targets heterogeneous inputs with unified parsing, content-adaptive poster planning, and editable HTML/CSS-based rendering.

System	Beyond Paper	Unified Parsing	Adaptive Planning	HTML/CSS Rendering
Paper2Poster (Pang et al., 2025)	✗	✗	✗	✗
Paper2Slides (HKUDS, 2025)	✓	✗	✗	✗
PPTAgent (Zheng et al., 2025)	✓	✗	✗	✗
Any2Poster Agent	✓	✓	✓	✓

Table 1. Comparison with representative poster and presentation generation systems.

Automatic poster generation has been studied in scientific, artistic, and commercial design settings. Scientific-poster systems such as PosterBot, Paper2Poster, P2P, PosterGen, PosterSum, and earlier probabilistic approaches focus on transforming academic papers into posters (Xu & Wan, 2022; Pang et al., 2025; Sun et al., 2025; Zhang et al., 2026; Saxena et al., 2025; Qiang et al., 2017). These methods provide useful insights into poster summarization, layout, and visual aesthetics, but they typically assume paper-specific inputs or narrower design settings. LLM-based agents that combine reasoning, tool use, and iterative repair are well suited to address these limitations, as poster generation requires parsing heterogeneous inputs, planning layouts, rendering editable outputs, and repairing visual failures. Evaluation is equally important because a poster must be both visually readable and faithful to its source. Paper2Poster introduces PaperQuiz and VLM-as-judge metrics for paper-to-poster evaluation (Pang et al., 2025). More broadly, model-based evaluation has been studied in G-Eval, MT-Bench/Chatbot Arena, and HELM, while VQA and GQA inspire question-answering evaluation for visual artifacts (Liu et al., 2023; Zheng et al., 2023; Liang et al., 2022; Antol et al., 2015; Hudson & Manning, 2019); unlike these benchmarks, which evaluate understanding over fixed inputs, Any2Poster Bench measures whether a generation system successfully transfers information from heterogeneous sources into a visual artifact. Any2Poster Bench extends these ideas to any-source poster generation by evaluating both quiz-based information recovery and VLM-as-judge visual communication quality across multiple modalities and content domains.

3. Any2Poster Bench

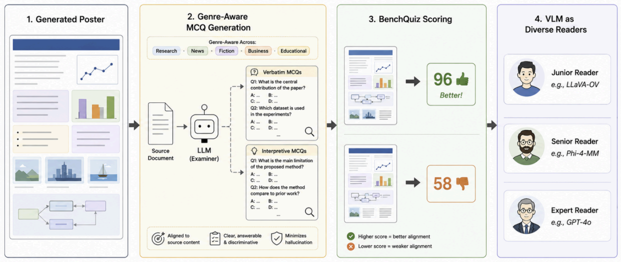


Figure 2. Overview of **Any2Poster Bench**. The benchmark covers eight input modalities and five content domains, and evaluates generated posters through *BenchQuiz*, a genre-aware multiple-choice reading-comprehension protocol answered by VLM readers.

We introduce **Any2Poster Bench**, a benchmark for evaluating any-source poster generation across heterogeneous source formats and content domains. As shown in Figure 2, the benchmark is designed to test whether a poster-generation system can handle not only different input modalities, but also different kinds of source material. The benchmark contains approximately 160 source documents spanning eight modalities: PDF, URL, PPTX, DOCX, Markdown, LaTeX, notebook, and video. It further covers five content domains: research, news, educational, business, and fiction. The goal is to move beyond paper-only poster evaluation and measure whether a system can robustly transform diverse real-world sources into informative, readable, and visually coherent single-page posters.

3.1. Benchmark Construction

The benchmark sources are selected according to three principles: *public accessibility*, *sufficient information density*, and *diversity in both modality and domain*. Public accessibility ensures that sources can be documented and, when licensing permits, redistributed or referenced. Information density ensures that each source contains enough substantive content to support a meaningful poster and a non-trivial reading-comprehension evaluation. Modality and domain diversity ensure that the benchmark tests more than paper-specific parsing or scientific summarization.

Research PDFs and LaTeX projects are drawn from public scientific sources, while news, educational, business, and fiction sources are collected from publicly accessible webpages, documents, slide decks, notebooks, and videos. Video inputs are included only when transcripts are available or can be reliably extracted. Because naturally occurring Markdown documents that are both long-form and content-rich are relatively rare outside software repositories, the Markdown subset is authored specifically for the benchmark to ensure sufficient information density and domain diversity.

All inputs are stored with a structured manifest that records modality, content domain, source, license or access information when available, and processing status. Before a source enters the benchmark, it is manually verified to ensure that the parser produces valid structured output and that the source contains enough substantive content to support a full poster. This verification step is important because any-source poster generation depends not only on visual generation quality, but also on upstream content extraction and structural recovery.

3.2. BenchQuiz Evaluation Protocol

To evaluate whether a generated poster preserves and communicates its source content, we construct **BenchQuiz**, a genre-aware reading-comprehension protocol inspired by PaperQuiz (Pang et al., 2025). For each source, an LLM examiner generates multiple-choice questions in two categories: *verbatim* and *interpretive*. Verbatim questions test whether explicitly stated facts can be recovered from the poster, such as names, numbers, datasets, methods, events, or claims. Interpretive questions test whether the poster communicates higher-level meaning, relationships, or takeaways. In our setup, the examiner generates 20 verbatim and 20 interpretive multiple-choice questions per source.

Questions are answered by VLM readers using only the generated poster. Scores are computed programmatically and aggregated by modality and by content domain. We use raw accuracy as the primary BenchQuiz metric because it directly measures information recoverability from the final poster. For compatibility with prior paper-to-poster evaluation, we also report density-augmented scoring when applicable, following the same principle used in Paper2Poster: a poster should be rewarded for preserving answerable information while remaining compact (Pang et al., 2025).

To reduce evaluation bias, answer options are randomly shuffled before each VLM answering pass so that the correct choice appears uniformly across positions. VLM readers are instructed to answer `NA` rather than guess when the poster does not contain enough evidence for a confident answer. This keeps the benchmark focused on what the poster actually communicates rather than what a strong model may infer from prior knowledge. The protocol is related to visual question-answering evaluation (Antol et al., 2015; Hudson & Manning, 2019), but differs in that the questions are source-grounded and are designed to test information transfer through the generated poster.

3.3. VLM-as-Judge Evaluation

BenchQuiz measures information recoverability, but posters must also be visually readable and well organized. We therefore complement BenchQuiz with a VLM-as-judge protocol following recent model-based evaluation practice (Liu et al., 2023; Zheng et al., 2023; Liang et al., 2022). The judge scores each poster along six criteria: element quality, layout balance, engagement, clarity, content completeness, and logical flow. We average element quality, layout balance, and engagement as the aesthetic score, and average clarity, content completeness, and logical flow as the information score. The overall score averages all six dimensions. Together, these metrics assess both informational effectiveness and visual communication quality.

4. Any2Poster Agent

4.1. Framework

Any2Poster Agent is a reference system for any-source poster generation. As shown in Figure 3, it converts heterogeneous source inputs into editable, visually coherent posters through three components: *unified parsing*, *content-adaptive poster planning*, and *HTML/CSS-based rendering with VLM-guided repair*. Internally, these components are instantiated through six operational stages: *parse*, *chunk*, *analyze*, *plan*, *generate*, and *compile*. The first component normalizes arbitrary inputs into a shared structured representation. The second component performs content understanding, weak-structure segmentation when needed, and poster planning conditioned on content type, content domain, and visual requirements. The third component renders the poster as editable HTML/CSS, exports browser-rendered outputs, and iteratively improves panel quality through localized VLM-guided edits.

This design differs from prior poster-generation and presentation-generation systems in three ways. First, it supports *unified input parsing* rather than restricting the pipeline to academic papers or a narrow template-retrieval setting. Second, it performs *content-adaptive planning*, so panel structure, visual style, and figure usage are selected according to the source content rather than imposed by a fixed template. Third, it uses *code as the rendering interface*: poster text is rendered as real HTML/CSS while visual assets are handled separately, enabling targeted editing and efficient VLM-guided refinement instead of expensive full-poster regeneration. By decoupling text rendering from visual asset generation and applying localized VLM-guided repair rather than full-poster regeneration, Any2Poster Agent is suitable as a strong and efficient baseline for Any2Poster Bench, where systems must handle diverse input modalities and content domains.

Code is available at `https://github.com/Any2Poster/Any2Poster`.

4.2. Unified Parsing

The first stage converts each supported input modality into a common `ParsedDocument` representation. We use dedicated parsers for different formats but normalize their outputs into the same schema, consisting of structured sections, extracted figures with captions, extracted tables, and raw text. PDF files are parsed with document-conversion tools inspired by recent work in document layout analysis and structured document understanding (Zhong et al., 2019; Pfitzmann et al., 2022; Blecher et al., 2023; Livathinos et al., 2025). DOCX and PPTX files are handled with dedicated document parsers. Markdown and plain text are parsed directly into sectioned text. URLs are processed with boilerplate removal. Video inputs are converted into transcripts using subtitles or speech transcription. We also support LaTeX projects and Jupyter notebooks.

The goal of this stage is not merely format conversion, but structural normalization. Regardless of whether the input is a paper, business report, story webpage, notebook, or slide deck, downstream stages receive the same type of object. This unified interface makes later reasoning more robust than format-specific pipelines, because planning and rendering operate over content structure instead of ad hoc modality-dependent heuristics.

4.3. Content-adaptive Poster Planning

The planning stage converts the parsed source into a poster-level content plan. A global analysis pass predicts the title, core message, section importance, panel categories, source figures to preserve, and candidate visual suggestions. A section-level pass then converts each section into panel-ready text and assigns a layout mode, such as prose only, bullets only, side by side, or figure dominant. For weakly structured inputs such as fiction or long-form essays, the agent first synthesizes narrative segments so that continuous prose does not collapse into a single oversized panel.

The planner also decides whether to preserve source visuals or synthesize new ones. Extracted figures are retained when they satisfy resolution and caption-relevance checks; otherwise, the agent generates a visual suggestion adapted to the content domain. This allows research inputs to preserve technical diagrams, while business, educational, and narrative inputs can use more explanatory or story-oriented visuals.

4.4. HTML/CSS-based Rendering

The generation and compilation stages use code, rather than an image model, as the primary interface for poster construction. All poster text is rendered in real HTML/CSS, which gives character-level fidelity that image-based text rendering often cannot guarantee. Each panel is instantiated as a structured HTML block whose layout is determined by the planning stage. Visuals are either preserved source figures or newly generated assets. For generated visuals, the system constructs prompts from the panel concept, visual

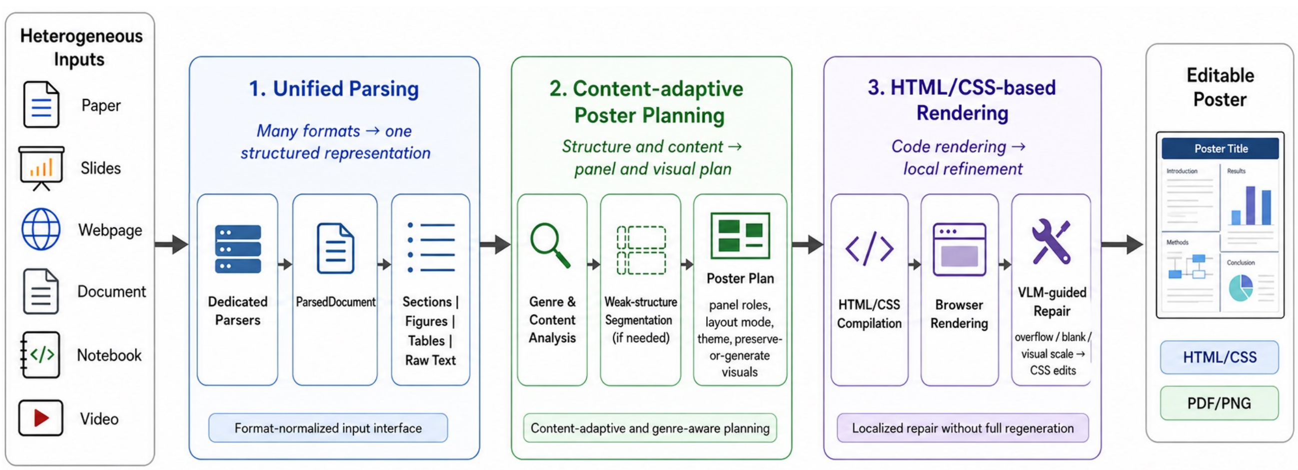


Figure 3. Overview of **Any2Poster Agent**. The pipeline has three main components: **(1) Unified Parsing**, which converts heterogeneous inputs into a shared structured representation; **(2) Content-adaptive Poster Planning**, which analyzes source content, assigns panel roles and layout modes, and decides whether to preserve or generate visuals; and **(3) HTML/CSS-based Rendering**, which compiles an editable poster and applies VLM-guided local repair without full regeneration.

type, validated numeric data, and active design theme.

This code-based interface is related to recent design-to-code and automatic layout-generation work (Lin et al., 2023; Yang et al., 2024; Si et al., 2025; Lu et al., 2023). In our setting, code-based rendering gives the agent fine-grained control over text, layout, and visual assets. Since these components are separated, the system can edit typography, spacing, and figure sizing directly through code without discarding the rest of the poster. The renderer embeds figures as data URIs, assembles a complete HTML poster, and uses a browser renderer to export both PDF and PNG outputs at poster resolution.

We also add a lightweight validation step for data-dependent visuals such as bar charts, line charts, comparisons, and tables. For these cases, the model is asked to return only data points explicitly supported by the source text; unsupported values are discarded before image generation. This reduces hallucinated quantitative content and improves the trustworthiness of synthesized visuals.

4.5. Visual Feedback Loop with VLMs

After rendering, Any2Poster Agent applies an optional VLM-based feedback loop for localized visual repair. The agent crops each poster panel, asks a VLM to detect issues such as overflow, excessive blank space, or poorly scaled visuals, and maps the diagnosis to targeted HTML/CSS edits before re-rendering. Unlike prior systems that rely on broader panel regeneration (Pang et al., 2025), this mechanism preserves already correct content while improving readability and layout quality. It also differs from presentation-generation systems that lack a comparable single-page poster repair loop (Zheng et al., 2025; HKUDS, 2025).

5. Experiments

We evaluate Any2Poster Agent on both Any2Poster Bench and the existing PaperQuiz-style paper-to-poster setting. The experiments address four questions: (i) whether Any2Poster Agent generalizes across heterogeneous input modalities, (ii) whether it generalizes across content domains, (iii) whether it remains competitive with prior paper-to-poster agents, and (iv) whether the generated posters preserve both recoverable information and visual communication quality. We use quiz-based accuracy as the primary measure of information recovery and complement it with VLM-as-judge assessment.

5.1. Experimental Setup

Any2Poster Bench evaluation. Any2Poster Bench evaluates any-source poster generation across eight input modalities: PDF, URL, PPTX, DOCX, Markdown, LaTeX, notebook, and video. For each evaluation instance, we generate 20 verbatim and 20 interpretive multiple-choice questions and measure the fraction answered correctly from the generated poster. Verbatim questions test whether factual details such as titles, datasets, numbers, and named methods are explicitly recoverable from the poster. Interpretive questions test whether the poster communicates higher-level understanding, such as the source's main contribution, conclusions, and implications. We aggregate accuracy by input modality and by content domain.

Baselines. We compare Any2Poster Agent with general-purpose multimodal models, including GPT-4o, Gemini 2.5 Flash, and GPT-5 (OpenAI, 2024; Comanici et al., 2025; OpenAI, 2025). We also compare with Paper2Poster when the input format is supported (Pang et al., 2025). Since Paper2Poster is designed for paper-like inputs, it is reported as N/A on unsupported modalities. For PaperQuiz-style evaluation, we compare against Paper2Poster-Qwen and PosterAgent-4o following the prior paper-to-poster setting (Pang et al., 2025; Bai et al., 2025).

VLM-as-judge evaluation. In addition to quiz-based information recovery, we use VLM-as-judge scoring to assess visual and communicative quality. The judge scores each poster along six criteria: element quality, layout balance, engagement, clarity, content completeness, and logical flow. We additionally report an aesthetic score, computed from element quality, layout balance, and engagement; an information score, computed from clarity, content completeness, and logical flow; and an overall score averaging all six dimensions.

5.2. Cross-Modality Evaluation

A central goal of Any2Poster Bench is to evaluate whether a poster-generation system can handle heterogeneous source formats rather than only scientific papers. Table 2 reports quiz accuracy across eight input modalities. Any2Poster Agent achieves the best average accuracy, 87.25%, outperforming GPT-4o, Gemini 2.5 Flash, GPT-5, and Paper2Poster. The strongest results are obtained on Markdown and LaTeX inputs, both above 93%. PPTX remains the most challenging modality for Any2Poster Agent at 75.00%, suggesting that slide inputs require further layout- and structure-aware parsing. Although GPT-5 performs slightly better on PDF and PPTX, Any2Poster Agent is more robust on average across the full modality spectrum.

5.3. Cross-Domain Evaluation

We next evaluate whether Any2Poster Agent generalizes across content domains. Table 3 reports quiz accuracy across fiction, research, business, educational, and news sources. Any2Poster Agent achieves an average accuracy of 87.28%, compared with 75.50% for GPT-4o and 68.50% for Paper2Poster on its supported domains. Performance is consistent across domains, with all domain-level accuracies above 86%. These results show that Any2Poster Bench evaluates a broader capability than scientific paper summarization alone: systems must extract, organize, and visually communicate information from sources with different discourse structures.

5.4. Comparison with Prior Paper-to-Poster Agents

Table 4 compares Any2Poster Agent with prior paper-to-poster agent baselines under PaperQuiz-style evaluation. Any2Poster Agent achieves substantially higher overall accuracy than both Paper2Poster-Qwen and PosterAgent-4o. In particular, it improves over PosterAgent-4o from 51.06–51.33% to 72.58% overall accuracy and from 116–121 to 145.16 in density-augmented score. The gain is especially large on interpretive questions, increasing from 49.17% for PosterAgent-4o to 84.00% for Any2Poster Agent. This suggests that Any2Poster Agent better communicates the high-level meaning of the source rather than only preserving isolated facts.

5.5. VLM-as-Judge Results

Table 5. VLM-as-judge evaluation. Element, layout, and engagement form the aesthetic score; clarity, content completeness, and logical flow form the information score. Higher is better.

Method	Elem.	Layout	Engage.	Clarity	Content	Flow	Aesthetic	Info.	Overall
Paper2Poster	3.94	3.77	2.91	3.99	3.91	3.64	3.54	3.85	3.69
Any2Poster Agent	**4.04**	**3.93**	**3.13**	**4.04**	**4.07**	**4.96**	**3.70**	**4.36**	**4.03**

Table 5 reports VLM-as-judge results. Compared with Paper2Poster, Any2Poster Agent improves the overall VLM-as-judge score from 3.69 to 4.03. The largest gain appears in logical flow, which increases from 3.64 to 4.96, indicating that the generated poster better organizes information into a coherent visual narrative. Any2Poster Agent also improves element quality, layout balance, engagement, clarity, content completeness, aesthetic score, and information score.

5.6. Ablation Study

We ablate two key components of Any2Poster Agent: unified input parsing and visual feedback refinement. Removing unified input parsing reduces verbatim accuracy from 60.80% to 31.00%, showing that structure-aware parsing is critical for preserving fine-grained source facts. Removing visual feedback refinement reduces overall accuracy from 72.58% to 55.50%, indicating that visual quality control directly affects whether downstream readers can recover both factual and interpretive information. The full ablation table is provided in Appendix E.

We report qualitative examples and token consumption in Appendices G and F. Overall, the results show that Any2Poster Bench reveals capabilities not captured by paper-only benchmarks, and that Any2Poster Agent provides a strong reference system for any-source poster generation.

Table 2. Cross-modality quiz accuracy (%) on Any2Poster Bench. Each modality is evaluated using verbatim and interpretive questions. N/A indicates that the method does not support the corresponding input modality.

Method	PDF	URL	PPTX	DOCX	MD	LaTeX	Notebook	Video	Avg.
GPT-4o	75.50	72.50	74.00	75.00	72.50	80.00	62.50	80.00	74.00
Gemini 2.5 Flash	72.50	71.20	73.80	73.80	80.00	80.00	79.00	78.80	76.14
GPT-5	**86.20**	82.50	**78.80**	82.50	85.20	83.70	85.00	87.40	83.90
Paper2Poster	57.60	N/A	N/A	N/A	N/A	59.00	N/A	N/A	58.30
Any2Poster Agent	85.63	**88.13**	75.00	**86.88**	**93.00**	**93.13**	**86.88**	**89.38**	**87.25**

Table 3. Cross-domain quiz accuracy (%) on Any2Poster Bench. Higher is better. †Paper2Poster average is computed over supported domains only because the educational setting is unsupported.

Method	Fiction	Research	Business	Educational	News	Avg.
GPT-4o	80.00	72.50	82.50	62.50	80.00	75.50
Paper2Poster	61.00	74.00	68.00	N/A	71.00	68.50†
Any2Poster Agent	**89.38**	**86.57**	**87.15**	**86.79**	**86.50**	**87.28**

Table 4. PaperQuiz-style comparison with prior paper-to-poster agents. Verbatim, interpretive, and overall values are accuracies in percentage points. Higher is better for all metrics.

Method	Verbatim	Interpretive	Overall Acc.	Density-Aug.
Paper2Poster-Qwen	51.81	48.79	50.30–50.57	114–115
PosterAgent-4o	52.95	49.17	51.06–51.33	116–121
Any2Poster Agent	**60.80**	**84.00**	**72.58**	**145.16**

6. Conclusion

We introduced Any2Poster Bench, a benchmark for evaluating any-source poster generation across eight input modalities and five content domains. By combining BenchQuiz with VLM-as-judge assessment, Any2Poster Bench measures both information recoverability and visual communication quality. We further provided Any2Poster Agent as a reference system that combines unified parsing, content-adaptive planning, HTML/CSS rendering, and VLM-guided repair. Experiments show that Any2Poster Agent generalizes across modalities and domains on Any2Poster Bench and improves over prior paper-to-poster agents under PaperQuiz-style evaluation. We hope Any2Poster Bench supports future work on reliable, multimodal, and domain-general visual communication systems.

Acknowledgements

We thank the reviewers for their helpful feedback. We also acknowledge Princeton University for supporting the computational resources used in this work.

References

Akhtar, M., Benjelloun, O., Conforti, C., Foschini, L., Giner-Miguelez, J., et al. Croissant: A metadata format for ml-ready datasets, 2024. URL https://arxiv.org/abs/2403.19546.

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., and Parikh, D. VQA: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, 2015.

Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., et al. Qwen2.5-VL technical report, 2025. URL https://arxiv.org/abs/2502.13923.

Bandyopadhyay, S., Maheshwari, H., Natarajan, A., and Saxena, A. Enhancing presentation slide generation by LLMs with a multi-staged end-to-end approach. In *Proceedings of the 17th International Natural Language Generation Conference*, pp. 222–229. Association for Computational Linguistics, 2024. URL https://aclanthology.org/2024.inlg-main.18/.

Blecher, L., Cucurull, G., Scialom, T., and Stojnic, R. Nougat: Neural optical understanding for academic documents. *arXiv preprint arXiv:2308.13418*, 2023. URL https://arxiv.org/abs/2308.13418.

Chen, H., Xu, X., Li, W., Ren, J., Ye, T., Liu, S., Chen, Y.-C., Zhu, L., and Wang, X. POSTA: A go-to framework for customized artistic poster generation. *arXiv preprint arXiv:2503.14908*, 2025. URL https://arxiv.org/abs/2503.14908.

Comanici, G. et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. URL https://arxiv.org/abs/2507.06261.

Gao, L., Madaan, A., Zhou, S., Alon, U., Liu, P., Yang, Y., Callan, J., and Neubig, G. PAL: Program-aided language models. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., and Crawford, K. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

HKUDS. Paper2Slides: From paper to presentation in one click. https://github.com/HKUDS/Paper2Slides, 2025. Software project.

Hudson, D. A. and Manning, C. D. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6700–6709, 2019.

Jin, C., Xu, H., Song, R., and Lu, Z. Text2Poster: Laying out stylized texts on retrieved images. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4823–4827. IEEE, 2022. doi: 10.1109/ICASSP43922.2022.9747465. URL https://doi.org/10.1109/ICASSP43922.2022.9747465.

Kaluarachchi, T. and Wickramasinghe, M. WebDraw: A machine learning-driven tool for automatic website prototyping. *Science of Computer Programming*, 233: 103056, 2024. doi: 10.1016/j.scico.2023.103056. URL https://www.sciencedirect.com/science/article/pii/S0167642323001387.

Li, F., Liu, A., Feng, W., Zhu, H., Li, Y., Zhang, Z., Lv, J., Zhu, X., Shen, J., Lin, Z., and Shao, J. Relation-aware diffusion model for controllable poster layout generation. *arXiv preprint arXiv:2306.09086*, 2023. URL https://arxiv.org/abs/2306.09086.

Li, Z., Li, F., Feng, W., Zhu, H., Li, Y., Zhang, Z., Lv, J., Shen, J., Lin, Z., Shao, J., and Yang, Z. Planning and rendering: Towards product poster generation with diffusion models. *arXiv preprint arXiv:2312.08822*, 2024. URL https://arxiv.org/abs/2312.08822.

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., et al. Holistic evaluation of language models, 2022. URL https://arxiv.org/abs/2211.09110.

Lin, J., Guo, J., Sun, S., Yang, Z. J., Lou, J.-G., and Zhang, D. LayoutPrompter: Awaken the design ability of large language models. *arXiv preprint arXiv:2311.06495*, 2023. URL https://arxiv.org/abs/2311.06495.

Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., and Zhu, C. G-Eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522, 2023.

Livathinos, N., Auer, C., Lysak, M., Nassar, A., Dolfi, M., Vagenas, P., Ramis, C. B., Omenetti, M., Dinkla, K., Kim, Y., Gupta, S., de Lima, R. T., Weber, V., Morin, L., Meijer, I., Kuropiatnyk, V., and Staar, P. W. J. Docling: An efficient open-source toolkit for ai-driven document conversion. *arXiv preprint arXiv:2501.17887*, 2025. URL https://arxiv.org/abs/2501.17887.

Lu, P., Chen, B., Liu, S., Thapa, R., Boen, J., and Zou, J. OctoTools: An agentic framework with extensible tools for complex reasoning. *arXiv preprint arXiv:2502.11271*, 2025. URL https://arxiv.org/abs/2502.11271.

Lu, Y., Tong, Z., Zhao, Q., Zhang, C., and Li, T. J.-J. UI layout generation with LLMs guided by UI grammar. *arXiv preprint arXiv:2310.15455*, 2023. URL https://arxiv.org/abs/2310.15455.

Ma, J., Deng, Y., Chen, C., Du, N., Lu, H., and Yang, Z. GlyphDraw2: Automatic generation of complex glyph posters with diffusion models and large language models. *arXiv preprint arXiv:2407.02252*, 2024. URL https://arxiv.org/abs/2407.02252.

Mondal, I., S, S., Natarajan, A., Garimella, A., Bandyopadhyay, S., and Boyd-Graber, J. Presentations by the humans and for the humans: Harnessing LLMs for generating persona-aware slides from documents. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 2664–2684. Association for Computational Linguistics, 2024. URL https://aclanthology.org/2024.eacl-long.163/.

OpenAI. GPT-4o system card. https://openai.com/index/gpt-4o-system-card/, 2024.

OpenAI. GPT-5 system card. https://openai.com/index/gpt-5-system-card/, 2025.

Pang, W., Lin, K. Q., Jian, X., He, X., and Torr, P. Paper2Poster: Towards multimodal poster automation from scientific papers, 2025. URL https://arxiv.org/abs/2505.21497.

Paranjape, B., Lundberg, S., Singh, S., Hajishirzi, H., Zettlemoyer, L., and Ribeiro, M. T. ART: Automatic multi-step reasoning and tool-use for large language models. *arXiv preprint arXiv:2303.09014*, 2023. URL https://arxiv.org/abs/2303.09014.

Paruchuri, V. marker: Convert PDF to Markdown and JSON Quickly with High Accuracy. https://github.com/VikParuchuri/marker, 2025. Software project.

Pfitzmann, B., Auer, C., Dolfi, M., Nassar, A. S., and Staar, P. W. J. DocLayNet: A large human-annotated dataset for document-layout analysis. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 3743–3751, 2022.

Pushkarna, M., Zaldivar, A., and Kjartansson, O. Data cards: Purposeful and transparent dataset documentation for responsible ai. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1776–1826, 2022.

Qiang, Y.-t., Fu, Y., Yu, X., Guo, Y., Zhou, Z.-H., and Sigal, L. Learning to generate posters of scientific papers by probabilistic graphical models, 2017. URL https://arxiv.org/abs/1702.06228.

Rodriguez, J. A., Jian, X., Panigrahi, S. S., Zhang, T., Feizi, A., Puri, A., Suresh, A. K., Savard, F., Masry, A., Nayak, S., et al. BigDocs: An open dataset for training multimodal models on document and code tasks. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=b1ivBPLb1n.

Saxena, R., Minervini, P., and Keller, F. PosterSum: A multimodal benchmark for scientific poster summarization. *arXiv preprint arXiv:2502.17540*, 2025. URL https://arxiv.org/abs/2502.17540.

Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Hambro, E., Zettlemoyer, L., Cancedda, N., and Scialom, T. Toolformer: Language models can teach themselves to use tools. *arXiv preprint arXiv:2302.04761*, 2023. URL https://arxiv.org/abs/2302.04761.

Sefid, A., Mitra, P., and Giles, L. SlideGen: An abstractive section-based slide generator for scholarly documents. In *Proceedings of the 21st ACM Symposium on Document Engineering*. Association for Computing Machinery, 2021. doi: 10.1145/3469096.3474939. URL https://doi.org/10.1145/3469096.3474939.

Si, C., Zhang, Y., Li, R., Yang, Z., Liu, R., and Yang, D. Design2Code: Benchmarking multimodal code generation for automated front-end engineering. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3956–3974. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.naacl-long.199/.

Sun, E., Hou, Y., Wang, D., Zhang, Y., and Wang, N. X. R. D2S: Document-to-slide generation via query-based text summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1405–1418. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.naacl-main.111. URL https://aclanthology.org/2021.naacl-main.111/.

Sun, T., Pan, E., Yang, Z., Sui, K., Shi, J., Cheng, X., Li, T., Huang, W., Zhang, G., Yang, J., and Li, Z. P2P: Automated paper-to-poster generation and fine-grained benchmark, 2025. URL https://arxiv.org/abs/2505.17104.

Xu, S. and Wan, X. PosterBot: A system for generating posters of scientific papers with neural models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 13233–13235, 2022. doi: 10.1609/aaai.v36i11.21738. URL https://ojs.aaai.org/index.php/AAAI/article/view/21738.

Yang, T., Luo, Y., Qi, Z., Wu, Y., Shan, Y., and Chen, C. W. PosterLLaVA: Constructing a unified multimodal layout generator with LLM. *arXiv preprint arXiv:2406.02884*, 2024. URL https://arxiv.org/abs/2406.02884.

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K. R., and Cao, Y. ReAct: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

Zhang, Z., Zhang, X., Wei, J., Xu, Y., and You, C. PosterGen: Aesthetic-aware paper-to-poster generation via multi-modal large language model. In *Findings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://arxiv.org/abs/2508.17188.

Zheng, H., Guan, X., Kong, H., Zheng, J., Zhou, W., Lin, H., Lu, Y., He, B., Han, X., and Sun, L. PPTAgent: Generating and evaluating presentations beyond text-to-slides. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 914–934. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.53. URL https://aclanthology.org/2025.emnlp-main.53/.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., and Stoica, I. Judging LLM-as-a-judge with MT-bench and chatbot arena, 2023. URL https://arxiv.org/abs/2306.05685.

Zhong, X., Tang, J., and Jimeno Yepes, A. PubLayNet: Largest dataset ever for document layout analysis. In *2019 International Conference on Document Analysis and Recognition*, pp. 1015–1022, 2019.

A. Dataset Details

A.1. Benchmark Composition

Any2Poster Bench is designed to evaluate any-source poster generation across both input modality and content domain. The benchmark contains approximately 160 source documents spanning eight input modalities—PDF, URL, PPTX, DOCX, Markdown, LaTeX, notebook, and video—and five content domains: research, news, educational, business, and fiction.

Of these, 32 fully evaluated instances are released here, with one instance for each modality–domain cell, covering all eight modalities and all five domains. The remaining instances are withheld because their source licenses do not permit redistribution of extracted text, such as news articles, institutional presentations, and proprietary reports. Poster outputs and evaluation scores for those withheld instances will be included in the full public release.

Source attribution and URLs for all 32 released instances are provided in `benchmark_manifest.csv`. Each source is paired with BenchQuiz questions and metadata describing its modality, content domain, source type, processing status, and associated evaluation files.

The benchmark construction follows three principles. First, sources should be publicly accessible or otherwise documentable. Second, each source should contain enough information density to support a meaningful single-page poster. Third, the collection should cover heterogeneous structures, including well-structured scientific papers, semi-structured documents and slide decks, webpage content, executable notebooks, video transcripts, and weakly structured long-form prose. This diversity is intended to test whether a poster-generation system can generalize beyond paper-only inputs.

A.2. Source Processing

Each source is converted into a structured representation before poster generation. For document-like inputs, the parser extracts section text, figures, captions, tables, and raw text when available. For URL inputs, boilerplate content is removed before extraction. For video inputs, transcripts are obtained from subtitles or speech transcription when available. For notebooks, markdown cells, code cells, outputs, and figures are converted into a document-like representation. For LaTeX inputs, source files are parsed into sectioned text and associated assets when possible.

Before inclusion in the benchmark, each source is manually checked to ensure that the parser produces valid structured output and that the source contains enough substantive content to support poster generation and question-answering evaluation. Sources that fail parsing or contain insufficient content are excluded or replaced.

A.3. Dataset Documentation and Responsible Release

Any2Poster Bench is intended for evaluating systems that transform diverse source materials into single-page visual posters. It is not intended to evaluate factual correctness beyond the provided source content, nor should it be used as a general-purpose measure of visual design ability independent of information fidelity. The benchmark also does not claim to exhaustively cover all possible poster domains, languages, or visual styles; rather, it provides a controlled testbed for measuring cross-modality and cross-domain generalization in poster generation.

Because Any2Poster Bench is a dataset-centered contribution, dataset documentation and responsible release are important. Following documentation practices such as Datasheets for Datasets, Data Cards, and Croissant metadata (Gebru et al., 2021; Pushkarna et al., 2022; Akhtar et al., 2024), the release includes a manifest documenting source metadata, modality, content domain, processing status, and evaluation files. For dataset-centered submission and release, we also prepare Croissant-compatible metadata and Responsible AI metadata describing intended uses, limitations, source provenance, licensing considerations when available, and recommended evaluation practices.

B. Extended Related Work

Any-source poster generation requires robust document understanding: tools such as PubLayNet, DocLayNet, Nougat, Docling, marker, and BigDocs support layout analysis, document conversion, and multimodal document understanding (Zhong et al., 2019; Pfitzmann et al., 2022; Blecher et al., 2023; Livathinos et al., 2025; Paruchuri, 2025; Rodriguez et al., 2025), motivating the unified parsing stage of Any2Poster Agent, which normalizes heterogeneous inputs into a shared structured representation.

Poster generation is also related to document-to-slide and presentation generation. D2S, SlideGen, multi-stage slide-generation pipelines, persona-aware slide generation, and PPTAgent study how to convert documents into presentations (Sun et al., 2021; Sefid et al., 2021; Bandyopadhyay et al., 2024; Mondal et al., 2024; Zheng et al., 2025). However, slides distribute information across multiple pages, whereas posters compress content into a single canvas. This makes poster generation especially sensitive to information density, spatial hierarchy, local readability, and cross-panel balance. Other work studies stylized, artistic, glyph, product, or controllable poster design (Jin et al., 2022; Chen et al., 2025; Ma et al., 2024; Li et al., 2024; 2023).

Poster construction also requires visual layout and rendering. LayoutPrompter, PosterLLaVA, and UI layout generation work study how language or multimodal models generate structured layouts (Lin et al., 2023; Yang et al., 2024; Lu et al., 2023). Design2Code and WebDraw are related to our code-based rendering interface, where visual artifacts are represented as executable front-end code (Si et al., 2025; Kaluarachchi & Wickramasinghe, 2024). Unlike these works, any-source poster generation must jointly extract source information, summarize it, select or synthesize visuals, and render an editable single-page poster.

LLM-based agents combine reasoning, tool use, code generation, and iterative repair. ReAct, Toolformer, PAL, ART, and OctoTools show that language models can benefit from decomposing tasks, calling tools, and executing intermediate programs (Yao et al., 2023; Schick et al., 2023; Gao et al., 2023; Paranjape et al., 2023; Lu et al., 2025). This paradigm is well suited to poster generation, where a system must parse heterogeneous inputs, plan layouts, render editable outputs, inspect visual artifacts, and repair failures such as overflow or poorly scaled figures.

C. BenchQuiz Prompts and Rubrics

C.1. Question Generation

BenchQuiz evaluates whether a generated poster preserves and communicates the source content. For each source, an LLM examiner generates 20 verbatim and 20 interpretive multiple-choice questions. Verbatim questions target explicitly stated source facts, such as names, numbers, datasets, methods, events, or claims. Interpretive questions target higher-level understanding, such as main takeaways, relationships between concepts, conclusions, implications, or domain-specific reasoning.

A representative prompt template for question generation is:

> You are constructing a reading-comprehension quiz for evaluating whether a visual poster preserves information from a source document. Given the source content, generate multiple-choice questions that can be answered only if the poster contains the relevant information. Generate two categories of questions: (1) verbatim questions that ask about explicitly stated facts, names, numbers, methods, events, or claims; and (2) interpretive questions that ask about higher-level takeaways, relationships, implications, or conclusions. Each question should have one correct answer and multiple plausible distractors. Avoid questions that require external knowledge.

C.2. VLM Reader Protocol

Generated posters are evaluated by VLM readers using only the poster image. The reader receives the poster and one multiple-choice question at a time. The answer options are randomly shuffled before each pass to reduce positional bias. Readers are instructed to answer `NA` if the poster does not contain enough evidence for a confident answer, rather than guessing. Accuracy is computed programmatically by comparing the predicted option with the ground-truth answer.

A representative VLM-reader instruction is:

> You are answering a multiple-choice question using only the provided poster. Read the poster carefully. Select the option that is best supported by the poster. If the poster does not contain enough information to answer confidently, respond with `NA`. Do not rely on outside knowledge.

C.3. VLM-as-Judge Rubric

For visual and communicative quality, we use a VLM-as-judge rubric with six criteria:

- **Element Quality**: clarity and consistency of figures, graphics, icons, and visual elements.
- **Layout Balance**: alignment, spacing, panel organization, and overall structural balance.
- **Engagement**: visual appeal and ability to attract attention.
- **Clarity**: readability of text and phrasing.
- **Content Completeness**: coverage of essential source information.
- **Logical Flow**: coherence of the narrative across poster panels.

We compute an aesthetic score by averaging element quality, layout balance, and engagement. We compute an information score by averaging clarity, content completeness, and logical flow. The overall VLM-as-judge score averages all six criteria.

D. Implementation Details

D.1. Any2Poster Agent

Any2Poster Agent follows a parse–chunk–analyze–plan–generate–compile pipeline. The parser converts heterogeneous inputs into a shared `ParsedDocument` schema. After parsing, long sections are split into chunks of roughly 800 tokens. Chunk boundaries prefer paragraph breaks and topic transitions, while equations and algorithmic blocks are kept intact when possible. Each chunk stores a small left and right context window so later LLM calls can access local context without repeatedly processing the full source document. The global analysis stage predicts the poster title,

Table 6. Ablation study of Any2Poster Agent under the PaperQuiz-style setting with 50 verbatim and 50 interpretive questions. Higher is better for all metrics.

Variant	Verbatim	Interpretive	Overall Acc.	Density-Aug.
Full Any2Poster Agent	**60.80**	**84.00**	**72.58**	**145.16**
w/o unified input parsing	31.00	83.00	58.00	114.00
w/o visual feedback refinement	42.00	69.00	55.50	111.00

core message, section importance, candidate panels, and source figures to preserve. The section-level analysis stage produces panel-ready summaries, panel roles, layout modes, and candidate visual suggestions.

The poster is rendered as HTML/CSS. Text is rendered directly in HTML to preserve character-level fidelity, while visual assets are either extracted from the source or generated from panel-specific prompts. The final poster is compiled through a browser renderer and exported as PDF and PNG. A VLM-based feedback loop optionally crops poster panels, diagnoses common visual issues, and applies localized HTML/CSS edits.

D.2. Visual Feedback Rules

The VLM feedback loop returns coarse diagnostic labels for each panel: *overflow*, *too blank*, *visual too small*, *visual too large*, or *good*. These labels are mapped to deterministic repair actions. Overflow reduces font scale and may remove the least important final bullet. Blank panels increase visual prominence or spacing. Undersized figures receive larger flex allocation, while oversized figures are scaled down. The poster is then re-rendered. This local repair strategy avoids regenerating the full poster and helps preserve already correct content.

E. Ablation Study

Removing unified input parsing causes the largest drop in verbatim accuracy, indicating that structured source recovery is especially important for fine-grained facts. Removing visual feedback refinement also reduces overall accuracy, showing that visual quality control affects whether downstream readers can recover information from the final poster.

F. Efficiency

We report token consumption to characterize the computational cost of different poster-generation pipelines. As shown in Table 7, Any2Poster Agent-4o uses 138.34K tokens per run on average, mainly from LLM calls. This is higher than PosterAgent-4o and the direct 4o-HTML baseline, but lower than PPTAgent-4o and OWL-4o. The direct 4o-HTML baseline has the lowest token consumption, but it reflects a simpler direct-generation setting rather than a full agentic poster-generation pipeline. Overall, Any2Poster Agent-4o has moderate computational cost while supporting heterogeneous input parsing, poster planning, generation, compilation, and feedback refinement.

Table 7. Average token consumption per run across poster-generation methods. Lower is better.

Method	Avg. Tokens
4o-HTML	20.67K
PosterAgent-Qwen	47.55K
PosterAgent-4o	101.10K
Any2Poster Agent-4o	138.34K
PPTAgent-4o	255.73K
OWL-4o	361.10K

G. Additional Qualitative Examples

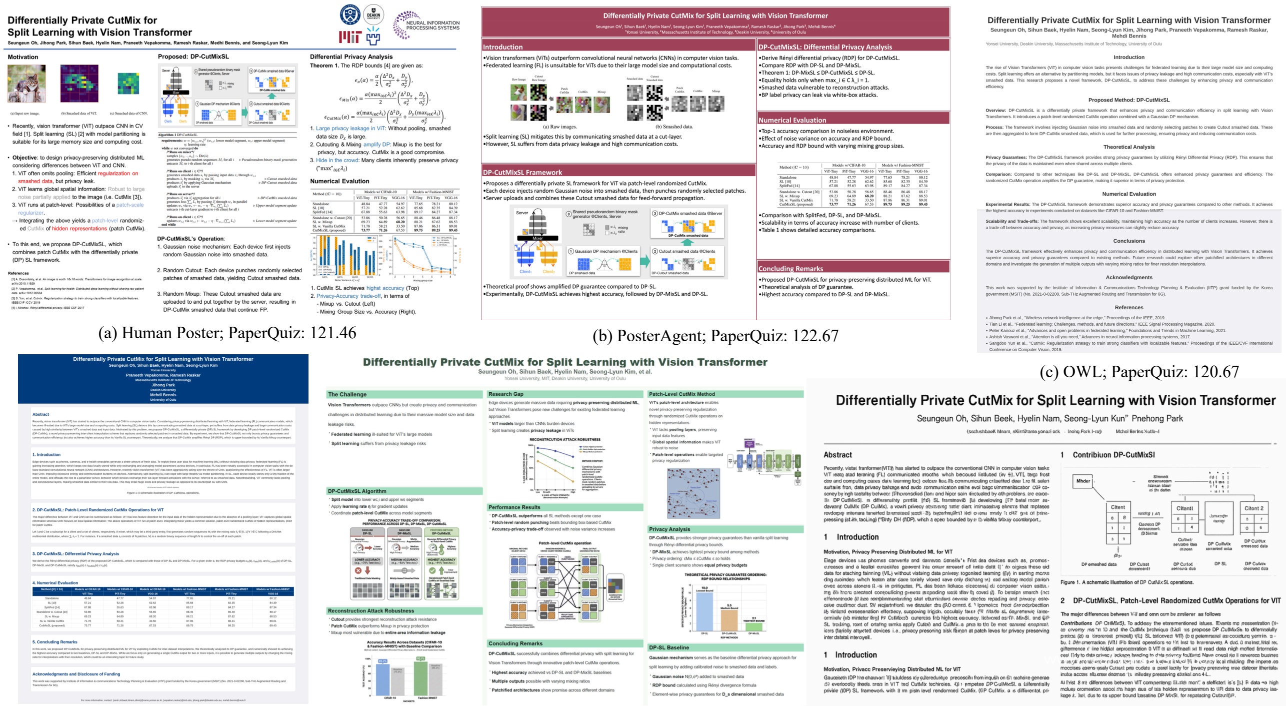
We include additional qualitative examples to illustrate the diversity of Any2Poster Bench across modalities and domains. These examples are intended to show how poster-generation systems handle heterogeneous source structures, such as technical papers, webpages, notebooks, slide decks, and narrative prose.

H. Limitations

Any2Poster Bench broadens poster-generation evaluation beyond paper-only inputs, but it is not exhaustive. The benchmark currently focuses on English-language sources and a finite set of eight input modalities and five content domains. Future versions could expand to more languages, more specialized domains, and additional source types such as interactive webpages, spreadsheets, or multi-file project repositories.

The evaluation protocol also has limitations. BenchQuiz measures whether information can be recovered from a generated poster by VLM readers, but VLM performance may vary across models and may not perfectly match human reader behavior. VLM-as-judge scores provide scalable visual-quality assessment, but they may reflect model-specific preferences and should not be interpreted as a complete replacement for human evaluation. To reduce bias, we randomize answer-option order and allow `NA` responses, but residual model bias may remain.

Any2Poster Agent is a reference system rather than a final solution to any-source poster generation. Its performance depends on parser quality, LLM planning quality, visual-generation quality, and the reliability of the VLM feedback loop. Difficult inputs such as heavily formatted slide decks, noisy videos, documents with complex tables, or sources requiring precise visual reproduction remain challenging. In addition, the use of large multimodal models introduces computational cost and may limit accessibility for some



(a) Human Poster; PaperQuiz: 121.46

(b) PosterAgent; PaperQuiz: 122.67

(c) OWL; PaperQuiz: 120.67

(d) 4o-HTML; PaperQuiz: 116.02

(e) Any2Poster; PaperQuiz: 147.0

(f) 4o-Image; PaperQuiz: 107.67

Figure 4. Qualitative examples comparing generated posters from different systems on the same source.

users.

I. Responsible Release

The benchmark is intended for research on evaluating and improving poster generation systems. It should not be used to claim factual correctness beyond the provided source content, nor should it be used as a standalone measure of general design ability. For release, we document source provenance, modality, content domain, processing status, and evaluation files. When source licensing does not permit redistribution, we provide metadata or pointers rather than redistributing the original content.

We also prepare Croissant-compatible metadata and Responsible AI metadata describing intended uses, limitations, source provenance, licensing considerations when available, and recommended evaluation practices. Generated posters may summarize or visually transform copyrighted source materials, so users should ensure that downstream usage complies with the source license and applicable policies.