PANEL2PAGE: STRUCTURED VISUAL NARRATIVE GENERATION WITH OMNI-PANEL ALIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Structured visual generation requires models to satisfy hierarchical constraints beyond global text-image alignment, including global composition, local semantic correspondence, and cross-panel relational consistency. We study this challenge through text-to-comic page generation, where a model directly synthesizes a complete comic page from a structured prompt specifying layouts, ordered events, recurring characters, textual elements, and cross-panel relations. We propose Panel2Page, a two-stage framework that improves structured correspondence by separating global composition learning from hierarchical semantic alignment while preserving single-pass generation. Page-Level Supervised Fine-Tuning first learns a global structural prior from complete prompt-page pairs, capturing layout organization, reading order, and page-level composition. We then introduce Omni-Panel Alignment (OPA), a training-time strategy that regularizes unified page predictions using detached localized predictions constructed from complementary text-aware, panel-event, and cross-panel views, without introducing additional inference modules. To enable structured supervision, we construct a hierarchical dataset of approximately 20,000 manually verified prompt-page pairs and a benchmark of 320 structured prompts covering diverse layouts and narratives. Extensive experiments demonstrate that Panel2Page achieves the best overall performance among the evaluated full-page and panel-wise generation methods, showing the complementary benefits of global structural learning and hierarchical semantic alignment.

1 INTRODUCTION

Recent advances in text-to-image generation have achieved remarkable progress in synthesizing high-quality images from natural-language descriptions Nichol et al. (2021); Rombach et al. (2022); Podell et al. (2023); Wu et al. (2025a); Z-Image Team et al. (2025). However, generating structured visual content remains challenging because the output must satisfy multiple interconnected constraints beyond global image–text correspondence.

Unlike conventional image synthesis, structured visual generation requires models to coordinate constraints across multiple semantic granularities, including global organization, local semantic correspondence, and cross-panel relational consistency. For example, a comic page must simultaneously preserve layout organization, panel-level events, textual fidelity, and narrative consistency across panels. Existing text-to-image models mainly optimize global image–text alignment and may produce visually plausible results while omitting requested events, changing recurring identities, or misplacing textual elements. This motivates training objectives that explicitly connect structured semantic units to their corresponding visual regions while retaining full-page generation.

We study this problem through text-to-comic page generation as a representative structured visual narrative task. A comic page naturally exhibits hierarchical semantic organization: the page defines global composition and reading order, individual panels describe localized events, and relations across panels describe narrative transitions and recurring entities. Given a structured prompt specifying the page layout, ordered panel events, recurring characters, textual elements, and cross-panel relations, the goal is to directly generate a complete comic page within a unified canvas. This setting provides a challenging testbed for studying how generative models coordinate global structure, local semantics, and cross-panel dependencies.

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

Existing approaches address only partial aspects of these requirements. Controllable text-to-image methods improve spatial grounding Gafni et al. (2022); Li et al. (2023); Chefer et al. (2023); Du et al. (2026), while text-rendering methods Ma et al. (2023); Tuo et al. (2023); Chen et al. (2023a;b) and visual storytelling approaches improve textual fidelity or consistency across generated sequences Li et al. (2018); Maharana et al. (2022); Rahman et al. (2022); Gong et al. (2023); Zhou et al. (2024); Tao et al. (2024); Park et al. (2025); Dinkevich et al. (2025). However, these methods typically focus on local control, sequential image generation, or post-processing, without explicitly aligning global page structure, localized semantic units, and cross-panel relationships within a unified generation process.

To address this challenge, we propose **Panel2Page**, a two-stage framework for structured visual generation that separates global composition learning from semantic-unit-aware hierarchical alignment. Stage 1, **Page-Level Supervised Fine-Tuning**, learns a global structural prior from complete prompt–page pairs, capturing layout organization, reading order, and page-level composition. Stage 2 introduces **Omni-Panel Alignment (OPA)**, a training-time alignment strategy that regularizes unified page predictions using detached localized predictions constructed from complementary text-aware, panel-event, and cross-panel views. Page-Level SFT uses standard domain adaptation; the methodological focus is OPA's construction of matched semantic conditions and spatial supports. OPA belongs to the family of consistency regularization methods: the shared model supplies detached local targets, while the full-page denoising objective anchors training to the data. Its text, panel, and relational views constrain a single page prediction without adding an inference-time teacher or composition module.

To support this study, we construct approximately 20,000 manually verified prompt–page pairs with hierarchical annotations derived from structured prompts. We further establish a benchmark of 320 structured prompts covering four regular-grid layouts and four narrative categories.

Our contributions are summarized as follows:

- We formulate structured visual generation as a hierarchical semantic alignment problem and propose **Panel2Page**, a unified framework that combines global structural learning with semantic-unit-aware alignment for complete-page generation.
- We introduce **Omni-Panel Alignment (OPA)**, a training-time alignment strategy that regularizes unified page predictions using detached localized predictions from text-aware, panel-event, and cross-panel views while preserving the original single-pass inference process.
- We construct a manually verified hierarchical prompt–page dataset and establish a page-oriented benchmark and evaluation protocol covering textual fidelity, semantic correspondence, narrative consistency, and page quality.

2 Related Work

Text-to-Image Generation. Recent text-to-image models based on diffusion architectures have achieved remarkable progress in high-resolution synthesis and prompt following Nichol et al. (2021); Ramesh et al. (2022); Saharia et al. (2022); Podell et al. (2023); Rombach et al. (2022). Recent foundation models further improve complex instruction following and visual text rendering Wu et al. (2025a); Z-Image Team et al. (2025). Meanwhile, controllable generation methods introduce additional spatial or semantic constraints through layouts, regions, masks, and attention manipulation Gafni et al. (2022); Li et al. (2023); Bar-Tal et al. (2023); Zhang et al. (2023); Chefer et al. (2023), while text rendering methods improve the generation of visual text Ma et al. (2023); Chen et al. (2023a;b); Tuo et al. (2023). However, these methods mainly focus on image-level alignment or local control, and do not explicitly model hierarchical dependencies among global layouts, local events, and cross-region relations required by structured visual generation.

Visual Storytelling and Comics. Visual storytelling studies the generation of coherent image sequences from textual narratives. Early works such as StoryGAN and StoryDALL-E explore sequential story visualization Li et al. (2018); Maharana et al. (2022), while later diffusion-based approaches improve consistency through memory mechanisms, layout conditioning, and language-model reasoning Rahman et al. (2022); Gong et al. (2023); Zhou et al. (2024); Tao et al. (2024);

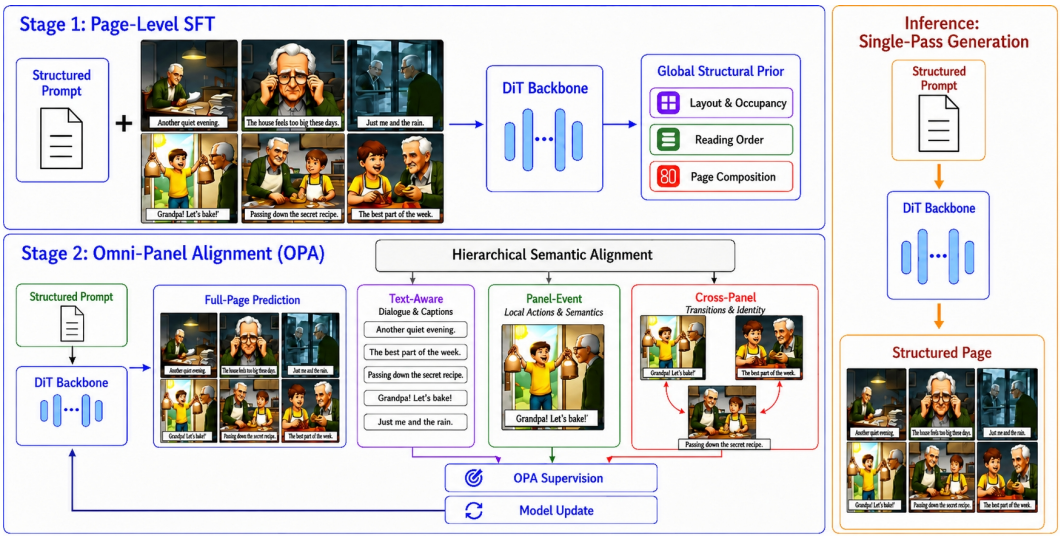


Figure 1: Overview of Panel2Page. Stage 1 performs Page-Level SFT to learn a global structural prior from structured prompt–page pairs, including layout organization, reading order, and global composition. Stage 2 introduces Omni-Panel Alignment (OPA), where the shared model generates detached localized predictions as online consistency targets for matched regions of the full-page prediction with hierarchical semantic views, including text-aware, panel-event, and cross-panel views. The resulting OPA-trained model enables single-pass complete comic page generation from a structured prompt.

Shen & Elhoseiny (2025); Park et al. (2025). Comic generation further requires page composition, panel ordering, character consistency, and text placement. Existing approaches study comic panels, manga pages, or story-to-manga pipelines Proven-Bessel et al. (2021); Wu et al. (2025b); Chen et al. (2024); Wang et al. (2026), but typically rely on separate panel generation, explicit layout control, or multi-stage composition. In contrast, Panel2Page directly generates a complete comic page from a unified structured prompt while jointly modeling page structure and hierarchical semantic relations.

Generation Evaluation. Existing text-to-image evaluation benchmarks mainly focus on compositional correctness, object binding, and image-text alignment, such as T2I-CompBench, GenEval, and DPG-Bench Huang et al. (2023); Ghosh et al. (2023); Hu et al. (2024). MLLM-based metrics further provide interpretable evaluation of conditional image quality and semantic consistency Ku et al. (2023); Lim et al. (2024). However, these protocols do not jointly evaluate page composition, panel semantics, narrative coherence, and text rendering in structured visual generation. We therefore introduce a page-oriented evaluation protocol covering text fidelity, semantic alignment, event coherence, and page quality.

3 METHOD

Panel2Page aims to generate a complete comic page from a structured prompt, where global page organization, local panel semantics, and cross-panel consistency must be jointly modeled. Given a structured prompt specifying the page layout, ordered panel events, recurring characters, textual elements, and cross-panel relations, Panel2Page directly synthesizes the entire page on a unified canvas. Unlike panel-wise pipelines that independently generate and compose individual panels, our approach learns hierarchical semantic correspondence across multiple granularities within a single generation process.

We build Panel2Page upon a pretrained text-to-image diffusion backbone Wu et al. (2025a). Specifically, a structured prompt is represented as

$$c = \left(\ell, \{p_i\}_{i=1}^{N}, \mathcal{R}\right), \tag{1}$$

where ℓ specifies the target page layout, p_i describes the visual content and textual elements of panel i, and $\mathcal{R}$ denotes cross-panel relations, including reading order, event transitions, and recurring-entity consistency.

As illustrated in Figure 1, Panel2Page adopts a two-stage training framework to progressively learn structured page generation. The first stage, **Page-Level Supervised Fine-Tuning (SFT)**, learns a global page-generation prior from complete prompt–page pairs, including page layout organization, reading order, and overall composition. The second stage, **Omni-Panel Alignment (OPA)**, further introduces training-time semantic alignment across different granularities by constructing text-aware, panel-event, and cross-panel views. These localized views provide complementary supervision for the full-page prediction while preserving a unified generation process. During inference, Panel2Page performs single-pass full-page generation without panel-wise synthesis or post-hoc composition.

3.1 Page-Level Supervised Fine-Tuning

Given a structured prompt c and its target page x, let $z_0 = E(x)$ be the encoded page latent, z_t its noisy state at timestep t, and u_t the corresponding diffusion training target. The full-page prediction is

$$s = f_\theta(z_t, t, c), \tag{2}$$

where f_θ denotes the denoising network. Page-Level SFT optimizes the standard diffusion objective:

$$\mathcal{L}_{\text{page}} = \mathbb{E}_{x,c,t,\epsilon}\left[\|f_\theta(z_t, t, c) - u_t\|_2^2\right]. \tag{3}$$

Because each target is a complete page rather than an independently generated panel, this stage adapts the model to page-level regularities, including layout organization, panel placement, reading order, recurring-entity placement, and global composition. It also provides the structural initialization for the subsequent alignment stage.

3.2 Omni-Panel Alignment

The page-level objective jointly processes all semantic units in a structured prompt, but does not explicitly isolate the correspondence between each unit and its associated visual region. OPA complements page-level denoising with localized semantic-view predictions. It applies the shared denoiser to a localized noisy latent and its view-specific condition, and encourages the corresponding region of the full-page prediction to remain consistent with the resulting localized prediction.

Localized semantic views. A localized semantic view is represented as

$$v = (c_v, \Omega_v), \tag{4}$$

where c_v is its semantic condition and Ω_v denotes its spatial support in the complete page. Let P_v be the corresponding deterministic region-extraction operator. We define

$$z_{t,v} = P_v z_t, \qquad Y_v = P_v u_t, \qquad s_v = P_v s, \tag{5}$$

where Y_v is introduced for the theoretical interpretation in Appendix C and is not used as an additional training target.

OPA constructs three complementary types of views. **Text-aware views** associate dialogue or captions with OCR-localized regions in the training page Du et al. (2020). **Panel-event views** associate individual panel regions with descriptions of their characters, actions, objects, events, and scenes. **Cross-panel views** associate selected panel pairs with relational descriptions, including reading order, event transitions, narrative progression, and recurring-entity relations. The full-page and localized branches use the same diffusion timestep and underlying noisy page latent. Localized inputs are cropped from z_t, rather than independently noising clean crops, so their predictions remain comparable. Image-space boxes are mapped to latent coordinates after the dataset transformation. Crop boundaries are expanded to align with the backbone's 2×2 latent-patch structure, with each spatial dimension at least two latent cells wide. The full-page prediction crop and localized branch therefore share exactly the same canonical latent support. Further view-extraction details are provided in Appendix A.

Detached localized alignment. For each view, the shared denoiser produces a localized prediction

$$\widetilde{q}_v = f_\theta(z_{t,v}, t, c_v). \tag{6}$$

Because c_v focuses on a particular text region, panel, or panel relation, this forward pass exposes the prediction conditioned on the selected semantic unit with reduced interference from unrelated prompt components.

We use its detached value as an online consistency target:

$$q_v = \mathrm{sg}\left[\widetilde{q}_v\right], \tag{7}$$

where $\mathrm{sg}[\cdot]$ denotes stop-gradient. The corresponding region of the full-page prediction is aligned with this target through

$$D_v = \frac{1}{|\Omega_v|} \left\| s_v - q_v \right\|_2^2. \tag{8}$$

The normalization prevents larger regions from dominating the objective. During backpropagation through Eq. (8), gradients pass through s_v but not through q_v. Thus, OPA acts as an asymmetric online consistency regularizer rather than distillation from a separately trained teacher, and introduces no additional inference parameters.

Conditional projection interpretation. Under squared loss, the optimal regional predictor given the localized information $G_v = (z_{t,v}, t, c_v)$ is the conditional projection

$$m_v(G_v) = \mathbb{E}[Y_v \mid G_v]. \tag{9}$$

The localized prediction $\widetilde{q}_v$ approximates this projection. Let $\epsilon_v = \mathbb{E}[\|\widetilde{q}_v - m_v\|_2^2/|\Omega_v|]$ denote its approximation error and $\mathcal{D}_v = \mathbb{E}[D_v]$ the expected alignment discrepancy. Since q_v and $\widetilde{q}_v$ have identical forward values, the squared triangle inequality gives

$$\mathbb{E}\left[\frac{\|s_v - m_v\|_2^2}{|\Omega_v|}\right] \le 2\mathcal{D}_v + 2\epsilon_v. \tag{10}$$

The alignment objective directly reduces $\mathcal{D}_v$, while the quality of the localized reference is captured by ϵ_v. OPA therefore encourages agreement with a focused conditional estimate, while the continuing page-level objective preserves global context and composition. The full derivation and regional denoising-risk characterization are provided in Appendix C.

Training objective. Let $\mathcal{V}_{\text{text}}$, $\mathcal{V}_{\text{panel}}$, and $\mathcal{V}_{\text{cross}}$ denote the three sets of semantic views. For granularity g, its alignment loss is

$$D_g = \frac{1}{|\mathcal{V}_g|} \sum_{v \in \mathcal{V}_g} D_v, \qquad g \in \{\text{text}, \text{panel}, \text{cross}\}. \tag{11}$$

OPA equally averages the three granularities:

$$\mathcal{L}_{\text{OPA}} = \frac{1}{3}\left(D_{\text{text}} + D_{\text{panel}} + D_{\text{cross}}\right). \tag{12}$$

If a view type is unavailable for a page, we average over the available granularities only.

Stage 2 is initialized from the Page-Level SFT checkpoint and optimizes

$$\mathcal{L}_{\text{Stage2}} = \mathcal{L}_{\text{page}} + \lambda_{\text{OPA}} \mathcal{L}_{\text{OPA}}, \tag{13}$$

where λ_{OPA} controls the strength of OPA. The page-level objective preserves complete-prompt conditioning and global composition, while OPA regularizes localized correspondence at text, panel-event, and cross-panel granularities. After training, all localized forward passes are discarded, and the final model generates complete pages using the original full-page sampling process.

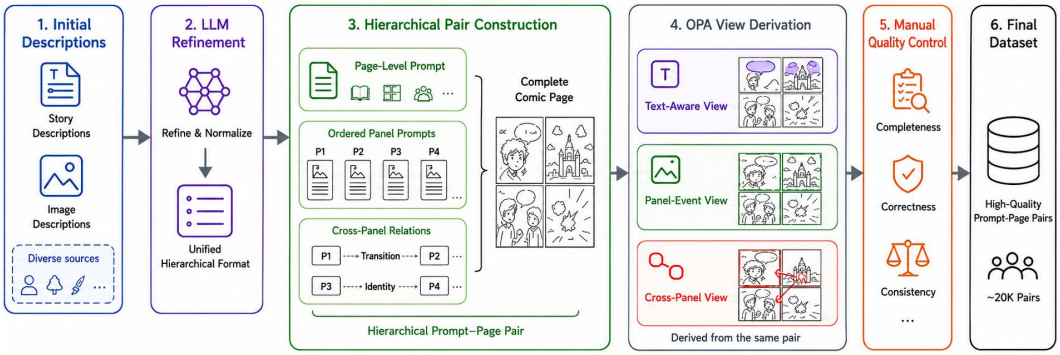


Figure 2: Overview of the hierarchical prompt–page data construction pipeline. Human-designed templates and LLM refinement produce structured prompts, from which Gemini 2.5 Flash Image synthesizes 100,000 candidate comic pages. After manual quality filtering, about 20,000 prompt–page pairs are retained to construct hierarchical representations and OPA supervision views.

4 Data Construction

4.1 Hierarchical Prompt–Page Pair Construction

To enable hierarchical semantic supervision for structured visual generation, we construct a prompt–page dataset containing approximately 20,000 high-quality training samples. Different from conventional image–text datasets, our dataset preserves semantic correspondence across three hierarchical levels: global page organization, local panel semantics, and cross-panel relations. As illustrated in Figure 2, we first design human-authored comic templates covering diverse narratives, characters, and visual scenarios. LLMs then refine these templates into structured prompts containing page layouts, ordered panel descriptions, textual elements, and recurring-entity constraints. Based on these prompts, Gemini 2.5 Flash Image generates complete comic pages, producing approximately 100,000 candidate prompt–page pairs through text-only conditioning without reference images or iterative editing Google (2025). After manual quality verification, each retained sample is paired with hierarchical annotations derived from the structured prompt. Page-level annotations describe global composition and reading order, panel-level annotations specify events, characters, actions, objects, and scenes, and cross-panel annotations describe relations between panels, including event transitions, narrative progression, and recurring entities. These annotations provide global structural supervision for Page-Level Supervised Fine-Tuning and localized semantic conditions for Omni-Panel Alignment.

4.2 Supervision View Construction for Omni-Panel Alignment

OPA does not require additional manual annotations. Instead, it derives three complementary supervision views from each hierarchical prompt–page pair: text-aware, panel-event, and cross-panel views. The text-aware view associates textual elements with OCR-localized regions to provide fine-grained text grounding. The panel-event view links each panel region with its corresponding semantic description, including characters, actions, objects, and events. The cross-panel view models relational dependencies between panel pairs, including event transitions, narrative progression, and recurring-entity consistency. These views provide complementary supervision across different semantic granularities while remaining consistent with the original full-page generation objective. View construction uses the existing structured prompts, layout regions, and OCR detections without additional manual view annotation; this is separate from manual page-quality verification.

4.3 Manual Quality Verification

To improve the reliability of visual supervision, we manually verify the generated prompt–page pairs according to semantic correspondence, event coherence, and page quality. Specifically, we check whether generated panels match their assigned descriptions, whether the complete page preserves narrative structure and cross-panel consistency, and whether the overall layout and visual

Table 1: Quantitative results on the 320-prompt benchmark. Panel2Page has the highest Text, Event, and Overall scores; Gemini 3.1 Flash Image Preview leads Semantic and Page Quality.

Method	Text ↑	Semantic ↑	Event ↑	Page Quality ↑	Overall ↑
DiffSensei Wu et al. (2025b)	0.6380	1.1051	1.1671	1.7615	1.1679
StoryDiffusion Zhou et al. (2024)	0.7390	1.3110	1.3959	1.9103	1.3391
Story2Board Dinkevich et al. (2025)	1.6080	2.1645	2.6076	2.0063	2.0966
Qwen-Image 2509 Base Wu et al. (2025a)	2.5760	2.7425	3.4532	2.7320	2.8759
Z-Image Z-Image Team et al. (2025)	3.2485	3.0653	3.7668	3.0485	3.2823
Qwen-Image 2512 Base Wu et al. (2025a)	3.6020	3.5276	4.0546	3.5150	3.6748
Gemini 2.5 Flash Image Google (2025)	3.5919	3.6326	4.1774	3.4863	3.7221
Gemini 3.1 Flash Image Preview Google (2026a)	3.7594	**3.8755**	4.1912	**3.9396**	3.9414
Panel2Page (ours)	**3.8275**	3.8384	**4.5028**	3.7633	**3.9830**



Figure 3: Qualitative comparison with strong direct full-page generators under the same structured prompts and target layouts. Panel2Page better preserves panel-event correspondence, reading order, textual elements, and recurring identities across panels.

quality satisfy the generation requirements. Candidates violating these criteria are discarded. After verification, approximately 20,000 hierarchical prompt–page pairs are retained for Page-Level Supervised Fine-Tuning and Omni-Panel Alignment, providing reliable supervision across page-level, panel-level, and cross-panel semantic granularities.

5 Experiments

5.1 Experimental Setup

5.1.1 Implementation Details

We build Panel2Page on the pretrained Qwen-Image 2512 backbone Wu et al. (2025a) and perform parameter-efficient adaptation with LoRA Hu et al. (2022) (rank 32). Following our two-stage framework, Page-Level SFT is first optimized with a learning rate of 1×10^{-5}, and Stage 2 is initialized from the SFT checkpoint and further optimized with the combined denoising and OPA objective using a learning rate of 1×10^{-6}. The OPA weight is set to 1, with three alignment terms equally weighted. During inference, we use 30 denoising steps with a classifier-free guidance scale of 4.0. We compare Panel2Page with direct full-page generators, including Qwen-Image 2509 Base, Qwen-Image 2512 Base Wu et al. (2025a), Z-Image Z-Image Team et al. (2025), Gemini 2.5 Flash Image Google (2025), and Gemini 3.1 Flash Image Preview Google (2026a), as well as panel-wise narrative generation systems, including DiffSensei Wu et al. (2025b), StoryDiffusion Zhou et al. (2024), and Story2Board Dinkevich et al. (2025). All methods are evaluated on a balanced benchmark of 320 prompts covering four layouts and four narrative categories.

Table 2: Ablation study on the combined 320-prompt benchmark. OPA from Base applies all three alignment views directly to the pretrained backbone. Other OPA variants are initialized from the Page-Level SFT checkpoint to evaluate the effect of hierarchical semantic alignment after structural adaptation.

Variant	Text ↑	Semantic ↑	Event ↑	Page Quality ↑	Overall ↑
Qwen-Image 2512 Base Wu et al. (2025a)	3.6020	3.5276	4.0546	3.5150	3.6748
OPA from Base	3.6231	3.5623	4.2450	3.5669	3.7493
Page-Level SFT	3.6500	3.6019	4.3085	3.6035	3.7910
Page-Level SFT + Text-aware	3.8247	3.6241	4.3186	3.6215	3.8472
Page-Level SFT + Panel-event	3.7243	3.7816	4.3479	3.6768	3.8827
Page-Level SFT + Cross-panel	3.7048	3.6572	4.4513	3.6619	3.8688
Page-Level SFT + Full OPA	**3.8275**	**3.8384**	**4.5028**	**3.7633**	**3.9830**



Figure 4: Qualitative ablation results under the same structured prompt. Page-Level SFT improves page organization, while Full OPA further enhances semantic consistency across panels.

5.1.2 Evaluation Protocol

We evaluate generated pages from four dimensions: Text, Semantic, Event, and Page Quality. Text measures the fidelity of rendered dialogue and captions using PaddleOCR Du et al. (2020) based on character-level edit similarity. Semantic evaluates whether each generated panel corresponds to its assigned description, including characters, actions, objects, and scenes. Event measures page-level narrative consistency, including reading order, temporal progression, causal relations, and recurring entity consistency across panels. Page Quality assesses overall composition, visual coherence, and generation quality. Semantic, Event, and Page Quality are evaluated by Gemini 3.1 Pro Preview Google (2026b) using standardized five-point prompts. The evaluator receives only the structured prompt and generated page without method identities, and identical instructions are applied to all methods. Overall is computed as the arithmetic mean of the four metrics.

5.2 Comparison with Existing Methods

Quantitative Results. Table 1 summarizes the comparison with existing structured visual generation approaches. Panel2Page achieves the best Overall performance, with leading results in Text and Event consistency while remaining competitive in Semantic and Page Quality. Notably, Panel2Page surpasses Gemini 2.5 Flash Image Google (2025), which is used to construct our training data, demonstrating that the proposed adaptation effectively improves structured page generation beyond the data-generation model. These results validate the effectiveness of combining page-level structural learning with multi-granularity semantic alignment. Unlike panel-wise generation approaches requiring post-composition, Panel2Page directly generates a complete page in a unified canvas while modeling cross-panel dependencies.

Qualitative Results. Figure 3 presents qualitative comparisons under identical structured prompts and target layouts. While existing generators can produce visually plausible pages, they may fail to preserve requested events, panel-level correspondence, or recurring character identities. In contrast, Panel2Page better maintains layout organization, panel-event alignment, textual elements, and cross-

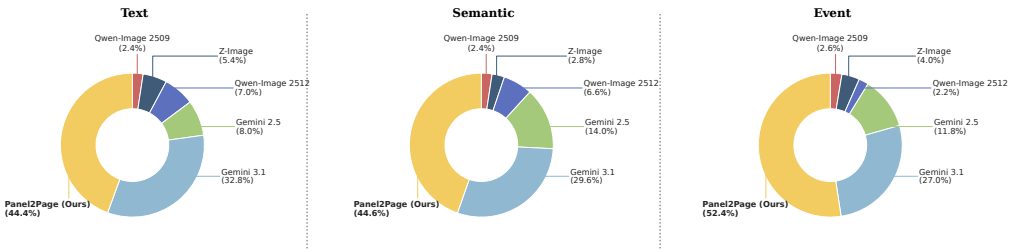


Figure 5: User study results.

panel consistency through hierarchical semantic alignment. Additional comparisons with panel-wise narrative generation systems are provided in the supplementary material.

5.3 Ablation Study and Analysis

Table 2 investigates the contributions of Page-Level SFT and different OPA granularities. Applying OPA on the pretrained backbone improves the Overall score from 3.6748 to 3.7493, indicating that localized semantic supervision is beneficial even without page-level adaptation. Page-Level SFT further increases the performance to 3.7910 by learning global page organization and narrative structure. Starting from this structural initialization, Full OPA further improves the performance to 3.9830, suggesting that hierarchical semantic alignment provides additional benefits beyond page-level structural adaptation. Different alignment views provide complementary advantages: text-aware alignment improves textual-visual correspondence, panel-event alignment enhances local event grounding, and cross-panel alignment improves relation-aware consistency across panels. Figure 4 further shows that Page-Level SFT improves global composition, while Full OPA enhances panel-level coherence and cross-panel consistency.

5.4 User Study

To further assess the perceptual quality of the generated comic pages, we conducted a user study with 25 participants. For each randomly sampled test case, participants were shown the results generated by Panel2Page and five baseline methods in randomized order, with the method identities hidden. They were asked to select the preferred result independently in terms of text-level fidelity, semantic coherence, and event-level narrative consistency. Figure 5 reports the percentage of preference votes received by each method. Panel2Page achieves the highest preference rates across all three dimensions, obtaining 44.4% for Text, 44.6% for Semantic, and 52.4% for Event. In particular, its advantage is most pronounced at the event level, indicating improved narrative consistency across panels. These human preferences align with the automatic evaluation on Text and Event, where Panel2Page also obtains the highest scores. In particular, the strongest human preference occurs for Event, reinforcing the improvement in cross-panel narrative consistency observed in the automatic metrics. Together, these results support the effectiveness of our hierarchical semantic alignment strategy in improving structured correspondence.

6 Conclusion

We present **Panel2Page**, a framework for structured visual generation studied through text-to-comic page generation. Given a structured prompt specifying layouts, ordered panel events, recurring characters, and textual elements, Panel2Page directly generates a complete page within a unified canvas. Our two-stage framework first learns a global structural prior through Page-Level SFT and then introduces **Omni-Panel Alignment (OPA)** to provide semantic-unit-aware hierarchical alignment across text-aware elements, panel events, and cross-panel relations while preserving the original single-pass generation process. On a balanced 320-prompt benchmark covering diverse layouts and narratives, Panel2Page achieves the highest Overall score and demonstrates stronger structured correspondence compared with existing approaches. These results show that global structural learning and hierarchical semantic alignment provide complementary supervision for addressing the multi-level constraints in structured visual generation.

AI USE STATEMENT

Generative AI was used to refine structured prompts and generate synthetic training pages, and a vision–language model was used for automatic evaluation, as described in the data and evaluation sections. An AI assistant also helped revise the manuscript's wording, organization, and interpretation of reported results, and correct LaTeX formatting. No additional experiments or measurements were generated by this assistance. The authors are responsible for the accuracy and integrity of the final manuscript.

ETHICS STATEMENT

Training pages are synthetically generated and manually screened. The method can also generate misleading visual narratives. The human preference study is described in the experimental section; its results concern the sampled participants and cases rather than universal preferences.

REPRODUCIBILITY STATEMENT

The main text specifies the OPA objective and evaluation protocol. Appendix A reports training and inference settings, region construction, and the training procedure; Appendix B provides an annotated data example and filtering criteria; Appendix C gives the mathematical interpretation and proof; and Appendices D and E describe evaluation details, computational cost, and data verification.

REFERENCES

Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. MultiDiffusion: Fusing diffusion paths for controlled image generation, 2023.

Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models, 2023.

Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. TextDiffuser: Diffusion models as text painters, 2023a.

Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. TextDiffuser-2: Unleashing the power of language models for text rendering, 2023b.

Siyu Chen, Dengjie Li, Zenghao Bao, Yao Zhou, Lingfeng Tan, Yujie Zhong, and Zheng Zhao. Manga generation via layout-controllable diffusion, 2024.

David Dinkevich, Matan Levy, Omri Avrahami, Dvir Samuel, and Dani Lischinski. Story2Board: A training-free approach for expressive storyboard generation, 2025.

Shiyan Du, Conghan Yue, Xinyu Cheng, and Dongyu Zhang. DEIG: Detail-enhanced instance generation with fine-grained semantic control, 2026.

Yuning Du, Chenxia Li, Ruoyu Guo, Xiaoting Yin, Weiwei Liu, Jun Zhou, Yifan Bai, Zilin Yu, Yehua Yang, Qingqing Dang, and Haoshuang Wang. PP-OCR: A practical ultra lightweight OCR system. *arXiv preprint arXiv:2009.09941*, 2020.

Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors, 2022.

Dhruba Ghosh, Hanna Hajishirzi, and Ludwig Schmidt. GenEval: An object-focused framework for evaluating text-to-image alignment, 2023.

Yuan Gong, Youxin Pang, Xiaodong Cun, Menghan Xia, Yingqing He, Haoxin Chen, Longyue Wang, Yong Zhang, et al. TaleCrafter: Interactive story visualization with multiple characters, 2023.

Google. Gemini 2.5 flash and gemini 2.5 flash image model card. Technical report, Google DeepMind, 2025. URL https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-2-5-Flash-Model-Card.pdf. Accessed: 2026-07-23.

Google. Gemini 3.1 flash image preview. https://ai.google.dev/gemini-api/docs/changelog, 2026a. Model ID: gemini-3.1-flash-image-preview.

Google. Gemini 3.1 pro model card. https://deepmind.google/models/model-cards/gemini-3-1-pro, February 2026b. Accessed: 2026-07-10.

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

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. ELLA: Equip diffusion models with LLM for enhanced semantic alignment, 2024.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747, 2023.

Max Ku, Dongfu Jiang, Cong Wei, Xiang Yue, and Wenhu Chen. VIEScore: Towards explainable metrics for conditional image synthesis evaluation, 2023.

Yitong Li, Zhe Gan, Yelong Shen, Jingjing Liu, Yu Cheng, Yuexin Wu, Lawrence Carin, David Carlson, et al. StoryGAN: A sequential conditional GAN for story visualization, 2018.

Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. GLIGEN: Open-set grounded text-to-image generation, 2023.

Youngsun Lim, Hojun Choi, and Hyunjung Shim. Evaluating image hallucination in text-to-image generation with question-answering, 2024.

Jian Ma, Mingjun Zhao, Chen Chen, Ruichen Wang, Di Niu, Haonan Lu, and Xiaodong Lin. GlyphDraw: Seamlessly rendering text with intricate spatial structures in text-to-image generation, 2023.

Adyasha Maharana, Darryl Hannan, and Mohit Bansal. StoryDALL-E: Adapting pretrained text-to-image transformers for story continuation, 2022.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models, 2021.

Jihun Park, Kyoungmin Lee, Jongmin Gim, Hyeonseo Jo, Minseok Oh, Wonhyeok Choi, Kyumin Hwang, Jaeyeul Kim, Minwoo Choi, and Sunghoon Im. Infinite-Story: A training-free consistent text-to-image generation, 2025.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis, 2023.

Ben Proven-Bessel, Zilong Zhao, and Lydia Chen. ComicGAN: Text-to-comic generative adversarial network, 2021.

Tanzila Rahman, Hsin-Ying Lee, Jian Ren, Sergey Tulyakov, Shweta Mahajan, and Leonid Sigal. Make-a-story: Visual memory conditioned consistent story generation, 2022.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents, 2022.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, et al. Photorealistic text-to-image diffusion models with deep language understanding, 2022.

Xiaoqian Shen and Mohamed Elhoseiny. StoryGPT-V: Large language models as consistent story visualizers, 2025.

Ming Tao, Bing-Kun Bao, Hao Tang, Yaowei Wang, and Changsheng Xu. StoryImager: A unified and efficient framework for coherent story visualization and completion, 2024.

Yuxiang Tuo, Wangmeng Xiang, Jun-Yan He, Yifeng Geng, and Xuansong Xie. AnyText: Multilingual visual text generation and editing, 2023.

Muyao Wang, Zeke Xie, Yanhao Chen, Lixin Xiu, and Hideki Nakayama. MangaFlow: An end-to-end agentic framework for controllable story to manga generation, 2026.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-Image technical report, 2025a.

Jianzong Wu, Chao Tang, Jingbo Wang, Yanhong Zeng, Xiangtai Li, and Yunhai Tong. DiffSensei: Bridging multi-modal LLMs and diffusion models for customized manga generation, 2025b.

Z-Image Team, Huanqia Cai, Sihan Cao, Ruoyi Du, Peng Gao, Aiming Hao, Steven Hoi, Zhaohui Hou, Shijie Huang, Dengyang Jiang, Yuming Jiang, Xin Jin, et al. Z-Image: An efficient image generation foundation model with single-stream diffusion transformer, 2025.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023.

Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. StoryDiffusion: Consistent self-attention for long-range image and video generation, 2024.

A Additional Implementation Details

A.1 Training Configuration

Panel2Page is built on the pretrained Qwen-Image 2512 backbone Wu et al. (2025a) and adapted using LoRA Hu et al. (2022) with rank 32. Page-Level SFT is trained for one epoch with a learning rate of 1×10^{-5}. Stage 2 is initialized from the Page-Level SFT checkpoint and trained for one epoch with a learning rate of 1×10^{-6} and a weight decay of 0.01. We train all models on eight NVIDIA H100 GPUs with bfloat16 precision, four gradient-accumulation steps, and gradient checkpointing. The LoRA adapters are applied to the attention projections and the image/text MLP and modulation layers of the diffusion transformer.

Table 3: Training configuration used by the current implementation.

Configuration	Stage 1	Stage 2
Initialization	Qwen-Image 2512	Page-Level SFT
Objective	Page denoising MSE	Page MSE + OPA
Learning rate	1×10^{-5}	1×10^{-6}
LoRA rank	32	32
Gradient accumulation	4	4
Precision	bfloat16	bfloat16

A.2 Stage 2 Objective

For a full-page latent z_0, timestep t, Gaussian noise ϵ, and complete structured prompt c, we construct the noisy latent z_t and obtain the online full-page prediction

$$s_\theta = f_\theta(z_t, t, c). \tag{14}$$

The standard page-level denoising objective is retained during Stage 2:

$$\mathcal{L}_{\text{page}} = \|s_\theta - y_t\|_2^2, \tag{15}$$

where y_t is the diffusion training target used by the backbone.

For each available OPA granularity $g \in \{\text{text}, \text{panel}, \text{cross}\}$, the implementation samples a localized view v, crops the corresponding noisy latent region $z_t^{\Omega_v}$, and evaluates the same model with its localized condition c_v:

$$q_v = \text{sg}\Big[f_\theta(z_t^{\Omega_v}, t, c_v)\Big]. \tag{16}$$

The stop-gradient operator makes the localized branch a target within the current optimization step. The full-page prediction is cropped to the same latent region and aligned by the region-normalized MSE

$$D_v = \frac{1}{|\Omega_v|}\left\| s_\theta^{\Omega_v} - q_v \right\|_2^2. \tag{17}$$

If multiple views of one granularity are sampled, their losses are first averaged:

$$\overline{D}_g = \frac{1}{|\mathcal{V}_g|}\sum_{v\in\mathcal{V}_g} D_v. \tag{18}$$

The OPA loss is then the mean over the granularities available for the current sample:

$$\mathcal{L}_{\text{OPA}} = \frac{1}{|\mathcal{G}_{\text{active}}|}\sum_{g\in\mathcal{G}_{\text{active}}} \overline{D}_g. \tag{19}$$

This treatment prevents samples with more annotations or larger regions from receiving proportionally larger weight. If a page has no matched text region, the mean is taken over panel-event and cross-panel supervision only. The Stage 2 objective is

$$\mathcal{L}_{\text{stage2}} = \mathcal{L}_{\text{page}} + \mathcal{L}_{\text{OPA}}. \tag{20}$$

Thus, page-level denoising remains an explicit anchor that preserves the full-page generation capability learned in Stage 1.

A.3 Localized View Construction

Shared noise and timestep. The full-page and localized branches use the same diffusion timestep and the same underlying noisy page latent. A localized teacher input is obtained by cropping z_t, rather than independently noising a clean local crop. This keeps the online and reference predictions comparable.

Text-aware views. Quoted target strings are extracted from the structured prompt. PaddleOCR Du et al. (2020) is run on the target comic page, and detections with confidence below 0.7 are discarded. Both target strings and OCR strings are Unicode-normalized, case-folded, and stripped of whitespace. Candidate pairs are scored using normalized Levenshtein similarity and matched greedily under a one-to-one constraint. Matches below 0.6 are discarded. The detected polygon is converted to a normalized bounding box, padded by 0.5% of the image size, and mapped through the same resize and center-crop transformation used by the training dataset. The localized condition asks the shared model to render the exact matched string clearly and legibly.

Panel-event views. The target page layout determines the normalized region of each panel. The implementation supports regular 2×2, 2×3, 2×4, and 3×3 layouts. A panel view uses one panel region and its corresponding event description, including the required characters, actions, objects, and scene.

Cross-panel views. Each relation annotation identifies a pair of related panels and provides a relational condition describing the transition, recurring identity, attribute, or state continuity. Its localized region is the union of the two panel boxes. This preserves the spatial relationship between the participating panels while excluding unrelated page content.

Latent canonicalization. Image-space boxes are mapped to latent coordinates after the dataset transformation. Crop boundaries are expanded to align with the Qwen-Image 2×2 latent-patch structure, and each spatial dimension is kept at least two latent cells wide. Both the full-page student crop and localized teacher branch therefore share exactly the same canonical latent support.

A.4 Stage 2 Training Procedure

Input: page image x, complete prompt c, panel annotations, relation annotations, and matched OCR regions.
1. Encode x into z_0; sample one timestep t and one noise tensor ϵ; construct z_t.
2. Predict the complete page $s_\theta = f_\theta(z_t, t, c)$ and compute $\mathcal{L}_{\text{page}}$.
3. Sample one available view from each of the text-aware, panel-event, and cross-panel sets.
4. For every sampled view v, crop $z_t^{\Omega_v}$, run the shared model with condition c_v under no-gradient mode, and detach the result to obtain q_v.
5. Crop the online prediction to $s_\theta^{\Omega_v}$ and compute the normalized regional MSE D_v.
6. Average losses within each granularity and then across active granularities to obtain $\mathcal{L}_{\text{OPA}}$.
7. Update the LoRA parameters using $\mathcal{L}_{\text{page}} + \mathcal{L}_{\text{OPA}}$.

A.5 Inference Configuration

OPA is used only during training. OCR, localized branches, annotations, and stop-gradient references are absent at inference. A single structured prompt is passed to the updated model to generate one complete comic page in one denoising trajectory.

Table 4: Inference configuration.

Configuration	Value
Denoising steps	30
Classifier-free guidance scale	4.0
4/6/8-panel resolution	1024×1536
9-panel resolution	1536×1536

B Data Construction Details

B.1 Representative JSONL Record

The following record is adapted from an actual training sample (`idiom_001023_idiom_001023_0_5228ec37`). All Chinese text fields are translated into English and long prompt strings are abbreviated for readability. The identifiers, relation types, coordinates, OCR confidence and match scores, and filtering thresholds retain their metadata values. Coordinates are normalized to the source page. The full record stores all six localized panel prompts and all available relations.


```
{
  "story_id": "001023_0_5228ec37",
  "image": "001023_0.png",
  "prompt": "Generate a 2x3 comic about a red fox in a blue sleeping
    cap and a magical wishing seed. The fox finds the seed, plants it
    carelessly, sleeps instead of caring for it, and finally loses the
     plant.",
  "panel_prompts": [
    "The fox happily holds the glowing wishing seed.",
    "The sleepy fox carelessly drops the seed into a pit.",
    "The fox sleeps in a hammock without watering the seed.",
    "An angry plant spirit confronts the sleeping fox.",
```

```
    "The plant spirit packs a bag and runs away.",
    "The fox stares helplessly at the empty pit."
  ],
  "aligned_edges": [
    {
      "panel_ids": [0, 1],
      "relation": "next",
      "bbox_norm": [0.0, 0.0, 0.6667, 0.5],
      "prompt": "The fox goes from finding the seed to planting it
   carelessly; preserve identity and action continuity."
    },
    {
      "panel_ids": [0, 3],
      "relation": "identity",
      "bbox_norm": [0.0, 0.0, 0.3333, 1.0],
      "prompt": "Preserve the same red fox, blue sleeping cap, and
   magical-seed story state across both panels."
    }
  ],
  "text_regions": [
    {
      "target_text": "Found it!",
      "ocr_text": "Found it!",
      "panel_id": 1,
      "ocr_confidence": 0.9843,
      "match_score": 1.0,
      "bbox_xyxy_norm": [0.2304, 0.2111, 0.3271, 0.3182]
    }
  ],
  "text_region_source_size": [2816, 1536],
  "text_region_annotation": {
    "target_count": 10,
    "ocr_detection_count": 14,
    "matched_count": 4,
    "confidence_threshold": 0.7,
    "match_threshold": 0.6
  }
}
```


B.2 Automatic Validation

Automatic validation checks file existence, the complete page-level prompt, the required panel count, panel and relation descriptions, valid normalized boxes, and non-empty OCR target strings. OCR matching further removes low-confidence detections and weak target–recognition matches.

B.3 Manual Verification and Filtering

Candidate prompt–page pairs undergo manual quality verification using the structured prompt as the reference. The screening criteria cover the following aspects:

- **Layout validity.** The page should follow the requested layout and panel count; pages with missing or merged panels are excluded.
- **Panel-event correspondence.** Visual events should match their assigned, ordered panel descriptions; pages with mismatched or duplicated events are excluded.
- **Cross-panel consistency.** Recurring characters and event transitions should remain coherent across panels; pages with inconsistent recurring characters or broken narrative transitions are excluded.
- **Visual and text quality.** Pages are checked for visual quality and text readability; pages with severe artifacts or unusable text regions are excluded.

Candidates that violate these criteria are discarded. Approximately 20,000 prompt–page pairs are retained from approximately 100,000 candidates. This step verifies the generated pages; hierarchical semantic annotations are derived from the structured prompts, and text-region correspondences are obtained through OCR matching.

C Interpretation of Localized Alignment

C.1 Conditional Projection Interpretation

We interpret OPA under squared loss. Let

$$G_v = (z_{t,v}, t, c_v) \tag{21}$$

denote the information available to the localized predictor. The minimum-risk predictor of the regional diffusion target Y_v based on G_v is the conditional projection

$$m_v(G_v) = \mathbb{E}[Y_v \mid G_v] = \arg\min_g \mathbb{E}\left[\|Y_v - g(G_v)\|_2^2\right]. \tag{22}$$

The localized prediction $\widetilde{q}_v$ can be viewed as a model-based approximation to m_v. Define its approximation error and the expected OPA discrepancy as

$$\epsilon_v = \mathbb{E}\left[\frac{\|\widetilde{q}_v - m_v\|_2^2}{|\Omega_v|}\right], \qquad \mathcal{D}_v = \mathbb{E}\left[\frac{\|s_v - q_v\|_2^2}{|\Omega_v|}\right]. \tag{23}$$

Since stop-gradient changes only gradient propagation, q_v and $\widetilde{q}_v$ have the same forward value. The squared triangle inequality therefore gives

$$\mathbb{E}\left[\frac{\|s_v - m_v\|_2^2}{|\Omega_v|}\right] \le 2\mathcal{D}_v + 2\epsilon_v. \tag{24}$$

Equation (24) clarifies both the role and limitation of OPA. The alignment objective directly reduces $\mathcal{D}_v$, whereas its benefit depends on the approximation quality ϵ_v of the localized predictor. Accordingly, the analysis does not assume that every localized prediction is necessarily more accurate than the full-page prediction. Instead, OPA encourages agreement with a focused conditional estimate, while the continuing page-level objective preserves global context and composition. The following elementary bound makes the dependence on reference quality explicit.

C.2 Dependence on Local Reference Quality

Let Y_v denote the ideal denoising target restricted to a view region Ω_v, and let $m_v = \mathbb{E}[Y_v \mid G_v]$ be the conditional mean under the localized information G_v. Let $s_v = s_\theta^{\Omega_v}$ be the crop from the full-page prediction and q_v be the stop-gradient localized reference. The following elementary bound explains the role of regional alignment.

Proposition 1. *For any s_v, q_v, and m_v in the same Euclidean latent space,*

$$\mathbb{E}\big[\|s_v - m_v\|_2^2\big] \le 2\mathbb{E}\big[\|s_v - q_v\|_2^2\big] + 2\mathbb{E}\big[\|q_v - m_v\|_2^2\big]. \tag{25}$$

Proof. Write $s_v - m_v = (s_v - q_v) + (q_v - m_v)$. The inequality $\|a+b\|_2^2 \le 2\|a\|_2^2 + 2\|b\|_2^2$ gives the pointwise bound. Taking expectations proves Eq. (25). □

The first term on the right is exactly the unnormalized regional alignment objective. The second term measures the quality of the localized reference. OPA does not assume that a same-step reference is a perfect teacher; rather, localized conditioning removes unrelated prompt and page content and therefore provides a more focused target for the selected scope. Retaining $\mathcal{L}_{\text{page}}$ is essential: it anchors the model to the ground-truth diffusion target and prevents a trivial agreement-only solution from replacing full-page denoising learning. The empirical ablations in the main paper complement this bound by testing OPA from the pretrained model, Page-Level SFT alone, individual scopes, and the complete two-stage objective.

D Evaluation Details

D.1 OCR-Based Text Score

Target strings are extracted from quoted dialogue and captions. PaddleOCR is applied panel by panel. For a target string T and recognized string $\widehat{T}$, characters and punctuation are separated and compared after lower-casing and removing spaces. Let r_{char} and r_{punct} be clipped normalized edit similarities. The OCR reward is

$$R_{\text{ocr}} = 0.9r_{\text{char}} + 0.1r_{\text{punct}}, \tag{26}$$

for a successfully recognized target panel. If target text is present but PaddleOCR returns no valid string, the current evaluation protocol assigns a reward of 0 to that panel. Rewards are averaged over text-bearing panels, and the paper reports

$$S_{\text{text}} = 5R_{\text{ocr}}. \tag{27}$$

Text is evaluated only when the prompt contains target dialogue or captions; samples without target text are excluded from the Text average rather than treated as failures. As a result, both R_{ocr} and S_{text} are bounded in $[0, 1]$ and $[0, 5]$, respectively.

D.2 VLM Evaluation Prompt

Semantic, Event, and Page Quality are obtained from the standardized VLM protocol used for all methods. The evaluator receives the structured prompt, target layout, ordered panel descriptions, and generated page, but not the method identity. The exact overall-evaluation prompt is reproduced below.

```
You are evaluating a generated N-panel comic page.

Story prompt:
{story_prompt}

Target number of panels: {target_panel_count}
Target layout: {layout}

Ordered panel descriptions:
{numbered_panel_prompts}

Evaluate:
1. panel count correctness
2. layout correctness
3. visual quality
4. panel-level semantic correctness
5. event order
6. transition coherence
7. character consistency
8. text readability if text is present

Return JSON only:
{
  "panel_count_score": 1,
  "layout_score": 1,
  "visual_quality_score": 1,
  "panel_semantic_score": 1,
  "event_order_score": 1,
  "transition_score": 1,
  "character_consistency_score": 1,
  "text_readability_score": null,
  "overall_score": 1,
  "main_failure_modes": [],
  "reason": "short explanation"
```

```
}

Scores must be integers from 1 to 5, except
text_readability_score may be null if no text is present.
Be strict: do not give full scores when panels are swapped,
repeated, unreadable, or semantically vague.
```

E SCOPE, COMPUTATIONAL COST, AND REPRODUCIBILITY

E.1 LIMITATIONS

The evaluation covers regular grids and an author-constructed benchmark. VLM scores and human preferences provide different evidence, without a measured calibration between them. The following subsections discuss these boundaries and synthetic-data provenance in detail.

E.2 RELATIONSHIP TO CONSISTENCY REGULARIZATION

Page-Level SFT is standard domain-specific adaptation on complete pages. OPA uses the familiar principle of detached consistency targets; its specific contribution is the organization of these targets around matched semantic conditions and regions in a structured page. Text-aware views address rendered strings, panel-event views address localized events, and cross-panel views retain a joint region for relational constraints. All views use the same model, timestep, and underlying noisy page latent. This differs operationally from a separately trained teacher or an inference-time panel-composition pipeline, but does not make self-distillation itself a new principle. The reported component ablations examine these views within Panel2Page; they do not establish superiority over all alternative consistency objectives.

E.3 TRAINING COST VERSUS INFERENCE COST

For a sample with all three view types available, the described implementation adds three localized forward evaluations to the full-page training pass, one per granularity. The local evaluations run without gradient tracking. The full-page branch still receives gradients from both the denoising and alignment losses. Actual overhead depends on crop sizes, sequence lengths, and execution details, so a count of forward passes is not a measured GPU-time ratio. At inference the localized branches are removed: generation uses the same full-page sampling structure, with 30 denoising steps, and no OCR or region-specific passes. This inference property should not be conflated with lower training cost or measured end-to-end speedups over other systems.

E.4 EVALUATION BOUNDARIES

The 320-prompt benchmark covers four regular-grid layouts and four narrative categories. It provides controlled comparisons within this scope rather than a representative sample of all comic design. The present region construction assumes regular panel boxes; irregular, overlapping, or non-rectangular panels are not validated. The Chinese-prompt examples provide qualitative evidence only.

The same VLM judge and hidden method identities improve protocol consistency but do not eliminate judge bias. In particular, the evaluator and some image-generation baselines belong to the Gemini family; possible family-specific preferences are not quantified. The human study measures preference among six displayed methods on sampled cases, whereas the automatic protocol assigns dimension scores. Preference shares are not pairwise win rates against Gemini 3.1 and are not a calibration of VLM scores. No judge–human correlation or confidence interval is reported. The small Overall margin should therefore be interpreted alongside the larger Event difference and the dimensions where Panel2Page does not lead.



Figure 6: Qualitative comparison on an eight-panel comic page involving temporal state changes. Panel2Page captures object evolution and maintains semantic continuity across different narrative stages.

E.5 Synthetic Data and Verification Scope

The approximate counts of 100,000 candidate pages and 20,000 retained pairs imply an aggregate retention rate of about 20%. This is an approximate pipeline-level statistic, not a per-criterion rejection breakdown. Manual verification screens generated pages for usability and correspondence; structured semantic annotations are inherited from prompts and text-region matches are derived using OCR. These are distinct sources of supervision, rather than independent human annotation of every field.

The present description supplies criteria and examples rather than annotator-agreement statistics.

F Additional Qualitative Comparisons

We provide additional qualitative comparisons of Panel2Page under both in-domain and out-of-domain settings. These examples illustrate page layouts, character consistency, and narrative progression; they are qualitative cases rather than a quantitative estimate of generalization.

F.1 English Test Set Comparisons

We first present additional qualitative comparisons on the English test set with diverse comic layouts and narrative structures. As shown in Figs. 7–9, Panel2Page directly generates complete comic pages in a single-pass manner, while existing methods often suffer from inconsistent character identities, broken event transitions, inaccurate textual elements, or degraded global page organization. In contrast, Panel2Page better preserves hierarchical dependencies among page-level structures, panel-level events, and cross-panel semantic relationships.

F.2 Chinese Prompt Generalization

Figures 10 and 11 show selected generations from Chinese structured prompts, whereas training uses English prompts. They illustrate the possibility of cross-language transfer in these examples. Without a systematic multilingual benchmark, they do not establish reliable generalization across languages, narrative distributions, or page layouts.



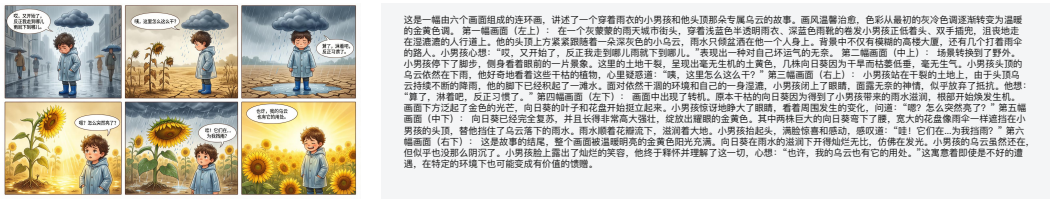
Figure 7: Qualitative comparison on a six-panel comic page generation example. Panel2Page maintains consistent character identities, coherent event progression, and aligned textual elements across the whole comic page.



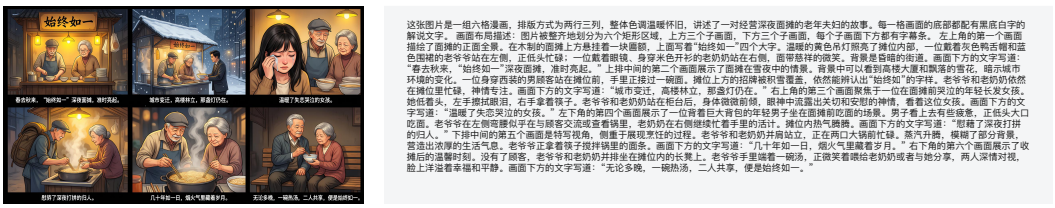
Figure 8: Qualitative comparison on a nine-panel comic page generation example. Panel2Page preserves global page organization and long-range narrative consistency under complex multi-panel layouts.



Figure 9: Qualitative comparison on a narrative-intensive comic example. Panel2Page preserves multi-character interactions, visual consistency, and story-level coherence throughout the generated page.



这是一幅由六个画面组成的连环画，讲述了一个穿着雨衣的小男孩和他头顶那朵专属乌云的故事。画风温馨治愈，色彩从最初的灰冷色调逐渐转变为温暖的金黄色调。第一幅画面（左上）：在一个灰蒙蒙的雨天城市街头，穿着浅蓝色半透明雨衣、深蓝色雨靴的卷发小男孩正低着头、双手插兜，沮丧地走在湿漉漉的人行道上。他的头顶上方紧紧跟随着一朵深灰色的小乌云，雨水只倾盆洒在他一个人身上。背景中不仅有模糊的高楼大厦，还有几个打着雨伞的路人，小男孩心想：“哎，又开始了，反正我走到哪儿雨就下到哪儿。”表现出一种对自己坏运气的无奈。第二幅画面（中上）：场景转换到了野外，小男孩停下了脚步，侧身看着眼前的一片景象。这里的土地干裂，呈现出毫无生机的土黄色，几株向日葵因为干旱而枯萎低垂，毫无生气。小男孩头顶的乌云依然在下雨，他好奇地看着这些干枯的植物，心里疑惑道：“咦，这里怎么这么干？”第三幅画面（右上）：小男孩站在干裂的土地上，由于头顶乌云持续不断的降雨，他的脚下已经积起了一滩水。面对依然干涸的环境和自己的一身湿漉，小男孩闭上了眼睛，面露无奈的神情，似乎放弃了抵抗。他想：“算了，淋着吧，反正习惯了。”第四幅画面（左下）：画面中出现了转机。原本干枯的向日葵因为得到了小男孩带来的雨水滋润，根部开始焕发生机。画面下方泛起了金色的光芒，向日葵的叶子和花盘开始挺立起来，小男孩惊讶地睁大了眼睛，看着周围发生的变化，问道：“嗯？怎么突然亮了？”第五幅画面（中下）：向日葵已经完全复苏，并且长得非常高大强壮，绽放出耀眼的金黄色。其中两株巨大的向日葵弯下了腰，宽大的花盘像雨伞一样遮挡在小男孩的头顶，替他挡住了乌云落下的雨水。雨水顺着花瓣流下，滋润着大地。小男孩抬起头，满脸惊喜和感动，感叹道：“哇！它们在...为我挡雨？”第六幅画面（右下）：这是故事的结尾。整个画面被温暖明亮的金黄色阳光充满。向日葵在雨水的滋润下开得灿烂无比，仿佛在发光。小男孩的乌云虽然还在，但似乎也没那么阴沉了。小男孩脸上露出了灿烂的笑容，他终于释怀并理解了这一切，心想：“也许，我的乌云也有它的用处。”这寓意着即使是不好的遭遇，在特定的环境下也可能变成有价值的馈赠。



这张图片是一组六格漫画，排版方式为两行三列，整体色调温暖怀旧，讲述了一对经营深夜面摊的老年夫妇的故事。每一格画面的底部都配有黑底白字的解说文字。画面布局描述：图片被整齐地划分为六个矩形区域，上方三个子画面，下方三个子画面，每个子画面下方都有字幕条。左上角的第一个画面描绘了面摊的正面全景。在木制的面摊上方悬挂着一块圆额，上面写着“始终如一”四个大字，温暖的黄色吊灯照亮了摊位内部，一位戴着灰色鸭舌帽和蓝色围裙的老爷爷站在左侧，正低头忙碌；一位戴着眼镜、身穿米色开衫的老奶奶站在右侧，面带慈祥的微笑。背景是昏暗的街道。画面下方的文字写道：“春去秋来，“始终如一”深夜面摊，准时亮起。”上排中间的第二个画面展示了面摊在雪夜中的情景。背景中可以看到高楼大厦和飘落的雪花，暗示城市环境的变化。一位身穿西装的男顾客站在摊位前，手里正接过一碗面，摊位上方的招牌被积雪覆盖，依然能辨认出“始终如”的字样。老爷爷和老奶奶依然在摊位里忙碌，神情专注。画面下方的文字写道：“城市变迁，高楼林立，那盏灯仍在。”右上角的第三个画面聚焦于一位在面摊前哭泣的年轻长发女孩，她低着头，左手擦拭眼泪，右手拿着筷子。老爷爷和老奶奶站在柜台后，身体微微前倾，眼神中流露出关切和安慰的神情，看着这位女孩。画面下方的文字写道：“温暖了失恋哭泣的女孩。”左下角的第四个画面展示了一位背着巨大背包的年轻男子坐在面摊前吃面的场景。男子看上去有些疲惫，正低头大口吃面。老爷爷在左侧弯腰似乎在与顾客交流或查看锅里。老奶奶在右侧继续忙着手里的活计。摊位内热气腾腾。画面下方的文字写道：“慰藉了深夜打拼的归人。”下排中间的第五个画面是特写视角，侧重于展现烹饪的过程。老爷爷和老奶奶并肩站立，正在两口大锅前忙碌。蒸汽升腾，模糊了部分背景，营造出浓厚的生活气息。老爷爷正拿着筷子搅拌锅里的面条。画面下方的文字写道：“几十年如一日，熄火气里藏着岁月。”右下角的第六个画面展示了收摊后的温馨时刻。没有了顾客，老爷爷和老奶奶并排坐在摊位内的长凳上。老爷爷手里端着一碗汤，正微笑着喂给老奶奶或者与她分享，两人深情对视，脸上洋溢着幸福和平静。画面下方的文字写道：“无论多晚，一碗热汤，二人共享，便是始终如一。”

Figure 10: Additional qualitative results with Chinese structured prompts as out-of-domain inputs. Panel2Page preserves the requested layouts, character consistency, textual elements, and narrative coherence despite the language shift from English to Chinese prompts.



这张图片展示了一组连环画，采用了两行三列的排版方式，共计六个长方形的子画面。画面风格为手绘插画风，色彩温暖且富有质感。故事的主角是一个戴着棕色飞行员帽子、护目镜，脖子上围着红色围巾，身穿白色衬衫、蓝色裤子和棕色靴子的小男孩。每个画面的左上角都有一段黑色的中文叙述文本。第一排左侧的第一个画面中，小男孩站在地面上，仰望着四周。他显得非常渺小，因为周围矗立着巨大的、如同摩天大楼般的古老书籍堆，这些书堆构成了高耸的墙壁。直通云霄，阳光从上方洒下。画面左上角的文字写道：“迷途的小探险家，误入古老书卷编织的迷宫。”第一排中间的第二个画面展示了攀登的场景。小男孩正悬挂在由巨大书籍堆叠而成的陡峭悬崖上，他右手握着一把登山镐，将其深深插入书脊之中作为支撑，左手向上攀爬，表情显得有些吃力。背景是深不见底的书籍峡谷和缭绕的云雾。画面左上角的文字写道：“向着未知的山顶，寻找传说中的智慧之光。”第一排右侧的第三个画面描绘了惊险的一刻。小男孩似乎失足跌落，身体悬空，正惊慌地伸出双手。此时，一只巨大的白色折纸鹤从右侧的书架缝隙中飞出，向他飞来，仿佛要接住他。背景是堆满书籍的架子。画面左上角的文字写道：“危急时刻，书缝间的纸鹤，伸出援手。”第二排左侧的第四个画面中，小男孩已经骑在了巨大的白色纸鹤背上。他张开双臂，表情兴奋且快乐，红色的围巾在风中飘扬。纸鹤载着他飞翔在空中，下方是由无数书籍堆叠而成的山峦和云海。画面左上角的文字写道：“乘着纸鹤，飞越重岩叠嶂的书海。”第二排中间的第五个画面展示了旅程的终点。纸鹤载着小男孩飞抵一座高耸的山峰顶端，山顶上悬浮着一颗巨大的、散发着耀眼金黄色光芒的圆球。小男孩坐在纸鹤上，伸出手去触碰这团光芒。画面左上角的文字写道：“在山顶之巅，终于触碰到了那传说中的光芒。”第二排右侧的第六个画面也就是最后一个画面，场景回到了室内的图书馆环境。小男孩站在一张放着摊开书籍的桌子前，双手捧着那颗发光的金色圆球，脸上洋溢着温暖的笑容。光球照亮了周围摆满书籍的深色书架。那只白色的纸鹤在右上方轻轻飞舞。画面左上角的文字写道：“将智慧之光带回，照亮了被遗忘的角落。”



这是一幅温馨的六格漫画，讲述了一个小男孩想要留住夏天的充满童趣和想象力的故事。画面分为上下两排，每排三幅图。第一幅图（左上）：画面是一个明亮的白天背景。一个小男孩正跪在草地上，他戴着一顶圆圆的草帽，穿着白T恤和蓝色的背带裤，他的左手拿着一个巨大的捕虫网，右手拿着一把刷子，正从旁边的一个黄色油漆桶里蘸取金黄色的粘稠液体（看起来像蜂蜜或胶水），涂抹在捕虫网的网格上，他的表情专注而带有一丝期待。图中的文字写着：“夏天，夏天，不要走！”。表达了他想要通过这种方式捕捉夏天的愿望。第二幅图（中上）：时间到了黄昏，天空呈现出紫色和橙色的渐变，太阳正缓缓落下山头。小男孩站在草地上，双手高高举起那个涂了粘稠液体的捕虫网，正对着即将落下的夕阳。他试图用网兜“粘”住太阳。他的表情充满希冀和努力。图中的文字是：“粘住它，拜托了！”，展示了他天真地以为这样就能阻止太阳下山。第三幅图（右上）：夜幕降临，天空变成了深蓝色，星星点点，太阳已经完全下山了。小男孩垂头丧气地站在黑暗中，那个捕虫网空荡荡地举在手里，他的表情显得非常失落和伤心，因为没能留住太阳。图中的文字写着：“哎，失败了……”，宣告了他计划的落空。第四幅图（左下）：画面背景依然是深蓝色的夜空，但突然出现了转机。无数发光的流星和光点像雨一样落下，正好落入了小男孩高举的捕虫网中。网兜瞬间被点亮，发出金色的光芒。小男孩抬起头，眼睛睁得大大的，脸上写满了惊讶和不可思议。图中的文字是：“哇！这是什么？”，表现出他的惊奇。第五幅图（中下）：原本的捕虫网此刻变成了一个发光的半球形结构，放置在草地上，就像一个小小的发光帐篷或穹顶。网格线条都在发光，周围还有许多金色的星星在飞舞和环绕。小男孩开心地在这个发光的网罩周围奔跑玩耍，脸上洋溢着快乐的笑容。图中的文字写着：“流星游乐场！”，意味着他虽然没抓到太阳，但抓住了流星，创造了一个新的乐园。第六幅图（右下）：夜深了，画面宁静而祥和。那个发光的网罩像一床温暖的被子或蚊帐一样罩在草地上。小男孩正趴在这个发光的网罩下甜美地睡着了，嘴角带着微笑。几颗明亮的星星挂在网罩上方，仿佛在守护着他。图中的文字总结道：“留住夏天了！”，寓意着虽然太阳落山了，但他通过流星的光芒和温暖，在梦中留住了他心中的夏天。

Figure 11: Additional qualitative results with Chinese structured prompts as out-of-domain inputs. Panel2Page maintains structured correspondence between prompts and generated pages, including global composition, local events, and cross-panel transitions.