

PosterCopilot: Toward Layout Reasoning and Controllable Editing for Professional Graphic Design

Jiazhe Wei^{1*}, Ken Li^{1*}, Tianyu Lao², Haofan Wang², Yueming Lyu¹,
Liang Wang^{1,3}, Caifeng Shan¹, and Chenyang Si^{1†}

¹ PRLab, Nanjing University, Suzhou, China

² LibLib.ai, China

³ Institute of Automation, Chinese Academy of Sciences, Beijing, China

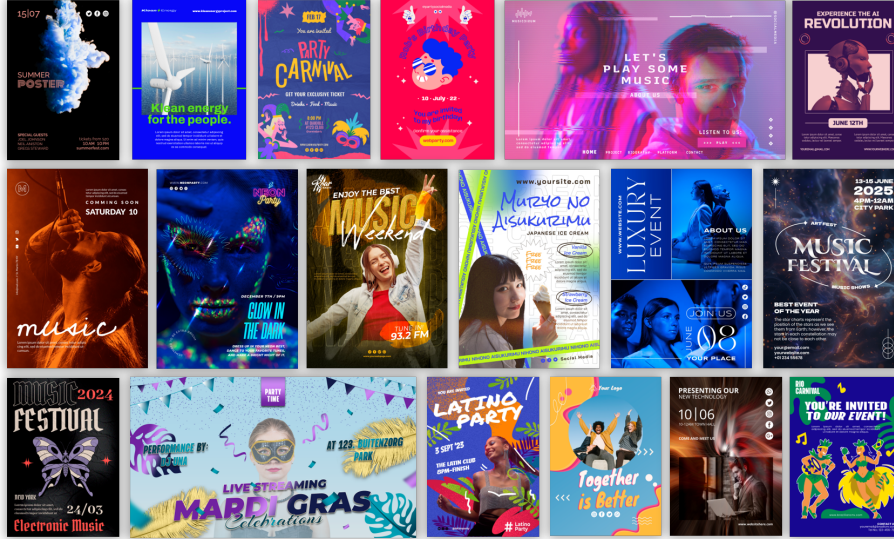


Fig. 1: Generated results from our **PosterCopilot**. PosterCopilot exhibits exceptional graphic design capabilities by creating artworks with professional-grade layout, compelling visuals, and cohesive themes.

Abstract. Graphic design forms the cornerstone of modern visual communication, serving as a vital medium for promoting cultural and commercial events. Recent advances have explored automating this process using Large Multimodal Models (LMMs), yet existing methods often produce geometrically inaccurate layouts and lack the iterative, layer-specific editing required in professional workflows. To address these limitations, we present PosterCopilot, a framework that advances layout reasoning and controllable editing for professional graphic design. Specifically, we

* Equal contribution.

† Corresponding author.

introduce a progressive three-stage training strategy that equips LMMs with geometric understanding and aesthetic reasoning for layout design, consisting of Perturbed Supervised Fine-Tuning, Reinforcement Learning for Visual-Reality Alignment, and Reinforcement Learning from Aesthetic Feedback. Furthermore, we develop a complete workflow that couples the trained LMM-based design model with generative models, enabling layer-controllable, iterative editing for precise element refinement while maintaining global visual consistency. Extensive experiments demonstrate that PosterCopilot achieves geometrically accurate and aesthetically superior layouts, offering fine-grained, layer-wise controllability for professional iterative design.

Keywords: Graphic Design · Large Multimodal Models · Reinforcement Learning

1 Introduction

Graphic design serves as a fundamental medium for visual communication [35], translating abstract ideas into clear and engaging visuals. It brings together images, text, and graphic elements in a deliberate way to create layouts that are both informative and visually appealing, bridging creativity with effective communication [2]. Recently, growing interest has emerged in automating the graphic design process through artificial intelligence. One major line of work explores diffusion-based generative models, which leverage their strong image synthesis capabilities to create visually rich posters [18, 30, 49]. However, because these models generate all image regions simultaneously, they struggle to preserve the structural integrity, texture fidelity, and stylistic consistency of user-provided assets, making local refinements prone to distortion [14, 34]. Another line leverages Large Multimodal Models (LMMs) to reason over design elements and predict their spatial and layer-wise arrangements, determining each element’s position, scale, and ordering within the composition [19, 39, 51]. These methods preserve the authenticity of visual assets and introduce interpretability and controllability into the design process, representing a promising step toward layout-centric and automation-oriented graphic design.

Despite these advances, current LMM-based methods still exhibit notable limitations when applied to professional design workflows: 1) when handling complex and numerous assets, existing methods often produce inaccurate and unaesthetic layouts [33, 50] as shown in Fig. 2. We identify that existing methods rely on supervised fine-tuning (SFT) over discrete textual tokens to represent continuous spatial coordinates, creating a mismatch between the model’s symbolic representation and the true Euclidean geometry of layout design [23, 29]. This mismatch leads to misalignment, distortion, and suboptimal compositions. Moreover, these models lack visual feedback during training, which limits their ability to perceive and reason about aesthetic layouts [39, 40]. 2) More critically, current LMM-based approaches merely generate initial drafts and lack interactive editing capabilities [25, 36]. However, professional designers refine the

drafts through multiple rounds of precise, layer-specific adjustments [3, 21, 37]. Therefore, enabling iterative refinement is a crucial requirement for advancing AI-assisted graphic design toward practical applications [24, 38].

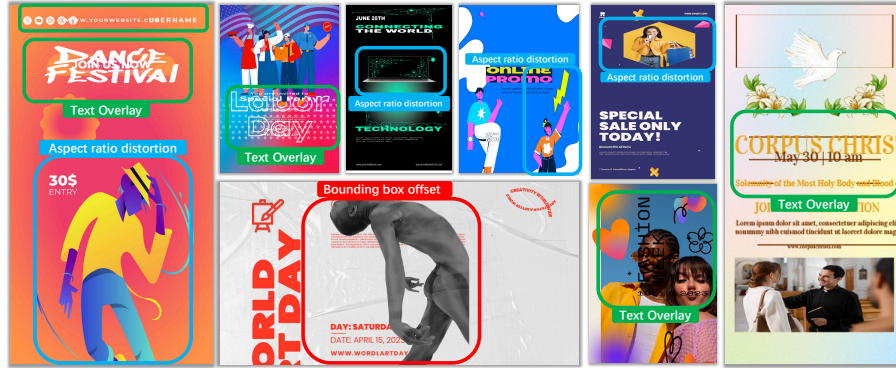


Fig. 2: Some failure cases created by existing design models in real-world, multi-asset scenarios, producing severe misalignments and visual discord.

To address these challenges, we propose **PosterCopilot**, which advances the field toward layout reasoning and controllable editing for professional graphic design. Specifically, to mitigate the inaccurate and unaesthetic layouts resulting from token-based coordinate representations, we propose Perturbed Supervised Fine-Tuning (PSFT), which reformulates coordinate regression into a distribution-based learning paradigm by introducing controlled perturbations to ground-truth coordinates. Compared to point-wise regression, learning a coordinate distribution allows the model to reason over continuous spatial relationships rather than memorizing discrete positions [27, 55], leading to more coherent and aesthetically balanced layouts. To overcome the model’s inability to perceive and reason about layout aesthetics, we introduce a two-stage reinforcement learning (RL) strategy: the first stage explicitly injects graphic design rules as verifiable rewards, while the second renders predicted layouts into poster images to obtain aesthetic feedback. In the first stage, Reinforcement Learning for Visual-Reality Alignment (RL-VRA) introduces verifiable geometric reward signals that, together with the rich layout priors acquired during PSFT, steer the model toward spatially coherent and rule-consistent layouts, helping it internalize broadly applicable layout design conventions. In the second stage, Reinforcement Learning from Aesthetic Feedback (RLAF) employs a learned aesthetic reward model to align the model with human aesthetic preferences, encouraging aesthetically coherent and diverse compositions rather than collapsing to a single canonical layout favored by purely geometric objectives.

Finally, to enable iterative and controllable refinement beyond initial poster draft, we develop a complete workflow that couples the trained LMM-based design model with the generative models, seamlessly integrating asset creation with precision editing. This workflow supports layer-specific, iterative editing,

allowing precise modification of individual elements without altering surrounding content. Experimental results indicate that the design model trained via our three-stage method produces layouts that are both accurate and visually appealing. More significantly, PosterCopilot’s integration of a generative agent provides precise layer-wise editing. This transforms it into a powerful assistant, allowing designers to take a well-composed draft as a starting point and have it further optimized for enhanced aesthetics and practical application. Our main contributions are summarized as follows:

- We introduce a progressive alignment training paradigm (PSFT $\rightarrow$ RL-VRA $\rightarrow$ RLAF) that stabilizes tokenized coordinate prediction and injects both verifiable geometric feedback and aesthetic feedback into LMM-based layout planning.
- We instantiate this paradigm in PosterCopilot for multi-layer poster layout, and deploy an inference-time generative agent to support layer-wise, multi-round editing in practical design scenarios.
- We contribute a large-scale multi-layer poster dataset with rational granularity, along with its construction pipeline, addressing critical gaps in data scarcity and layer segmentation to benefit future research and applications.

2 Related Work

Multi-layer Graphic Layout Planning prioritizes real-world practicality by first inferring layouts, then assembling layers for optimal flexibility. LMM-assisted approaches (LayoutPrompter [31], LayoutNUWA [48], PosterLlama [45], PosterLLaVA [57], PosterO [16]) employed in-context learning, while others specialized in asset integration (Graphist [10]), typography (POSTA [4]), or external generation (CreatiPoster [58], COLE [22]). Crucially, these methods mimic static datasets rather than learning from aesthetic outcomes. In a parallel line, end-to-end generative models synthesize posters holistically from user-provided assets and prompts (DreamPoster [17], PosterCraft [8], PosterOmni [7]); while some inject reward or aesthetic feedback to enhance visual quality, generating all regions jointly forgoes the explicit, layer-level controllability and asset fidelity that professional iterative editing demands. Our approach transcends these limitations by internalizing layout principles and visual aesthetics through direct generative feedback within a layer-wise, editable layout paradigm.

3 Methodology

In this section, we will first detail the training paradigm for the design model, and subsequently present the complete PosterCopilot inference pipeline. Our three-stage design model training paradigm is illustrated in Fig. 3.

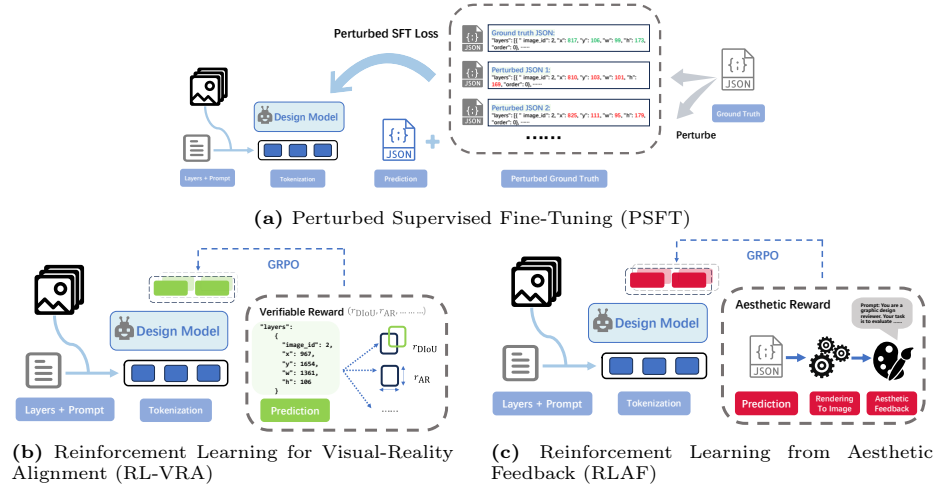


Fig. 3: Overview of the training paradigm of PosterCopilot. Instead of treating layout generation as a simple regression task, we train PosterCopilot through a three-stage paradigm that progressively develops both precise layout capabilities and human-aligned aesthetic sensibility.

3.1 Task Formulation

Our objective is to automatically arrange user-provided elements $E = \{e_1, \dots, e_N\}$ of types $\mathcal{T} = \{\text{image}, \text{text}, \text{shape}\}$ on a canvas, achieving aesthetic coherence while preserving asset fidelity. Text elements are rasterized into image layers for unified processing. The input elements, canvas dimensions (H_c, W_c) and user text prompt are encoded into a multimodal prompt P_{H_c, W_c}^E , which our design model $\mathcal{M}$ processes to generate the final layout $\mathcal{M}(P_{H_c, W_c}^E) \rightarrow \mathbf{G}$, where $\mathbf{G} = \{(\mathbf{b}_i, l_i)\}_{i=1}^N$ specifies each element’s bounding box (BBox) $\mathbf{b}_i$ and layer order l_i . Following prior work, we represent the layout $\mathbf{G}$ as structured JSON text encoding each layer’s BBox and order.

3.2 Perturbed Supervised Fine-Tuning

When LMMs represent numerical coordinates as text, each number is tokenized into sub-word units (e.g., “517” $\rightarrow$ “5”, “17”), whose embeddings bear no inherent metric structure [11, 44]. As a result, numerically adjacent coordinates such as “499” and “500” may be mapped to distant regions in embedding space due to entirely different token decompositions, while distant values sharing token prefixes appear deceptively close [20]. This broken topology turns coordinate regression via standard SFT into optimization over a fractured, non-smooth landscape, leading to systematic localization errors. A natural remedy is to soften the supervision signal: rather than forcing the model to memorize a single discrete target, we encourage it to learn a distribution over spatially neighboring coordinates, thereby smoothing out the token-boundary discontinuities.

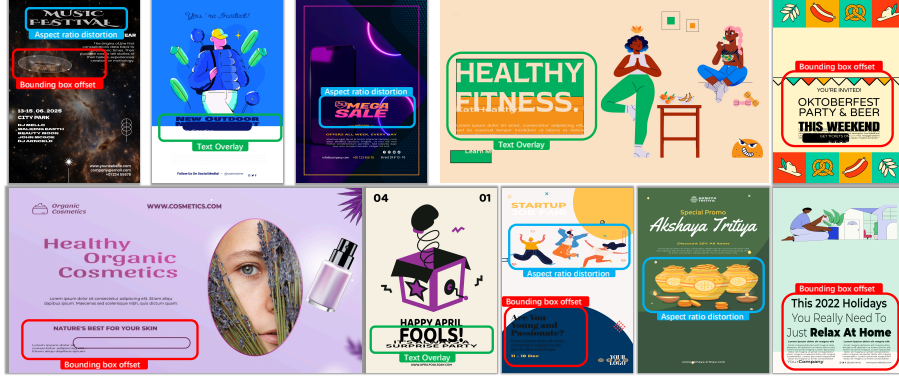


Fig. 4: Our motivation for visual-reality alignment and aesthetic feedback stems from the observation that design models frequently produce works that violate fundamental graphic design principles, as well as exhibit serious aesthetic flaws. We use red, green, and blue boxes to mark the error areas in the figure.

Based on this insight, we propose **Perturbed Supervised Fine-Tuning (PSFT)**. Standard SFT supervises the model with a single ground-truth layout $\mathbf{G}_{\text{gt}} = \{(b_i, l_i)\}_{i=1}^N$, performing point-wise regression on discrete JSON tokens. Instead, we perturb the BBox numerical values in the JSON text to construct a soft supervision neighborhood around the ground truth. Concretely, for each training sample, we generate n perturbed variants by injecting Gaussian noise to every BBox coordinate:

$$\mathbf{G}_{\text{pert}}^{(j)} \sim \mathcal{N}(\mathbf{G}_{\text{gt}}, \sigma^2 \mathbf{I}), \quad j = 1, 2, \dots, n \quad (1)$$

where σ controls the perturbation magnitude. In practice, perturbed values are rounded to integers and clamped to the valid canvas range to ensure physically meaningful coordinates. The training objective $\mathcal{L}_{\text{PSFT}}$ combines the standard cross-entropy loss on the original ground truth with an averaged loss over the n perturbed targets:

$$\mathcal{L}_{\text{PSFT}} = L_{\text{CE}}(\hat{\mathbf{G}}, \mathbf{G}_{\text{gt}}) + \lambda_{\text{pert}} \cdot \frac{1}{n} \sum_{j=1}^n L_{\text{CE}}(\hat{\mathbf{G}}, \mathbf{G}_{\text{pert}}^{(j)}) \quad (2)$$

where $\hat{\mathbf{G}}$ is the model prediction and λ_{pert} balances the two terms. By supervising the model against multiple spatially neighboring layouts rather than a single ground truth, PSFT encourages it to learn a continuous spatial distribution centered on the target, effectively alleviating the inherent discretization bottleneck of text-based layout regression.

3.3 Reinforcement Learning for Visual-Reality Alignment

While PSFT offers a robust spatial prior, its dependence on supervised learning without visual feedback results in geometric flaws, such as BBox drift and aspect ratio distortion. Critically, these layout errors, though evident in Fig. 4, cannot

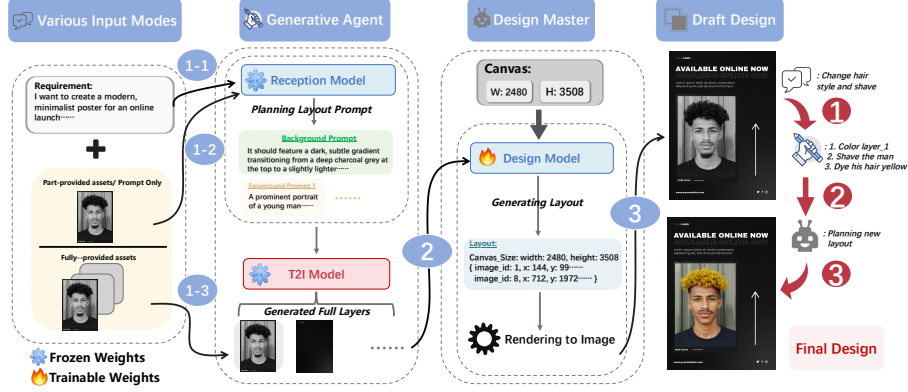


Fig. 5: Overview of PosterCopilot’s Inference and Editing Pipeline. The *standard inference* and *multi-round editing* pipelines are marked by blue and red numbers, respectively. Before layout design, PosterCopilot can supplement new assets when design materials are insufficient. **Generative Agent** first processes user requirements, undergoes professional planning, and delivers complete assets. **Design Master** then generates optimal compositions based on the assets and requirements, ultimately rendering the **Draft design**. The draft design will be revised into the **final design** after multiple rounds of editing by the collaboration of both generative agent and design master.

be captured easily within the SFT paradigm itself. To bridge this gap and align model outputs with graphic design principles, we introduce **Reinforcement Learning for Visual-Reality Alignment (RL-VRA)** phase.

We frame RL-VRA as an online policy optimization task under a single-step Markov Decision Process (MDP). The state s corresponds to the input prompt P_{H_c, W_c}^E , while the action a represents the layout generation $\mathbf{G} = \{(b_i, l_i)\}_{i=1}^N$. Our objective is to refine the pre-trained SFT policy $\pi_{\text{ref}}(\mathbf{G} | s)$ into an enhanced policy $\pi_{\theta}(\mathbf{G} | s)$ by maximizing the expected return under a geometry-aware reward signal:

$$J_{VRA}(\theta) = \mathbb{E}_{\mathbf{G} \sim \pi_{\theta}(\cdot | s)} [r(\mathbf{G})] - \beta D_{KL}(\pi_{\theta}(\cdot | s) || \pi_{\text{ref}}(\cdot | s)) \quad (3)$$

where $J_{VRA}(\theta)$ balances reward maximization against policy conservatism, with π_{ref} serving as the frozen reference policy, β controlling the KL regularization strength [43], and $r(\mathbf{G})$ providing dense **verifiable geometric visual feedback**. To ensure stable policy updates for high-dimensional discrete action spaces, we employ **Group Relative Policy Optimization (GRPO)** [46], which operates without explicit value function estimation. For each group of K policy rollouts, we compute:

$$A_i = r(\mathbf{G}_i) - \frac{1}{K} \sum_{j=1}^K r(\mathbf{G}_j), \quad r_i(\theta) = \frac{\pi_{\theta}(\mathbf{G}_i | s)}{\pi_{\theta_{\text{old}}}(\mathbf{G}_i | s)} \quad (4)$$

where A_i represents the advantage of action $\mathbf{G}_i$ relative to the group, and $r_i(\theta)$ is the probability ratio between the new and old policies. Our reward

function $r(\mathbf{G}) = r_{\text{Spatial}} + r_{\text{Element}} + r_{\text{format}}$ provides multi-scale geometric supervision, decomposing layout quality into spatial coherence and element-level fidelity components.

The spatial reward r_{Spatial} addresses layout misalignment through Distance Intersection over Union (DIOU) [60]:

$$r_{\text{Spatial}} = r_{\text{DIOU}} = \sum_i \left(\text{IoU}(\mathbf{b}_i, \mathbf{b}_i^{\text{gt}}) - \frac{\rho^2(\mathbf{b}_i, \mathbf{b}_i^{\text{gt}})}{c^2} \right) \quad (5)$$

where ρ denotes the center distance, c represents the diagonal of the minimal enclosing box, and $\mathbf{b}_i^{\text{gt}}$ is the ground-truth box from $\mathbf{G}_{\text{gt}}$.

The element-level reward $r_{\text{Element}} = r_{\text{AR}} + r_{\text{size}}$ penalizes geometric distortions that compromise visual integrity. The aspect ratio reward:

$$r_{\text{AR}} = - \sum_i \left| \log \left(\frac{w_i/h_i}{w_i^{\text{gt}}/h_i^{\text{gt}}} \right) \right| \quad (6)$$

preserves element proportions, while the size reward:

$$r_{\text{size}} = - \sum_i \left[\text{smooth}_\delta \left(\frac{w_i - w_i^{\text{gt}}}{w_i^{\text{gt}}} \right) + \text{smooth}_\delta \left(\frac{h_i - h_i^{\text{gt}}}{h_i^{\text{gt}}} \right) \right] \quad (7)$$

maintains original dimensions using the Huber loss [13]:

$$\text{smooth}_\delta(d) = \begin{cases} 0.5d^2/\delta & |d| < \delta \\ |d| - 0.5\delta & \text{otherwise} \end{cases} \quad (8)$$

where δ controls the transition between quadratic and linear regimes, preventing reward domination during extreme size distortions.

We further incorporate r_{format} to enforce JSON-structured outputs. r_{format} returns a large penalty if the model output fails JSON validation, and full score otherwise. The complete reward formulation:

$$r(\mathbf{G}) = \underbrace{r_{\text{DIOU}}}_{\text{Spatial Coherence}} + \underbrace{\lambda_{\text{size}}r_{\text{size}} + \lambda_{\text{AR}}r_{\text{AR}}}_{\text{Element Fidelity}} + r_{\text{format}} \quad (9)$$

where $\lambda_{\text{size}}, \lambda_{\text{AR}} > 0$ balance reward components. This geometrically-grounded reward structure injects explicit visual-reality constraints directly into the policy gradient updates, breaking the limitations of previous methods that lack direct design principle guidance.

3.4 Reinforcement Learning from Aesthetic Feedback

While prior stages enforce graphic design rules based on a single ground-truth, this is just one of many aesthetically valid solutions. To align with broader human aesthetic preferences, we introduce the **Reinforcement Learning from**

Aesthetic Feedback (RLAF) stage. This stage explores a wider design space using a new subjective reward, $r_{\text{aes}}(\mathbf{G})$, provided by a pre-trained LMM (acting as an aesthetic judge) that evaluates the final *rendered image*. This aesthetic score is combined with our format reward r_{format} :

$$r_{\text{RLAF}}(\mathbf{G}) = r_{\text{format}} + \lambda_{\text{aes}} r_{\text{aes}}(\mathbf{G}) \quad (10)$$

where $\lambda_{\text{aes}} > 0$. This stage encourages the model to discover novel, high-appeal layouts rather than collapsing to a single canonical layout favored by purely geometric objectives. To mitigate potential reward hacking, we retain KL regularization against the frozen reference policy in Eq. 3, enforce structural validity via r_{format} , and restrict RLAF to a single epoch on a small curated set of 1K samples, limiting the policy’s opportunity to overfit to the reward model. Full details on all reward scaling are provided in the supplementary material.

3.5 Generative Agent for Asset Completion and Iterative Layer-wise Editing

With the design model as the core, we integrate an inference-time generative agent to complete PosterCopilot in practical settings. This agent introduces no additional trainable components and supports two functions (Fig. 5): it synthesizes missing assets when only partial materials are provided, and enables fine-grained, multi-round, layer-wise editing from user instructions. For asset completion, we employ a reception model to generate a layer-specific text prompt for each missing layer by conditioning on the *entire poster context*, including the poster-level description and a structured summary of existing layers. The resulting prompt is combined with the provided assets as style references and fed into a text-to-image model to synthesize the missing element. For iterative refinement, the agent applies edits only to the user-specified target layer(s) while keeping all other layers fixed, enabling stable multi-round updates without drifting edit scopes. We use a fixed prompt template and consistent decoding settings for the reception model across experiments.

4 Application

Harnessing its powerful reasoning capability and fine-grained layer-wise architecture, PosterCopilot unlocks diverse applications in professional design scenarios. Due to the limited space, here we briefly introduce several typical applications, leaving more application presentations in the supplementary material.

Poster Generation from Insufficient Assets: As shown in Fig. 6a, PosterCopilot’s generative agent synthesizes missing layers (*e.g.*, backgrounds or foregrounds) with stylistic consistency, enabling rapid draft generation from incomplete assets.

Multi-round Fine-grained Single-layer Edit: As shown in Fig. 6b, PosterCopilot supports iterative edits on individual layers (*e.g.*, material or background

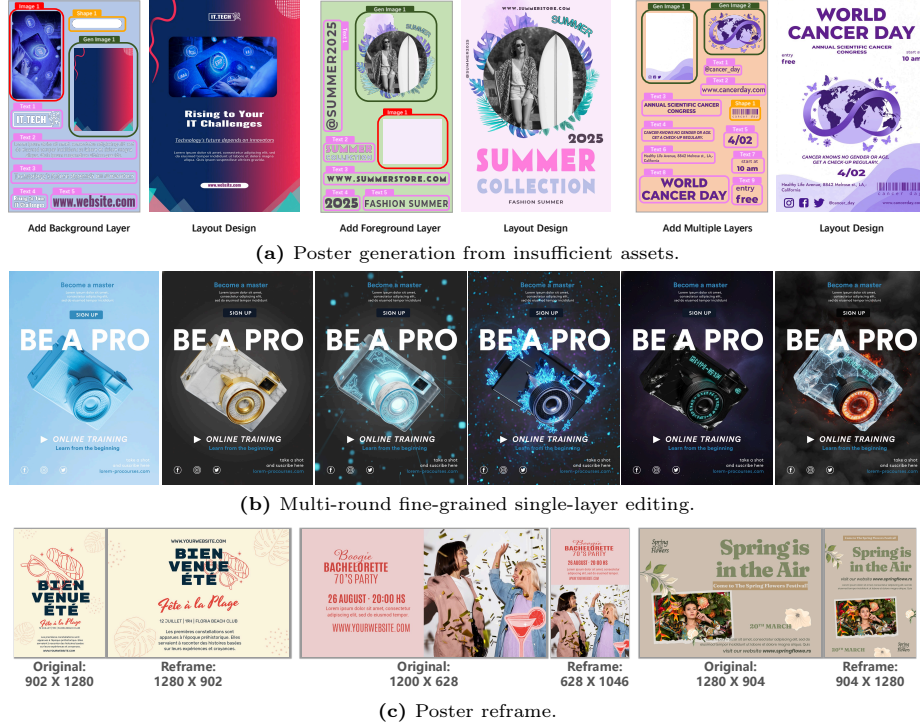


Fig. 6: Representative applications of PosterCopilot.

changes) while strictly preserving all other elements, avoiding the global distortion common in diffusion-based editing.

Poster Reframe: As shown in Fig. 6c, PosterCopilot intelligently regenerates layouts for new canvas sizes while maintaining layout harmony.

5 Experimental Details

5.1 MLP-169K Datasets

A long-standing challenge in multi-layer poster dataset construction is over-segmentation, where a single visual element is fragmented across multiple layers [61]. To address this, we develop a dedicated pipeline (Fig. 7) that leverages OCR-based fine-grained bounding boxes to merge over-segmented layers and filter redundant ones, yielding Multi-Layer-Poster-169K (**MLP-169K**): **169K** posters with **2.6M** layers (1.2M text, 1.4M image/decorative). Fig. 8 compares MLP-169K with existing datasets: CGL-V2 [28], PosterLayout [15], QB-Poster [57], WebUI [53], and Crello [56]. PosterLayout lacks explicit layer z-ordering; CGL-V2 covers only e-commerce; Crello is well-structured but limited in scale; and WebUI, primarily targeting UI design, has an average layer count far exceeding typical real-world posters. In contrast, MLP-169K offers precise

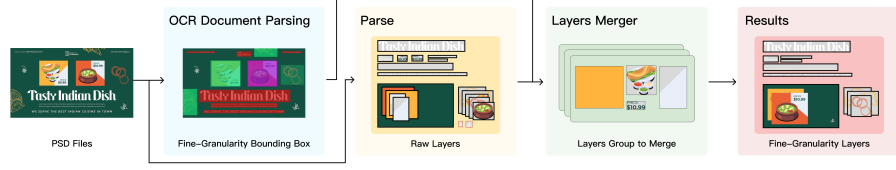


Fig. 7: Dataset construction pipeline for our PosterCopilot. We merged numerous scattered layers with OCR-based fine-granularity bounding box.

Dataset	Ordered-layers	Volume	Avg. Layers	Domain
CGL V2	×	60K	4.80	E-commerce
PosterLayout	×	10K	4.73	Not Quantified
QB-Poster	×	5.1K	15.17	Chinese Social Media
WebUI	×	400K	32.00	Not Quantified
Crello	✓	23K	9.70	24 Diverse Themes
MLP-169K (Ours)	✓	169K	15.94	42 Diverse Themes

169K Layered Posters

2.6M Total Layers

15 Avg layers per Layout

42 Topic Domains

37+ Layout Sizes

Fig. 8: Overview of our MLP-169K. *Left:* comparison between MLP-169K and previous assets. *Middle:* the word cloud to visualize theme distribution of our datasets. *Right:* basic statistical information about MLP-169K.

z-ordering, broad thematic diversity, large scale, and a realistic number of layers per poster. We will publicly release both MLP-169K and the construction pipeline code, and more details about MLP-169K are included in the supplementary material.

5.2 Experimental Setup

Implementation: PosterCopilot employs Qwen-2.5-VL-7B-Instruct [1] as backbone; the generative agent employs Qwen-Image-Edit-2509 [52] as T2I model; the reception model uses Qwen-2.5-7B [41]; and RLAF utilizes VisualQuality-R1 [54] as reward model. MLP-169K is split into approximately 160K samples for training and 9K samples for test. Training proceeds in three stages: (1) PSFT for 3 epochs on 140K samples with perturbation parameters $\sigma = 2.5, n = 5, \lambda_{pert} = 0.5$; (2) RL-VRA on 20K samples featuring complex layout rules from MLP-169K with $\lambda_{size} = 0.6, \lambda_{AR} = 0.4, \delta = 0.6$; and (3) RLAF on 1K expert-validated samples with $\lambda_{aes} = 2$. All experiments run on 8×RTX H20 GPUs. Analysis of all hyperparameters and GRPO settings are detailed in the supplementary material.

Metrics. We adopt a two-stage evaluation framework that consolidates and extends prior evaluation protocols. The first stage subsumes all geometric metrics adopted by AesthetiQ [39] and Graphist [9], quantitatively assessing layout accuracy using Mean-IoU (IoU), Inverse Order Pair Ratio (IoPR) [10], and Aspect Ratio Distortion (ARD) [60] on three benchmarks: Crello and WebUI, where PosterCopilot is evaluated in a zero-shot setting, and the MLP-169K test set. These metrics jointly capture spatial overlap, layer ordering correctness, and geometric fidelity, respectively. The second stage builds upon the evaluation set-

tings of CreatiPoster [58] and POSTA [4] with broader and stricter criteria, assessing overall aesthetic quality through both human judgment and GPT-5 scoring across seven established design dimensions: Layout Rationality [12, 59], Text Legibility [6], Element Preservation [30], Style Consistency [47], Instruction Following [42], Visual Appeal [42], and Design Utility [26]. For GPT-5 evaluation, we exclude Element Preservation and Instruction Following, as reliably quantifying exact asset retention and abstract instruction alignment remains challenging for current Vision Language Models (VLM), which may compromise the fairness of evaluation. These two criteria are assessed solely by human evaluators. For brevity, we abbreviate each metric by its initials in subsequent tables (e.g., IF for Instruction Following).

Baselines: In the first stage, we compare PosterCopilot against academic methods that generate explicit layouts for poster composition. In the second stage, we move beyond geometric metrics to assess real-world design quality, as graphic design fundamentally prioritizes visual impact over calculated scores. Here, we expand the comparison to include the strongest academic baseline: LaDeCo [32], as well as leading commercial products across different paradigms not limited to VLM-based methods: CreatiPoster [58], Microsoft Designer, Nano-Banana, and its Pro variant. We additionally benchmark against Qwen-VL-2.5-72B-Instruct [1], a general-purpose VLM ten times the scale of PosterCopilot, to further underscore our model’s design effectiveness.

User study and GPT-5 evaluation procedure: For human evaluation, we conducted pairwise, binary-choice comparisons against each baseline. We sampled 30 examples per baseline—10 each from Crello, WebUI, and MLP-169K—all generated using identical prompts and fully provided assets from their respective test sets. We collected 5 judgments per example, totaling 750 responses from 40 professional graphic designers and 40 general users. To our best knowledge, our user evaluation is among the largest in scale within this field, ensuring convincing and rigorous assessment. For poster generation, T2I-based baselines (Microsoft Designer, Nano-Banana, and Nano-Banana Pro) are conditioned on all user assets with target canvas dimensions and design prompts to directly generate poster images. Layout-based methods (CreatiPoster [58], LaDeCo [32], Qwen-VL-2.5-72B-Instruct [1], and PosterCopilot) output layouts in JSON format, which are then converted to poster images via a unified lossless rendering script. Note that CreatiPoster requires ground-truth foreground layouts as input, placing our method at a comparative disadvantage, yet PosterCopilot still achieves a significant lead across all metrics. For GPT-5 evaluation, we used in-context learning to align the model with our scoring criteria and prompted GPT-5 to evaluate all results ten times, taking the average as the final score (Temp=0.7, Top-p=0.9). Please refer to the supplementary material for more details such as GPT-5 prompt, version details and access methods for all proprietary models.

Table 1: Main Results on Crello [56], WebUI [53], and MLP-169K. We evaluate models across spatial geometry dimensions. The best and second best performances are highlighted. $\uparrow$ indicates higher is better, and $\downarrow$ indicates lower is better.

Method	Crello			WebUI			MLP-169K		
	IoU(%) $\uparrow$	ARD $\downarrow$	IoPR $\downarrow$	IoU(%) $\uparrow$	ARD $\downarrow$	IoPR $\downarrow$	IoU(%) $\uparrow$	ARD $\downarrow$	IoPR $\downarrow$
LACE [5]	23.18	2.19	0.76	17.88	2.34	0.87	12.47	1.79	0.61
PosterLLaVa [57]	25.18	1.72	0.68	30.19	1.56	0.72	10.75	0.97	0.73
LayoutNUWA [48]	25.74	1.68	0.59	32.16	1.38	0.63	12.61	1.03	0.44
LaDeCo [32]	41.09	0.51	0.34	46.36	0.93	0.39	22.14	0.78	0.35
AesthetiQ [39]	42.83	x	x	48.29	x	x	x	x	x
PosterCopilot (Ours)	54.31	0.12	0.03	57.19	0.19	0.02	34.23	0.04	0.04

Table 2: Main Results of User Study and GPT-5 Evaluation. We evaluate models across layout and aesthetic dimensions. For the User Study, values represent each baseline’s win rate (%) against PosterCopilot. $\uparrow$ indicates a higher value means a stronger competitor. For GPT-5 evaluation, $\uparrow$ indicates higher scores are better. The best and second best performances in each column are highlighted.

Method	Win Rate Against PosterCopilot (%) $\uparrow$								GPT-5 Evaluation $\uparrow$				
	EP	LR	TL	SC	DU	VA	IF	AWR	LR	TL	SC	DU	VA
Proprietary Models													
Microsoft Designer	7.2	12.0	12.0	14.4	8.8	12.8	0.0	9.6	5.57	5.65	5.91	5.91	4.74
Nano-Banana Pro	11.3	28.9	26.1	35.6	38.9	35.4	23.7	28.6	6.37	5.49	6.47	6.30	5.88
Nano-Banana	6.4	23.2	27.2	16.0	19.2	20.0	12.0	17.7	6.50	6.08	6.23	6.27	5.46
CreatiPoster	7.2	30.4	33.6	25.6	23.2	29.6	30.4	25.7	6.19	5.38	5.96	6.23	5.23
Open-Source Models													
Qwen-2.5-vl-72B	3.2	0.8	4.0	1.6	1.6	3.2	0.8	2.2	3.85	3.27	4.42	4.58	3.42
LaDeCo	16.8	17.6	16.8	20.0	14.4	16.8	11.2	16.2	5.23	4.96	5.27	5.38	4.00
PosterCopilot (Ours)	-	-	-	-	-	-	-	-	6.52	5.96	6.68	6.60	6.16

6 Results and Analysis

6.1 Results of geometric metrics

As shown in Table 1, PosterCopilot achieves the best performance across all datasets and metrics. On Crello and WebUI, it outperforms the second-best method by 11.48% and 8.90% in IoU, while reducing ARD and IoPR by over 76% and 91%, respectively. MLP-169K, our constructed dataset with higher layout complexity, poses a greater challenge for all methods, yet PosterCopilot still leads LaDeCo by 12.09% in IoU with an ARD of only 0.04. Since AesthetiQ has not released its model or source code, we directly cite the IoU results reported in the original paper for Crello and WebUI, and other metrics cannot be reproduced. Even so, PosterCopilot surpasses AesthetiQ by a clear margin on both benchmarks. These results confirm the effectiveness of our approach in generating accurate and well-ordered layouts.

6.2 Results of human evaluation and VLM judge

As shown in Tab. 2, PosterCopilot achieves dominant win rates against all baselines in the user study, with even the strongest competitor Nano-Banana Pro reaching only 28.6% Average win rate (AWR). Notably, while Nano-Banana Pro and CreatiPoster perform competitively on individual aesthetic dimensions such as SC and DU, their overall performance still falls substantially behind PosterCopilot. In the GPT-5 evaluation, PosterCopilot ranks first on four out of five metrics. Nano-Banana achieves the highest Text Legibility, but this comes at the cost of discarding user-provided elements to reduce layout complexity, a trade-off evidenced by its notably low EP score (6.4%) in the user study. PosterCopilot, by contrast, faithfully preserves all assets while maintaining competitive legibility.

6.3 Ablation Study

Ablation on training stages. To investigate the impact of the three-stage training strategy on enhancing the model’s layout reasoning capability, we conducted an ablation study on MLP-169K test set. As shown in Tab.3a, leveraging SFT alone lead to a drop in all geometric metrics compared to PSFT strategy. Skipping the RL-VRA stage to directly leverage the RLAF stage leads to a significant drop in all metrics. Surprisingly, jointly training the RL-VRA and RLAF stages results in an even steeper decline. In stark contrast, our PosterCopilot demonstrates steady performance gains through a strictly sequential, progressive training pipeline (PSFT, RL-VRA, then RLAF), ultimately achieving state-of-the-art results across all metrics. Furthermore, the slight decrease in the IoU metric observed after the RL-VRA stage is entirely reasonable. RLAF stage inherently encourages the model to explore aesthetically pleasing layouts rather than strictly adhering to existing ground-truth designs. We also conduct user study on core training stages. As shown in Tab. 4, human aesthetic scores improve significantly following both the RL-VRA and RLAF stages, demonstrating the effectiveness of human aesthetic guidance in the RLAF phase.

Ablation on RL-VRA rewards. We conducted an ablation study on MLP-169K test set to examine the effects of different rewards of RL-VRA stage. Results in Tab. 3b on the first two rewards reveal the spatial coherence reward (r_{DIOU}) substantially enhances layout accuracy, while the element fidelity reward ($r_{AR} + r_{size}$) improves preservation of element sizes and proportions. Their combination yields optimal performance.

6.4 Justification for reward model in RLAF

To validate VisualQuality-R1 as the aesthetic reward model, we conducted a human alignment study. We sampled 200 layout pairs generated by PosterCopilot from MLP-169K and rendered them into posters, and then collected preference annotations from 40 general participants following the same protocol as

Table 3: Ablation studies on training stages and RL-VRA components. Tab. 3a shows the progressive improvements after different training stages. Tab. 3b reveals the effects of three individual reward functions during the RL-VRA stage. The **best** and **second best** performances are highlighted.

(a) Ablation on Core Training Stages.

Config	Training Stage			Metrics		
	PSFT	RL-VRA	RLAF	IoU(%) $\uparrow$	ARD $\downarrow$	IoPR $\downarrow$
SFT	$\times$	$\times$	$\times$	28.53	0.851	0.42
PSFT	$\checkmark$	$\times$	$\times$	31.14	0.699	0.38
PSFT+RLAF	$\checkmark$	$\times$	$\checkmark$	27.12	0.531	0.33
PSFT+Joint	$\checkmark$	$\checkmark^\dagger$	$\checkmark^\dagger$	24.71	0.712	0.57
PSFT+RL-VRA	$\checkmark$	$\checkmark$	$\times$	34.72	0.061	0.18
Ours (Full)	$\checkmark$	$\checkmark$	$\checkmark$	34.23	0.045	0.04

$\dagger$ Joint: RL-VRA and RLAF trained jointly rather than sequentially.

(b) Ablation on RL-VRA reward components.

ID	Reward Component			Metrics		
	r_{fmt}	r_{DIOU}	$r_{\text{AR}}+r_{\text{size}}$	IoU(%) $\uparrow$	ARD $\downarrow$	IoPR $\downarrow$
I	$\checkmark$	$\times$	$\times$	31.73	0.707	0.37
II	$\checkmark$	$\checkmark$	$\times$	33.91	0.734	0.21
III	$\checkmark$	$\checkmark$	$\checkmark$	34.72	0.061	0.18

$\dagger$ Joint: RL-VRA and RLAF trained jointly rather than sequentially.

Table 4: Human evaluation comparison across different training stages.

Method	Win Rate Against PosterCopilot (%) $\uparrow$							
	EP	LR	TL	SC	DU	VA	IF	AWR
PSFT	16.3	15.8	35.9	15.1	28.7	18.7	39.1	24.3
PSFT+RL-VRA	38.8	40.6	37.0	27.6	35.1	30.5	44.7	36.4

Sec. 6.2. Each pair was independently scored by VisualQuality-R1. We find that VisualQuality-R1’s outputs aligned with human preferences on 82.5% of the total pairs. Further analysis of the remaining 17.5% disagreement cases reveals that the majority involve layout pairs of comparable quality where human annotators themselves showed low inter-rater agreement (Fleiss’ $\kappa < 0.3$), suggesting that most discrepancies stem from inherent ambiguity rather than systematic bias. These results confirm that VisualQuality-R1 serves as a reliable proxy for human aesthetic judgment in our RLAF framework.

7 Conclusion

PosterCopilot advances automated poster design through a progressive training paradigm that injects geometric rules and human aesthetics into LMM-based layout planning, a large-scale multi-layer dataset MLP-169K with rational layer granularity, and a generative agent for lossless, multi-round, layer-wise editing. Current limitations include the absence of a community-recognized evaluation benchmark and a poster-specific aesthetic reward model. We hope PosterCopilot and MLP-169K provide a useful foundation for future work in this direction.

8 Acknowledgement

This study is supported by the Fundamental Research Funds for the Central Universities (KG2025XX) and Jiangsu Science and Technology Major Project (BG2025035). This research is also supported by cash and in-kind funding from Nanjing Kunpeng & Ascend Center of Cultivation and industry partner(s).

References

1. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
2. Baldwin, J., Roberts, L.: Visual communication: from theory to practice, vol. 30. Ava Publishing (2006)
3. Cha, I., Wong, R.Y.: Understanding socio-technical factors configuring ai non-use in ux work practices. In: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems. pp. 1–17 (2025)
4. Chen, H., Xu, X., Li, W., Ren, J., Ye, T., Liu, S., Chen, Y.C., Zhu, L., Wang, X.: Posta: A go-to framework for customized artistic poster generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 28694–28704 (2025)
5. Chen, J., Zhang, R., Zhou, Y., Jain, R., Xu, Z., Rossi, R., Chen, C.: Towards aligned layout generation via diffusion model with aesthetic constraints. arXiv preprint arXiv:2402.04754 (2024)
6. Chen, J., Huang, Y., Lv, T., Cui, L., Chen, Q., Wei, F.: Textdiffuser: Diffusion models as text painters. *Advances in Neural Information Processing Systems* **36**, 9353–9387 (2023)
7. Chen, S., Lai, J., Gao, J., Shi, H., Liu, Z., Ye, T., Luo, J., Wei, X., Zhu, L.: Posteromni: Generalized artistic poster creation via task distillation and unified reward feedback. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5978–5987 (2026)
8. Chen, S., Lai, J., Gao, J., Ye, T., Chen, H., Shi, H., Shao, S., Lin, Y., Fei, S., Xing, Z., et al.: Postercraft: Rethinking high-quality aesthetic poster generation in a unified framework. arXiv preprint arXiv:2506.10741 (2025)
9. Cheng, Y., Zhang, Z., Yang, M., Hui, N., Li, C., Wu, X., Shao, J.: Graphic design with large multimodal model. arXiv preprint arXiv:2404.14368 (2024)
10. Cheng, Y., Zhang, Z., Yang, M., Nie, H., Li, C., Wu, X., Shao, J.: Graphic design with large multimodal model. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 2473–2481 (2025)
11. Davies, A.O., Nzoyem, R., Ajmeri, N., et al.: Language models do not embed numbers continuously. arXiv preprint arXiv:2510.08009 (2025)
12. Feng, W., Zhu, W., Fu, T.j., Jampani, V., Akula, A., He, X., Basu, S., Wang, X.E., Wang, W.Y.: Layoutgpt: Compositional visual planning and generation with large language models. *Advances in Neural Information Processing Systems* **36** (2024)
13. Girshick, R.: Fast R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 1440–1448 (2015)
14. Hou, C., Wei, G., Chen, Z.: High-fidelity diffusion-based image editing. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 2184–2192 (2024)

15. Hsu, H., He, X., Peng, Y., Kong, H., Zhang, Q.: Posterlayout: A new benchmark and approach for content-aware visual-textual presentation layout. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 6018–6026 (2023)
16. Hsu, H., Peng, Y.: Postero: Structuring layout trees to enable language models in generalized content-aware layout generation. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 8117–8127 (2025)
17. Hu, X., Chen, H., Qi, Z., Zhang, H., Hong, D., Shao, J., Wu, X.: Dreamposter: A unified framework for image-conditioned generative poster design. *arXiv preprint arXiv:2507.04218* (2025)
18. Huang, R., Cai, K., Han, J., Liang, X., Pei, R., Lu, G., Xu, S., Zhang, W., Xu, H.: Layerdiff: Exploring text-guided multi-layered composable image synthesis via layer-collaborative diffusion model. In: *European Conference on Computer Vision*. pp. 144–160. Springer (2024)
19. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. *arXiv preprint arXiv:2410.21276* (2024)
20. Inoue, N., Kikuchi, K., Simo-Serra, E., Otani, M., Yamaguchi, K.: Layoutdm: Discrete diffusion model for controllable layout generation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 10167–10176 (2023)
21. Janusz, M., Wojnar, T., Li, Y., Benini, L., Adamczewski, K.: One shot vs. iterative: Rethinking pruning strategies for model compression. *arXiv preprint arXiv:2508.13836* (2025)
22. Jia, P., Li, C., Yuan, Y., Liu, Z., Shen, Y., Chen, B., Chen, X., Zheng, Y., Chen, D., Li, J., et al.: Cole: A hierarchical generation framework for multi-layered and editable graphic design. *arXiv preprint arXiv:2311.16974* (2023)
23. Jiang, Q., Huo, J., Chen, X., Xiong, Y., Zeng, Z., Chen, Y., Ren, T., Yu, J., Zhang, L.: Detect anything via next point prediction. *arXiv preprint arXiv:2510.12798* (2025)
24. Karimi, P., Rezwana, J., Siddiqui, S., Maher, M.L., Dehbozorgi, N.: Creative sketching partner: an analysis of human-ai co-creativity. In: *Proceedings of the 25th international conference on intelligent user interfaces*. pp. 221–230 (2020)
25. Khan, A., Shokrizadeh, A., Cheng, J.: Beyond automation: How designers perceive ai as a creative partner in the divergent thinking stages of ui/ux design. In: *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. pp. 1–12 (2025)
26. Kong, X., Jiang, L., Chang, H., Zhang, H., Hao, Y., Gong, H., Essa, I.: Blt: Bidirectional layout transformer for controllable layout generation. In: *European Conference on Computer Vision*. pp. 474–490. Springer (2022)
27. Li, C., Hu, X., Abousamra, S., Xu, M., Chen, C.: Spatial diffusion for cell layout generation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 481–491. Springer (2024)
28. Li, F., Liu, A., Feng, W., Zhu, H., Li, Y., Zhang, Z., Lv, J., Zhu, X., Shen, J., Lin, Z.: Relation-aware diffusion model for controllable poster layout generation. In: *Proceedings of the 32nd ACM International Conference on Information & Knowledge Management*. pp. 1249–1258 (2023)
29. Li, S.: Llms as layout designers: A spatial reasoning perspective. *arXiv e-prints pp. arXiv-2509* (2025)

30. Li, Y., Liu, H., Wu, Q., Mu, F., Yang, J., Gao, J., Li, C., Lee, Y.J.: Gligen: Open-set grounded text-to-image generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22511–22521 (2023)
31. Lin, J., Guo, J., Sun, S., Yang, Z., Lou, J.G., Zhang, D.: Layoutprompter: Awaken the design ability of large language models. *Advances in Neural Information Processing Systems* **36**, 43852–43879 (2023)
32. Lin, J., Sun, S., Huang, D., Liu, T., Li, J., Bian, J.: From elements to design: A layered approach for automatic graphic design composition. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 8128–8137 (2025)
33. Liu, X., Zhang, T., Gu, Y., Iong, I.L., Xu, Y., Song, X., Zhang, S., Lai, H., Liu, X., Zhao, H., et al.: Visualagentbench: Towards large multimodal models as visual foundation agents. *arXiv preprint arXiv:2408.06327* (2024)
34. Malhi, I., Dutta, P., Talius, E., Ma, S., Driscoll, B., Holden, K., Pruthi, G., Narayanaswamy, A.: Preserving product fidelity in large scale image recontextualization with diffusion models. *arXiv preprint arXiv:2503.08729* (2025)
35. Meggs, P.B., Purvis, A.W., Maxa, S., Sanders, M.: *Meggs’ history of graphic design*. John Wiley & Sons (2025)
36. Muehlhaus, M., Steimle, J.: Interaction design with generative ai: An empirical study of emerging strategies across the four phases of design. *arXiv preprint arXiv:2411.02662* (2024)
37. Ocampo Blanco, R., Bown, O.: Integrating generative ai into creative workflows: Dealing with consistency, scene control, and refinement in a professional image generation case study. In: International Conference on Computational Creativity (2024)
38. Oksanen, J.: Bridging the integrity gap: Towards ai-assisted design research. In: Extended abstracts of the CHI conference on human factors in computing systems. pp. 1–5 (2024)
39. Patnaik, S., Jain, R., Krishnamurthy, B., Sarkar, M.: Aesthetiq: Enhancing graphic layout design via aesthetic-aware preference alignment of multi-modal large language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 23701–23711 (2025)
40. Qu, L., Li, H., Wang, W., Liu, X., Li, J., Nie, L., Chua, T.S.: Silmm: Self-improving large multimodal models for compositional text-to-image generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 18497–18508 (2025)
41. Qwen Team: Qwen2.5: A party of foundation models. Blog post (2024), accessed: 2024
42. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems* **35**, 36479–36494 (2022)
43. Schulman, J., Wolski, F., Dhara, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* (2017)
44. Selvam, K.P.: Why large language models fail at precision regression. Blog post (2025), accessed: 2025
45. Seol, J., Kim, S., Yoo, J.: Posterllama: Bridging design ability of language model to content-aware layout generation. In: European Conference on Computer Vision. pp. 451–468. Springer (2024)
46. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300* (2024)

47. Sohn, K., Ruiz, N., Lee, K., Chin, D.C., Blok, I., Chang, H., Barber, J., Jiang, L., Entis, G., Li, Y., et al.: Styledrop: Text-to-image generation in any style. arXiv preprint arXiv:2306.00983 (2023)
48. Tang, Z., Wu, C., Li, J., Duan, N.: Layoutnuwa: Revealing the hidden layout expertise of large language models. arXiv preprint arXiv:2309.09506 (2023)
49. Wang, X., Fu, S., Huang, Q., He, W., Jiang, H.: Ms-diffusion: Multi-subject zero-shot image personalization with layout guidance. arXiv preprint arXiv:2406.07209 (2024)
50. Wang, Y., Lin, Z., Teng, Y., Zhu, Y., Ren, S., Feng, J., Liu, X.: Bridging continuous and discrete tokens for autoregressive visual generation. arXiv preprint arXiv:2503.16430 (2025)
51. Wang, Z., Xie, E., Li, A., Wang, Z., Liu, X., Li, Z.: Divide and conquer: Language models can plan and self-correct for compositional text-to-image generation. arXiv preprint arXiv:2401.15688 (2024)
52. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., ming Yin, S., Bai, S., Xu, X., Chen, Y., Chen, Y., Tang, Z., Zhang, Z., Wang, Z., Yang, A., Yu, B., Cheng, C., Liu, D., Li, D., Zhang, H., Meng, H., Wei, H., Ni, J., Chen, K., Cao, K., Peng, L., Qu, L., Wu, M., Wang, P., Yu, S., Wen, T., Feng, W., Xu, X., Wang, Y., Zhang, Y., Zhu, Y., Wu, Y., Cai, Y., Liu, Z.: Qwen-image technical report (2025)
53. Wu, J., Wang, S., Shen, S., Peng, Y.H., Nichols, J., Bigham, J.: Webui: A dataset for enhancing visual ui understanding with web semantics. ACM Conference on Human Factors in Computing Systems (CHI) (2023)
54. Wu, T., Zou, J., Liang, J., Zhang, L., Ma, K.: Visualquality-r1: Reasoning-induced image quality assessment via reinforcement learning to rank. arXiv preprint arXiv:2505.14460 (2025)
55. Wu, X.K., Chen, M., Li, W., Wang, R., Lu, L., Liu, J., Hwang, K., Hao, Y., Pan, Y., Meng, Q., et al.: Llm fine-tuning: Concepts, opportunities, and challenges. Big Data and Cognitive Computing **9**(4), 87 (2025)
56. Yamaguchi, K.: Canvasvae: Learning to generate vector graphic documents. ICCV (2021)
57. Yang, T., Luo, Y., Qi, Z., Wu, Y., Shan, Y., Chen, C.W.: Posterllava: Constructing a unified multi-modal layout generator with llm. arXiv preprint arXiv:2406.02884 (2024)
58. Zhang, Z., Cheng, Y., Hong, D., Yang, M., Shi, G., Ma, L., Zhang, H., Shao, J., Wu, X.: Creatiposter: Towards editable and controllable multi-layer graphic design generation. arXiv preprint arXiv:2506.10890 (2025)
59. Zheng, G., Zhou, X., Li, X., Qi, Z., Shan, Y., Li, X.: Layoutdiffusion: Controllable diffusion model for layout-to-image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22490–22499 (2023)
60. Zheng, Z., Wang, P., Liu, W., Li, J., Ye, R., Ren, D.: Distance-IoU loss: Faster and better learning for bounding box regression. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 12993–13000 (2020)
61. Zou, X., Zhang, W., Zhao, N.: From fragment to one piece: A review on ai-driven graphic design. Journal of Imaging **11**(9), 289 (2025)