i-Design: Step-by-Step Graphic Layout Design with Progressive Aesthetic Policy Optimization

Sohan Patnaik[1*], Rishabh Jain[2*†],
Balaji Krishnamurthy, and Mausoom Sarkar[1]

[1] Media and Data Science Research, Adobe
[2] Tether
{soha@adobe.com, rishabh.jain@tether.io}

1 Introduction

The creation of visually appealing layouts is the cornerstone of graphic design applications. The growing demand for automated design systems underscores their broad impact across multiple domains, including poster design (20; 41), user interface design (10; 22), document layout generation (44; 50), and presentation slide creation (11). The quality of a layout not only determines its aesthetic appeal but also influences how viewers perceive information hierarchy, balance, and intent (18). Consequently, automating layout generation has emerged as a long-standing goal across the computer vision and design intelligence communities, with far-reaching implications for creative assistance and large-scale content generation (20; 26; 33; 34; 28).

In this work, we focus on *2D graphic layout generation*, where visual and textual elements (e.g., titles, images, shapes) are arranged on a canvas. Unlike other layout planning problems (e.g., 3D scene layout, object placement, or interior design), this task emphasizes perceptual qualities such as visual hierarchy, alignment, balance, and typography, central to human-centered design. Early rule-based (33; 34) and energy-based methods (24; 26) laid the groundwork for automatic layout synthesis but relied heavily on handcrafted heuristics and fixed constraints, limiting adaptability and scalability. With the advent of deep generative modeling, transformer- and diffusion-based architectures such as LayoutTransformer (15), BLT (25), and LayoutDiffusion (47; 24) enabled data-driven spatial reasoning over element coordinates. Multimodal large language models (LLMs) such as PosterLLaVa (45), LayoutNUWA (39), AestheticQ (35), have achieved strong performance by coupling visual understanding with aesthetic-aware fine-tuning. Further, hierarchical and structured approaches, PosterO (17), show that representing layouts as tree-structured compositions enhances content awareness and generalization across diverse design domains (14; 6).

Despite these advances, most existing methods formulate layout generation as a single-shot prediction problem, where all elements are placed simultaneously on the canvas. This oversimplified setting (Fig. 1) fails to capture the iterative

[*] These authors contributed equally to this work.
[†] Work done while at Adobe.

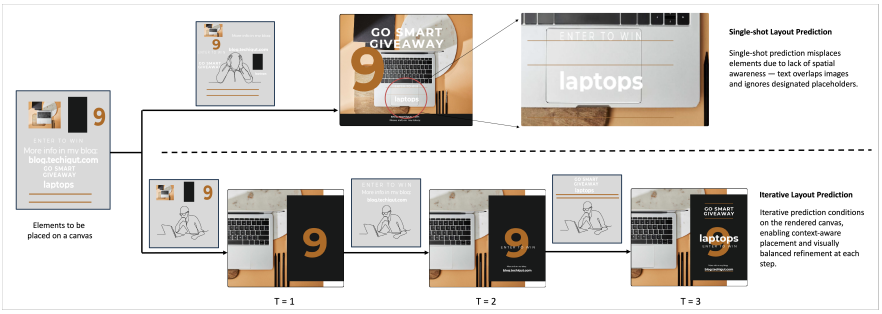

Fig. 1. Comparison between single-shot and iterative layout prediction. Single-shot prediction places all elements simultaneously, often misaligning text and images due to the lack of spatial feedback, resulting in overlaps and misplaced content. In contrast, iterative layout prediction conditions on the rendered canvas at each step ($T = 1, 2, 3$), progressively placing the elements to achieve better layout.

nature of human design, which evolves progressively through feedback, adjustment, and refinement (18; 25; 36). Humans seldom design in one pass. They iteratively position a few elements, evaluate visual balance, and refine compositions. In contrast, current models lack such progressive reasoning, often producing visually inconsistent or unbalanced layouts when all elements are predicted at once. Conventional training paradigms impose rigid supervision via coordinate- or token-level losses (5; 38; 7; 47; 43), penalizing valid aesthetic variations and limiting generalization across unseen design styles or complex templates. While preference-tuning based techniques (35; 5) have shown promise in related domains such as architectural or facility layout generation, their application to graphic layout design remains primitive and largely unexplored.

To this end, we pose a fundamental research question: *"Can layout generation be formulated as a progressive, iterative design process, guided by aesthetic feedback over layout trajectories, to yield stronger aesthetic-aware generators?"* Vision–Language Models (VLMs) possess remarkable spatial reasoning and visual understanding, yet their potential remains underutilized within current layout-training paradigms, which typically constrain models to predict all element coordinates in a single forward pass. Such formulations overlook the inherently sequential and feedback-driven nature of human design, where designers progressively place elements, evaluate visual balance, and refine compositions as the canvas evolves. Consequently, existing approaches lack the capacity to adapt or self-correct during generation. While some prior works explore iterative layout prediction (39; 31), our contribution lies not in stepwise prediction itself but in casting layout generation as a trajectory-level decision process optimized using rendered aesthetic feedback. Instead of relying on offline supervision or single-shot preference alignment, our framework performs online policy optimization over rendered layout rollouts, enabling aesthetic credit assignment across intermediate design states and encouraging globally coherent compositions.

In this paper, we introduce ***i-Design***, a framework that models layout generation as a human-like progressive construction process. Instead of placing all elements simultaneously, *i-Design* incrementally places k elements at a time, conditioned on previously rendered partial canvases. Our approach combines Progressive Imitation Learning (PIL) on partial layout states with a novel **Progressive Aesthetic Policy Optimization (PAPO)** stage that improves the model using aesthetic feedback from VLM layout judge (12). Unlike prior offline preference-alignment methods (35), PAPO performs online optimization by dynamically sampling multiple rollouts, enabling exploration of diverse layout trajectories and learning adaptive aesthetic preferences grounded in visual feedback. The judge, an open-source vision classifier, performs pairwise comparisons between rendered layouts to identify better visual composition. From these judgments, we construct a directed comparison graph and estimate a global aesthetic consensus by aggregating preferences across candidates. Ground-truth layouts are included as anchors to stabilize training and enrich reward signals, allowing consensus scores to guide the model toward aesthetic quality rather than mere geometric accuracy. Layouts are rendered using Skia (7; 35; 47), ensuring learning is grounded in actual visual composition. This formulation encourages context-aware prediction, enabling the model to adapt placements of remaining elements as the canvas evolves. Our main contributions are summarized below.

1. We introduce *i-Design*, a framework that formulates 2D graphic layout generation as a *trajectory-level decision process*, progressively placing element groups conditioned on rendered intermediate canvases.
2. We propose a training paradigm combining *Progressive Imitation Learning* (PIL) and *Progressive Aesthetic Policy Optimization* (PAPO), which enables online policy optimization using aesthetic feedback from rendered layout rollouts and assigns rewards across intermediate design states.
3. Extensive experiments on Crello and WebUI demonstrate that *i-Design* achieves SoTA performance in both geometric fidelity and aesthetic preference, validating the efficacy of trajectory-level aesthetic optimization for layouts.

2 Related Works

Graphic Layout Generation. Automatic layout generation aims to arrange visual and textual elements on a canvas to produce coherent and aesthetically balanced designs. With the advent of deep generative models, GAN-based methods (27; 24; 51) modeled geometric dependencies among layout elements, while transformer-based architectures such as LayoutTransformer (15), VTN (1), and BLT (25) captured long-range spatial dependencies. Multimodal encoder–decoder frameworks, including CanvasVAE (44), FlexDM (20), and ICVT (4), integrated textual and visual inputs for content-aware layout generation, though most remain constrained by limited design diversity.

Diffusion-guided Layout Generation. Diffusion models have demonstrated strong performance in layout design, offering smooth latent transitions and

higher diversity. PLay (9) employs latent diffusion guided by layout constraints, while LayoutDM (5) and LDGM (19) adopt discrete diffusion (2) to unify layout generation under attribute-specific corruption. Continuous diffusion models such as LACE (8) and LayoutDiT (29) introduce differentiable aesthetic constraints and content-aware representations, and Layout-Corrector (21) alleviates layout-sticking in discrete diffusion. However, these methods primarily optimize for geometric plausibility and overlook semantic intent or aesthetic coherence.

LLMs and Aesthetic-aware Layout Generation. Large Language Models (LLMs) have been adapted for layout generation by representing designs as structured sequences (e.g., JSON/XML trees) (48; 32; 16; 23). Early adaptations such as LayoutNUWA, based on LLaMA and CodeLLaMA (40; 37) focused on content-agnostic generation, while retrieval-augmented methods (3; 30) improved in-context reasoning. Although token-based layout representations align well with LLMs, conventional cross-entropy training wrongly penalizes good aesthetic variations. To bridge this gap, recent works incorporate aesthetic alignment, i.e, PosterLLaVA (45) adds visual-language conditioning. LaDeCo (31) introduces an iterative element placement strategy, while AesthetiQ (35) employs preference tuning with a vision–language judge. Nevertheless, most existing methods still operate under a single-shot generation paradigm and do not integrate iterative prediction with online reinforcement learning to capture the progressive nature of human design.

3 Methodology

We propose *i-Design* (Fig. 2), a unified framework that formulates layout generation as a progressive composition process, mirroring human design behavior. *i-Design* incrementally builds layouts by placing small groups of elements conditioned on the evolving visual canvas. This progressive formulation naturally embodies the iterative decision-making and visual feedback loop inherent to human design. The framework consists of two training stages: **(1) Progressive Imitation Learning (PIL)**, where the model learns to predict unplaced elements given partially rendered layouts, acquiring spatial and compositional priors; and **(2) Progressive Aesthetic Policy Optimization (PAPO)**, an online reinforcement learning stage that refines the model using aesthetic feedback from a learned visual judge.

3.1 Iterative Layout Generation

The objective of layout generation is to arrange design elements on an empty canvas to produce a visually coherent and aesthetically pleasing composition. Let a layout consist of N design elements $E = \{e_i = (p_i, s_i, \tau_i) \mid i = 1, \ldots, N\}$, where $p_i = (x_i, y_i)$ denotes the center coordinates, $s_i = (w_i, h_i)$ the element size, and $\tau_i \in \mathcal{T}$ the element type, with $\mathcal{T} = \{\texttt{text}, \texttt{image}, \texttt{shape}, \texttt{background}\}$. Given a canvas of width W_c and height H_c, the renderer $R(\cdot)$ composites all elements to form the final layout $G = R(E)$, where R is a non-differentiable

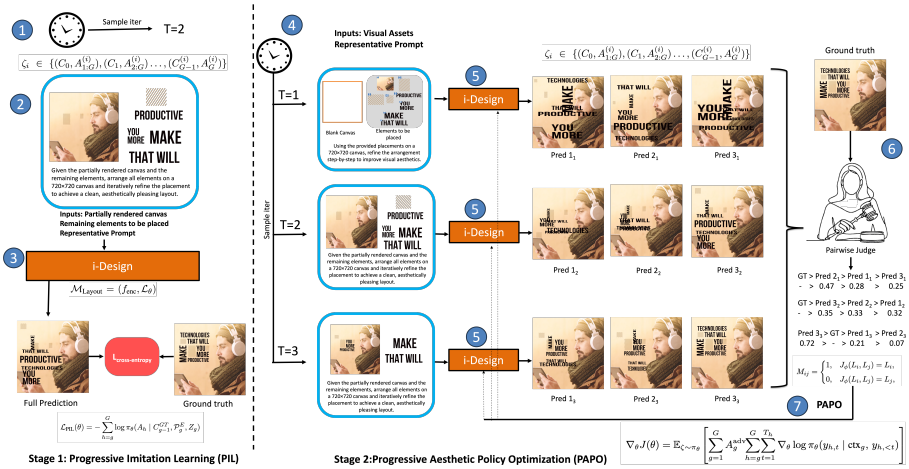

Fig. 2. Overview of the proposed *i-Design* framework. Following are the steps in the training pipeline of *i-Design*. **(1)** A training data is sampled from ζ_i which is defined as a sequence of partial canvas states and placement groups $\{(C_0, A_{1:G}), \ldots, (C_{G-1}, A_G)\}$. **(2)** At each iteration, a partial canvas is rendered from previously placed elements, forming the visual context for subsequent placements. **(3)** During **Progressive Imitation Learning (PIL)**, the model $\mathcal{M}_{\text{Layout}} = (f_{\text{enc}}, \mathcal{L}_\theta)$ predicts all remaining element groups conditioned on the current canvas, optimized via cross-entropy loss $\mathcal{L}_{\text{PIL}}$. **(4–5)** In **Progressive Aesthetic Policy Optimization (PAPO)**, the model samples multiple layout rollouts for each input, and predicts the positions in a similar fashion as in PIL. **(6)** A VLM judge (12) J_ϕ performs pairwise comparisons between rendered layouts to form a directed preference matrix M_{ij}, indicating visually superior layouts. **(7)** These preferences are aggregated into consensus-based rewards that compute the policy gradient $\nabla_\theta J(\theta)$ for training. Together, PIL and PAPO enable *i-Design* to generate visually coherent, aesthetically pleasing layouts.

rendering operator. To emulate the progressive nature of human design, we divide the layout elements into ordered groups placed sequentially. Formally, $\mathcal{A} = \{A_1, A_2, \ldots, A_G\}$, where $G = \lceil N/k \rceil$ is the number of groups and each group A_g contains k_g elements such that $1 \leq k_g \leq k$ and $\sum_{g=1}^{G} k_g = N$. Grouping is deterministic and follows the z-order sequence of elements in the layout, without any learned grouping mechanism. Consider $N = 10$ and $k = 3$, the groups will be $A_1 = \{e_1, e_2, e_3\}$, $A_2 = \{e_4, e_5, e_6\}$, $A_3 = \{e_7, e_8, e_9\}$, and $A_4 = \{e_{10}\}$. The group index g acts as the discrete time step in the progressive placement process. We represent each group A_g as a sequence of tokens $y_g = (y_{g,1}, \ldots, y_{g,T_g})$ (35), where each token encodes a discretized component of the element bounding box $(pos_x, pos_y, pos_w, pos_h)$ along with separators.

At the beginning of each group placement step g, the model has access to the already rendered elements, denoted by $E^{<g} = \bigcup_{h<g} A_h$, which form a part of the partial layout constructed so far. Each element e_i, whether already placed or yet to be positioned, is verbalized into a sequence of tokens that describe its

semantic and geometric attributes. The verbalization depends on whether the element is already rendered on the canvas as depicted by Eq 1

$$\mathcal{V}(e_i, g) = \begin{cases} [\tau_i,\, \text{Content}(e_i),\, \text{BBox}(p_i, s_i)], & e_i \in E^{<g}, \\ [\tau_i,\, \text{Content}(e_i)], & e_i \notin E^{<g}, \end{cases} \tag{1}$$

where τ_i specifies the element category (such as `text` or `image`), $\text{Content}(e_i)$ denotes its tokenized textual or visual representation, and $\text{BBox}(p_i, s_i)$ are normalized bounding-box coordinate tokens, which are only provided for the elements that have already been placed. Note that we use discrete position tokens to represent the bounding boxes as defined in (35). We construct the iteration-specific multimodal prompt as the ordered concatenation of all element descriptions in their z-order as depicted by Eq 2.

$$\mathcal{P}_g^E = [\,\mathcal{V}(e_1, g),\, \mathcal{V}(e_2, g),\, \ldots,\, \mathcal{V}(e_N, g)\,]. \tag{2}$$

This verbalised prompt, fed as input to the model, provides contextual awareness of the current layout state, while preserving an abstract representation and the design intent for placing the remaining elements. We implement the layout generation model as a vision–language architecture, denoted $\mathcal{M}_{\text{Layout}} = (f_{\text{enc}}, \mathcal{L}_\theta)$, which combines a modality-aware encoder f_{enc} with a large language model (LLM) decoder $\mathcal{L}_\theta$. The encoder processes both textual and visual inputs, mapping each verbalized element sequence $\mathcal{V}(e_i, g)$ into a latent embedding $z_i^{(g)} = f_{\text{enc}}(\mathcal{V}(e_i, g)) \in \mathbb{R}^d$, where d denotes the embedding dimension. We concatenate all per-element embeddings to obtain the multimodal layout representation at iteration g defined in Eq 3.

$$Z_g = [\, z_1^{(g)}; z_2^{(g)}; \ldots; z_N^{(g)} \,] \in \mathbb{R}^{N \times d}. \tag{3}$$

We feed this latent representation Z_g into the language model $\mathcal{L}_\theta$, which autoregressively decodes a sequence of layout tokens representing bounding-box coordinates (e.g. pos_x, pos_y, pos_w, pos_h) for the remaining unplaced elements. Through this process, the vision–language model acts as a parametric layout policy π_θ that predicts a probability distribution over element positions tokens conditioned on the current visual context and the multimodal prompt. At each placement step g, the VLM observes the partially rendered canvas $C_{g-1} = R(E^{<g})$, along with the corresponding prompt $\mathcal{P}_g^E$ and the multimodal embedding matrix Z_g. It then predicts the placements of all remaining unplaced elements by sampling from its policy distribution (Eq 4).

$$A_{g:G} \sim \pi_\theta\big(A_{g:G} \mid C_{g-1}, \mathcal{P}_g^E, Z_g\big), \tag{4}$$

where $A_{g:G} = \{A_g, A_{g+1}, \ldots, A_G\}$ denotes the sequence of future element placements. This formulation enables the model to reason across the full set of remaining elements, while operating in a recurrent, context-aware manner.

Execution during inference. While the model is trained to predict placements for *all remaining elements* at each step, the inference procedure follows a

progressive rollout strategy to better mimic human design iteration. Specifically, only the next group A_g (comprising $k_g \leq k$ elements) is rendered at each iteration: $C_g = R(C_{g-1}, A_g)$, after which the updated canvas C_g is provided back to the model to infer the placements of the remaining groups $\{A_{g+1}, \ldots, A_G\}$. This process repeats until all elements are placed, producing the complete trajectory of states and actions (Eq 5).

$$\tau_{\text{traj}} = \{(C_0, A_1), (C_1, A_2), \ldots, (C_{G-1}, A_G)\}, \tag{5}$$

This culminates into the final rendered layout $C_G = R(E)$. This iterative inference mechanism allows the model to continuously place the elements as the layout evolves, leveraging visual feedback from previously placed elements. Such progressive conditioning leads to improved compositional coherence, alignment, and visual balance compared to single-pass layout generators.

3.2 Progressive Imitation Learning

Before introducing aesthetic feedback through RL, we warm up the layout policy π_θ via supervised imitation on ground-truth layouts. This stage teaches the model to predict the positions of *all remaining* groups conditioned on a partially rendered canvas, which provides a stable initialization for subsequent aesthetic optimization. Given a ground-truth layout with N elements partitioned into ordered groups $\mathcal{A} = \{A_1, \ldots, A_G\}$, we simulate progressive design states by rendering the first $g-1$ groups to obtain the partial canvas $C_{g-1}^{GT} = R(A_1^{GT}, \ldots, A_{g-1}^{GT})$, where, GT denotes ground-truth positition coordinates for the elements in each group. In the training step g, the model is conditioned on $(C_{g-1}^{GT}, \mathcal{P}_g^E, Z_g)$ and trained to predict the *entire sequence of future groups* $A_{g:G}^{GT} = \{A_g^{GT}, \ldots, A_G^{GT}\}$, allowing global prediction over the remaining placements. We minimise the cross entropy loss over the policy's conditional probablity against the ground truth positions of future groups at each training step as depicted in Eq. 6

$$\mathcal{L}_{\text{PIL}}(\theta) = -\sum_{h=g}^{G} \log \pi_\theta(A_h \mid C_{g-1}^{GT}, \mathcal{P}_g^E, Z_g) \tag{6}$$

where, $\pi_\theta(A_h \mid C_{g-1}^{GT}, \mathcal{P}_g^E, Z_g) = \sum_{t=1}^{T_h} \log \pi_\theta\big(y_{h,t} \mid C_{g-1}^{GT}, \mathcal{P}_g^E, Z_g, y_{h,<t}\big)$ depicts the token level log-probability of each of the remaining groups.

3.3 Progressive Aesthetic Policy Optimization

While supervised imitation establishes geometric validity, it does not enable the model to discern what makes a layout aesthetically superior. To overcome this, we introduce **Progressive Aesthetic Policy Optimization (PAPO)**, an online reinforcement learning method that adapts Grouped Sequence Policy Optimization (GSPO) (49) for layout generation using visual preference feedback and consensus-based reward aggregation. This stage forms a closed loop between

the layout policy, renderer, and judge, allowing the model to learn aesthetic quality directly from rendered layouts.

At each optimization step, the current policy π_θ samples m layout rollouts $\mathcal{Z} = \{\zeta_1, \zeta_2, \ldots, \zeta_m\}$ for a given design prompt. Each trajectory ζ_i corresponds to a progressive layout construction sequence $\zeta_i \in \{(C_0, A^{(i)}_{1:G}), (C_1, A^{(i)}_{2:G}) \ldots, (C^{(i)}_{G-1}, A^{(i)}_G)\}$, and is rendered into a layout image $L_i = R(\zeta_i)$ using the renderer. To anchor the comparison, the ground-truth layout L^{GT} is also included, forming the set $\mathcal{L} = \{L^{GT}, L_1, \ldots, L_m\}$. Recent works have introduced specialized models for evaluating the aesthetic quality of graphic layouts (13; 12). These approaches train layout-aware scoring or reward models using curated design datasets and human preference annotations, enabling reliable assessment of visual balance, alignment, and compositional quality. Such dedicated judges have been shown to better capture design-specific aesthetic preferences compared to generic vision-language models. Motivated by these advances, we employ the open-source *DesignSense* (12) model as our aesthetic judge. The judge J_ϕ receives pairs of rendered layouts (L_i, L_j) and predicts the preferred design based on compositional quality, visual balance, alignment, and overall layout coherence. Using these pairwise preferences, we construct a directed comparison matrix (Eq. 7) that captures the relative aesthetic ordering among the sampled layouts.

$$M_{ij} = \begin{cases} 1, & J_\phi(L_i, L_j) = L_i, \\ 0, & J_\phi(L_i, L_j) = L_j, \end{cases} \qquad i \neq j. \tag{7}$$

Since aesthetic preference is asymmetric, both comparison orders (L_i, L_j) and (L_j, L_i) are evaluated, forming a directed graph where edges encode preference relations. Each layout L_i is assigned an aesthetic strength score $S(L_i)$, indicating how often it is preferred over others, computed via iterative preference propagation over the directed graph (Eq. 8).

$$S(L_i) = (1 - \gamma) \sum_{L_j \in \mathcal{N}^-(L_i)} \frac{S(L_j)}{|\mathcal{N}^+(L_j)|} + \frac{\gamma}{|\mathcal{L}|}, \tag{8}$$

where γ is the damping coefficient, and $\mathcal{N}^-(L_i)$ and $\mathcal{N}^+(L_i)$ denote the sets of incoming and outgoing neighbors, respectively. Scores $S(L_i)$ are obtained via power iteration until convergence, producing a global aesthetic ranking. To compute trajectory-level rewards, we combine the raw aesthetic scores returned by the judge with an exponential term based on deviation from the ground-truth score. Let $s_i = S(L_i)$ and $s^{GT} = S(L^{GT})$ denote the scores of the generated and reference layouts, respectively. Eq. 9 depicts the final reward for each rollout ζ_i.

$$r(\zeta_i) = s_i + \alpha \, \exp\big(\beta(s_i - s^{GT})\big) + r_{\text{fmt}}, \tag{9}$$

where $\alpha = 0.1$, $\beta = 1$ are reward scaling and sensitivity factors. Predictions adhering to the required output format receive a small bonus reward $r_{\text{fmt}} = 0.1$,

while unparseable outputs receive zero reward. The inclusion of ground-truth layout L^{GT} anchors the reward distribution to semantically valid references.

Policy optimization. Following (49), we optimize the layout policy at the token level, while maintaining groupwise credit assignment for stability across sequential layout steps. All notation for groups A_h and token sequences $\mathbf{y}_h$ follows their definitions introduced in Sec. 3.2. The model conditions its predictions on the multimodal context $\text{ctx}_g = (C^{GT}_{g-1}, \mathcal{P}^E_g, Z_g)$, where C^{GT}_{g-1} is the partially rendered canvas, $\mathcal{P}^E_g$ is the multimodal prompt describing all elements, and Z_g are the per-element latent embeddings obtained from the modality-aware encoder. The policy's conditional probability then factorizes autoregressively across tokens as given by Eq. 10

$$\log \pi_\theta(A_{g:G} \mid \text{ctx}_g) = \sum_{h=g}^{G} \sum_{t=1}^{T_h} \log \pi_\theta(y_{h,t} \mid \text{ctx}_g, y_{h,<t}), \tag{10}$$

where $y_{h,<t}$ denotes all previously generated tokens within group A_h. During reinforcement learning, all groups $\{A_g, A_{g+1}, \ldots, A_G\}$ originating from the same canvas state C^{GT}_{g-1} share the same scalar advantage, since the trajectory-level reward $R(\zeta)$ evaluates the complete rollout. For a sampled trajectory ζ, we define the advantage for each rollout as defined in Eq. 11.

$$A^{\text{adv}}_g = r(\zeta) - b_g, \tag{11}$$

where b_g is a running baseline computed across rollout batches to reduce variance. The overall policy gradient is computed by summing token-level gradients across all remaining groups that share this advantage (Eq. 12).

$$\nabla_\theta J(\theta) = \mathbb{E}_{\zeta \sim \pi_\theta} \left[\sum_{g=1}^{G} A^{\text{adv}}_g \sum_{h=g}^{G} \sum_{t=1}^{T_h} \nabla_\theta \log \pi_\theta(y_{h,t} \mid \text{ctx}_g,\, y_{h,<t}) \right]. \tag{12}$$

This formulation ensures that all future group predictions conditioned on a given canvas state receive the same feedback signal, encouraging consistent compositional reasoning and stable long-horizon credit assignment.

4 Experimental Details

Datasets and Evaluation: We evaluate *i-Design* and baselines on two benchmark graphic layout datasets: **Crello** (44) and **WebUI** (42). The Crello dataset contains 23,302 professionally designed templates across diverse formats such as posters and social media posts, posing challenges of high stylistic diversity, variable aspect ratios, and complex compositions. We adopt the standard split of 19,479 training, 1,852 validation, and 1,971 test samples. In contrast, the WebUI dataset consists of approximately 70k web page layouts collected from real-world websites, including both visual screenshots and structural metadata describing UI components. Layout quality is assessed using *Mean IoU* (mIoU),

Method	Mean IoU (↑)	Win Rate (%, ↑)	
		GPT-4o	o3
FlexDM (20)	12.71	–	–
LayoutDETR (46)	15.04	1.71	0.62
LACE (8)	23.18	0.85	0.09
PosterLLaVA (45)	25.18	2.19	1.01
LayoutNUWA (39)	25.74	4.18	1.93
LaDeCo (31)	35.25	14.42	8.73
AesthetiQ-8B (35)	42.83	15.74	8.04
i-Design–1B	23.83	2.85	1.13
i-Design–2B	30.96	6.74	1.98
i-Design–4B	43.47	17.59	10.37
i-Design–8B	**48.27**	**27.64**	**19.03**

table Layout generation performance on the **Crello** dataset.

Method	Mean IoU (↑)	Win Rate (%, ↑)
		GPT-4o
Desigen (42)	15.36	4.81
LayoutDETR (46)	16.11	5.27
LACE (8)	17.88	5.27
PosterLLaVA (45)	30.19	14.73
LayoutNUWA (39)	32.16	15.28
LaDeCo (31)	41.16	19.49
AesthetiQ-8B (35)	48.29	24.48
i-Design–1B	39.24	18.03
i-Design–2B	43.18	21.40
i-Design–4B	47.86	24.35
i-Design–8B	53.71	31.29

table Layout generation performance on the **WebUI** dataset.

which measures spatial alignment between predicted and ground-truth boxes, and *Win Rate* (35), which captures aesthetic preference via pairwise comparisons by GPT-4o and o3.

Implementation Details: Our layout generation model $\mathcal{M}_{\text{layout}}$ builds on the InternVL-3 8B backbone (52), selected for its strong multimodal grounding and visual–textual reasoning. We first pretrain on 148k graphic layouts (following (35)) to initialize spatial reasoning and compositional priors, then train in two stages:(1) **Progressive Imitation Learning (PIL)** on 75% of the Crello dataset to learn partial layout completion and placement consistency, and (2) **Progressive Aesthetic Policy Optimization (PAPO)** on the remaining 25% for one epoch, aligning the policy with aesthetic preferences via online reinforcement learning. Pretraining and fine-tuning run for 2 and 5 epochs, respectively, and PAPO for one epoch with groupwise rollouts and online reward feedback. We set α=0.1, β=1, γ=0.85, and rollouts m=4. All experiments use 8× A100 GPUs (80GB each) with an effective batch size of 128. We discretize element positions into 224 coordinate tokens (35), and render layouts using Skia.

4.1 Results and Analysis

Comparison with Baselines and Model Scaling: Table 4.1 and 4.1 report results on Crello and WebUI benchmarks respectively, comparing *i-Design* with recent state-of-the-art models. On Crello, *i-Design–8B* achieves the best overall performance with a Mean IoU of 48.27%, significantly outperforming **AesthetiQ–8B** (42.83%) and **LayoutNUWA** (25.74%). In terms of aesthetic preference, it attains the highest win rates of 27.64% and 19.03% as judged by GPT-4o and O3, respectively, reflecting superior visual balance, spatial harmony, and compositional coherence. These results confirm that Progressive Aesthetic Policy Optimization (PAPO) effectively aligns spatial reasoning with learned aesthetic preferences, improving both geometric fidelity and perceptual appeal. Unlike single-pass generators such as **FlexDM** (20), **LACE** (8), and **PosterLLaVA** (45), which often produce visually uneven layouts, *i-Design* incrementally predicts element placements through visual feedback from the evolving

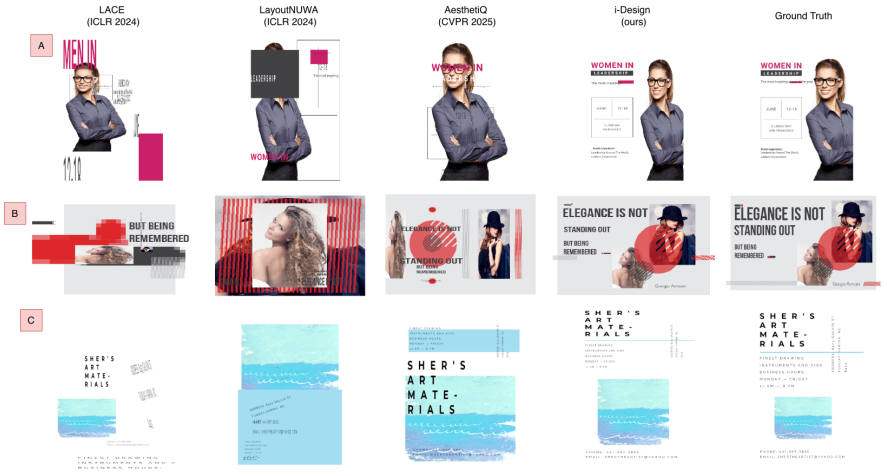

Fig. 3. Across all examples, i-Design produces layouts that exhibit superior compositional coherence and balance. In row (a), it effectively maintains spatial hierarchy and visual alignment between textual and visual elements, resulting in a well-organized poster closely matching the ground truth. In row (b), i-Design demonstrates consistent alignment and proportion across elements, avoiding clutter while maintaining semantic emphasis. Row (c) highlights the model's ability to preserve structure and alignment even in sparse compositions, achieving a balanced arrangement between text and imagery. Overall, i-Design generates layouts that are visually structured, semantically consistent, and aesthetically closer to human-designs compared to prior methods.

canvas. Furthermore, model scaling yields consistent improvements across both metrics. As shown in Table 4.1 and 4.1, scaling the model from 1B to 8B consistently improves Mean IoU and win rates, indicating stronger compositional reasoning and aesthetic understanding.

Qualitative Results: Figure 3 shows a qualitative comparison of *i-Design* with recent methods—LACE, LayoutNUWA, and AesthetiQ. In row (a), *i-Design* preserves visual hierarchy by placing key elements in salient regions, maintaining balance and clarity. Row (b) highlights its precise alignment and consistent spacing across elements, producing coherent, overlap-free layouts. In row (c), *i-Design* demonstrates strong compositional reasoning in sparse settings, achieving aesthetically harmonious and semantically consistent results. Overall, *i-Design* delivers visually superior and contextually aligned layouts, validating the benefits of its progressive generation and aesthetic optimization.

Effect of Training Stages: Across model scales (1B–8B), performance consistently improves as training progresses from pretraining to PIL and finally PAPO (Fig. 4). Mean IoU rises from 20.3% to 41.1% (1B–8B) with pretraining only, to 23.83% to 48.27% when all stages are combined. Similarly, GPT-4o Win Rate improves from 2.85% to 27.64%, and o3 Win Rate from 1.13% to 19.03%. Intermediate configurations (PIL-only, PIL+PAPO-only) outperform pretraining alone but remain below the full pipeline, indicating that pretraining provides essen-

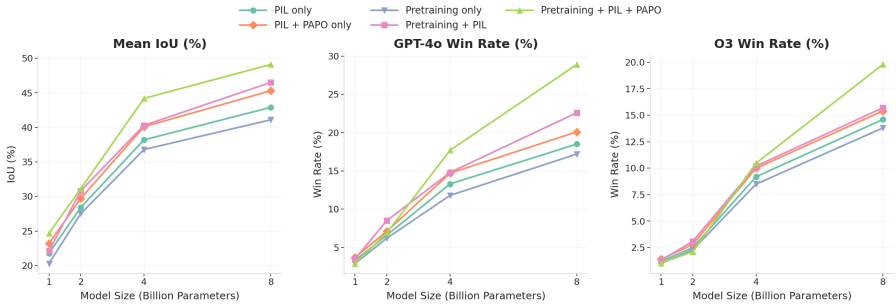

Fig. 4. Ablation of training stages across model scales (1B–8B) on the Crello benchmark. Left: Mean IoU shows progressive geometric improvements with iterative training. Middle and Right: GPT-4o and O3 Win Rates highlight stronger aesthetic alignment achieved through PIL and PAPO. Models trained with **Pretraining + PIL + PAPO** attain the highest performance, confirming the complementary roles of geometric grounding, iterative prediction, and aesthetic optimization in layout generation.

Training Strategy	**Mean IoU (↑)**	**Win Rate (%, ↑)**	
		GPT-4o	**O3**
i-Design-8B + PIL	44.39	20.41	15.28
i-Design–8B + PIL + AAPA (35))	46.62	25.60	16.79
i-Design–8B + PIL + PAPO (w/o GT)	45.19	23.04	15.39
i-Design–8B + PIL + PAPO (w GT)	**48.27**	**27.64**	**19.03**

Table 1. Comparison of reinforcement learning strategies and the effect of ground-truth anchoring for *i-Design–8B*. AAPA improves over supervised fine-tuning, while PAPO—with ground-truth references—achieves the highest geometric and aesthetic alignment across both GPT-4o and O3 evaluations.

tial grounding, while progressive imitation and aesthetic optimization contribute complementary gains in both spatial and aesthetic alignment across model size and scale.

Effect of Reinforcement Learning Strategy and Ground-Truth Anchoring: Table 1 compares different reinforcement learning strategies for *i-Design–8B*. Without RL training, the model achieves a Mean IoU of 44.39 and Win Rates of 20.41% (GPT-4o) and 15.28% (O3). Introducing Aesthetic-Aware Policy Alignment (AAPA) (35) enhances both spatial accuracy and aesthetic preference, improving Mean IoU to 46.62 and Win Rates to 25.60% and 16.79%. Our proposed Progressive Aesthetic Policy Optimization (PAPO) further advances these results, achieving the best Mean IoU of 48.27 and Win Rates of 27.64% (GPT-4o) and 19.03% (O3). Further, to assess the role of ground-truth anchoring, we ablate its inclusion in the preference graph. Removing the ground-truth reference causes a drop of 3.1% mIoU and over 5% in GPT-4o Win Rate, confirming that ground-truth layouts act as stabilizing anchors during reward

aggregation, preventing drift toward inconsistent or unstable aesthetic preferences. Together, these findings validate that PAPO with ground-truth anchoring provides the most stable and aesthetically aligned optimization.

Effect of group size on performance: Table 2 examines the influence of the group size k—the number of elements placed per iteration—on overall layout quality. When $k = N$ (single-pass generation), performance drops to a Mean IoU of 40.28 and win rates of 17.64% and 14.59%, revealing the limitations of one-shot prediction. However, this still surpasses AesthetiQ in both GPT-4o and O3 win rates, indicating that our progressive training improves even the model's single-shot generation ability. In contrast, both $k = 1$ and $k = \sqrt{N}$ achieve markedly higher scores, with $k = \sqrt{N}$ yielding the best balance between spatial fidelity (48.27 mIoU) and aesthetic alignment (27.64% / 19.03%). The minimal difference between $k = 1$ and $k = \sqrt{N}$ suggests that fine-grained iterative refinement provides diminishing returns beyond a moderate grouping size, while still confirming that progressive placement is critical for producing coherent and visually pleasing layouts.

Group Size (k)	**Mean IoU (↑)**	**Win Rate (%, ↑)**	
		GPT-4o	**O3**
$k = N$ (single pass)	40.28	17.64	13.59
$k = \sqrt{N}$ (default)	48.27	27.64	19.03
$k = 1$	47.72	27.13	18.31

Table 2. Effect of varying the group size k on layout quality. Smaller and intermediate group sizes ($k = 1$ and $k = \sqrt{N}$) achieve comparable performance, while large single-pass generation ($k = N$) degrades geometric fidelity and aesthetic alignment.

Effect of grouping strategy. We examine how the strategy used to partition elements into placement groups influences layout quality. Our default approach groups elements deterministically according to their z-order in the ground-truth layout. As shown in Table 3, this strategy achieves the best geometric fidelity (48.27 mIoU) while maintaining strong aesthetic alignment. In contrast, random grouping that mixes unrelated elements significantly degrades performance (43.27 mIoU), producing incoherent intermediate canvases that hinder progressive reasoning. We further evaluate semantic grouping using GPT-4o to cluster conceptually related elements. Semantic grouping yields slightly higher aesthetic win rates, likely because related elements (e.g., titles, images, and captions) are positioned jointly, improving local compositional coherence during intermediate placement steps. However, this requires an additional LLM call during inference to identify semantic clusters, introducing extra computational overhead. Despite not using explicit semantic grouping, deterministic z-order grouping achieves

Grouping Strategy	**Mean IoU (↑)**	**Win Rate (%, ↑)**	
		GPT-4o	**O3**
Z-order (default)	**48.27**	27.64	19.03
Random ($\sqrt{N}$ elems)	43.27	15.04	6.11
Semantic (GPT-4o)	46.21	**29.23**	**19.91**

Table 3. Effect of grouping strategy on layout quality.

competitive performance while avoiding this additional cost. Overall, these results indicate that i-Design is robust to semantically meaningful grouping strategies but sensitive to incoherent group compositions, highlighting the importance of structured intermediate states during progressive layout generation.

User Study: We conduct a user study to further evaluate the perceptual quality of generated layouts. Participants compared layouts generated by *i-Design* and other baselines. As shown in Fig. 5, *i-Design* achieved the highest preference rate of 52.9%, while AesthetiQ obtained 40.7%. Methods such as LACE and LayoutNUWA were rarely favored (2.8% and 3.6% respectively). Participants consistently preferred *i-Design*'s outputs for their superior balance, clear hierarchy, and harmonious composition, confirming that iterative generation with aesthetic feedback better reflects human design preferences.

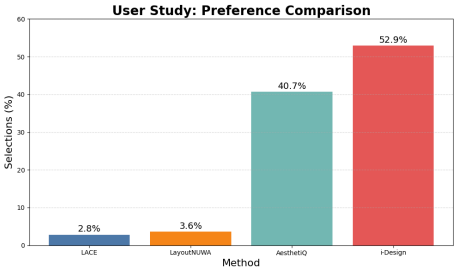

Fig. 5. User preference study comparing *i-Design* with baselines LACE, LayoutNUWA, and AesthetiQ. Participants selected the best layout among four methods. *i-Design* was preferred in **52.9%** of cases, surpassing prior SoTA **AesthetiQ** (40.7%).

5 Conclusion

We presented *i-Design*, a unified framework that redefines layout generation as a progressive, feedback-driven process combining geometric reasoning and aesthetic alignment. Through Progressive Imitation Learning (PIL) and Progressive Aesthetic Policy Optimization (PAPO), the model learns to compose layouts iteratively, guided by both spatial context and aesthetic feedback. Experiments on Crello and WebUI benchmarks establish new state-of-the-art results, demonstrating substantial gains in spatial coherence, visual balance, and human preference alignment. Overall, *i-Design* represents a step toward human-centric generative design systems that reason, adapt using visual context, and design aesthetically.

Bibliography

[1] Arroyo, D.M., Postels, J., Tombari, F.: Variational transformer networks for layout generation. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 13637–13647 (2021), https://api.semanticscholar.org/CorpusID:233033690

[2] Austin, J., Johnson, D.D., Ho, J., Tarlow, D., van den Berg, R.: Structured denoising diffusion models in discrete state-spaces. In: Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems (2021), https://openreview.net/forum?id=h7-XixPCAL

[3] Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. NIPS '20, Curran Associates Inc., Red Hook, NY, USA (2020)

[4] Cao, Y., Ma, Y., Zhou, M., Liu, C., Xie, H., Ge, T., Jiang, Y.: Geometry aligned variational transformer for image-conditioned layout generation. In: Proceedings of the 30th ACM International Conference on Multimedia. p. 1561–1571. MM '22, Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3503161.3548332, https://doi.org/10.1145/3503161.3548332

[5] Chai, S., Zhuang, L., Yan, F.: Layoutdm: Transformer-based diffusion model for layout generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 18349–18358 (June 2023)

[6] 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 28694–28704 (June 2025)

[7] Chen, J., Zhang, R., Zhou, Y., Chen, C.: Towards aligned layout generation via diffusion model with aesthetic constraints. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=kJ0qp9Xdsh

[8] Chen, J., Zhang, R., Zhou, Y., Jain, R., Xu, Z., Rossi, R., Chen, C.: Towards aligned layout generation via diffusion model with aesthetic constraints. In: Proceedings of the International Conference on Learning Representations (ICLR) (2024)

[9] Cheng, C.Y., Huang, F., Li, G., Li, Y.: Play: parametrically conditioned layout generation using latent diffusion. In: Proceedings of the 40th International Conference on Machine Learning. ICML'23, JMLR.org (2023)

[10] Deka, B., Huang, Z., Franzen, C., Hibschman, J., Afergan, D., Li, Y., Nichols, J., Kumar, R.: Rico: A mobile app dataset for building data-driven design applications. In: Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology. p. 845–854. UIST '17, Association for Computing Machinery, New York, NY, USA (2017). https://doi.org/10.1145/3126594.3126651, https://doi.org/10.1145/3126594.3126651

[11] Fu, T.J., Wang, W.Y., McDuff, D.J., Song, Y.: Doc2ppt: Automatic presentation slides generation from scientific documents. In: AAAI Conference on Artificial Intelligence (2021), https://api.semanticscholar.org/CorpusID:231719374

[12] Gopal, V., Jain, R., Mathur, A., SR, N., Patnaik, S., Yarram, S., Hemani, M., Krishnamurthy, B., Sarkar, M.: Designsense: A human preference dataset and reward modeling framework for graphic layout generation (2026), https://arxiv.org/abs/2602.23438

[13] Goyal, S., Mahajan, A., Mishra, S., Udhayanan, P., Shukla, T., Joseph, K., Srinivasan, B.V.: Design-o-meter: Towards evaluating and refining graphic designs. In: Proceedings of the Winter Conference on Applications of Computer Vision (WACV). pp. 5676–5686 (February 2025)

[14] Gu, S., Chen, D., Bao, J., Wen, F., Zhang, B., Chen, D., Yuan, L., Guo, B.: Vector quantized diffusion model for text-to-image synthesis. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10686–10696 (2022). https://doi.org/10.1109/CVPR52688.2022.01043

[15] Gupta, K., Lazarow, J., Achille, A., Davis, L.S., Mahadevan, V., Shrivastava, A.: Layouttransformer: Layout generation and completion with self-attention. 2021 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 984–994 (2020), https://api.semanticscholar.org/CorpusID:238227253

[16] He, J., Wang, Y., Wang, L., Lu, H., Li, C., Chen, H., Lan, J.P., He, J.Y., Luo, B., Geng, Y.: Gldesigner: Leveraging multi-modal llms as designer for enhanced aesthetic text glyph layouts. In: Proceedings of the 33rd ACM International Conference on Multimedia. p. 9996–10005. MM '25, Association for Computing Machinery, New York, NY, USA (2025). https://doi.org/10.1145/3746027.3755289, https://doi.org/10.1145/3746027.3755289

[17] Hsu, H., Peng, Y.: PosterO: Structuring Layout Trees to Enable Language Models in Generalized Content-Aware Layout Generation . In: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8117–8127. IEEE Computer Society, Los Alamitos, CA, USA (Jun 2025). https://doi.org/10.1109/CVPR52734.2025.00760, https://doi.ieeecomputersociety.org/10.1109/CVPR52734.2025.00760

[18] Huang, D., Guo, J., Sun, S., Tian, H., Lin, J., Hu, Z., Lin, C.Y., Lou, J.G., Zhang, D.: A survey for graphic design intelligence (2023), https://arxiv.org/abs/2309.01371

[19] Hui, M., Zhang, Z., Zhang, X., Xie, W., Wang, Y., Lu, Y.: Unifying layout generation with a decoupled diffusion model. 2023 IEEE/CVF Confer-

ence on Computer Vision and Pattern Recognition (CVPR) pp. 1942–1951 (2023), https://api.semanticscholar.org/CorpusID:257427355

[20] Inoue, N., Kikuchi, K., Simo-Serra, E., Otani, M., Yamaguchi, K.: Towards Flexible Multi-modal Document Models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14287–14296 (2023)

[21] Iwai, S., Osanai, A., Kitada, S., Omachi, S.: Layout-corrector: Alleviating layout sticking phenomenon in discrete diffusion models. In: Proceedings of the European Conference on Computer Vision (ECCV) (2024)

[22] Jiang, Z., Sun, S., Zhu, J., Lou, J.G., Zhang, D.: Coarse-to-fine generative modeling for graphic layouts. Proceedings of the AAAI Conference on Artificial Intelligence **36**(1), 1096–1103 (Jun 2022). https://doi.org/10.1609/aaai.v36i1.19994, https://ojs.aaai.org/index.php/AAAI/article/view/19994

[23] Kikuchi, K., Honda, U., Inoue, N., Otani, M., Simo-Serra, E., Yamaguchi, K.: Multimodal markup document models for graphic design completion. In: Proceedings of the 33rd ACM International Conference on Multimedia. p. 11022–11031. MM '25, Association for Computing Machinery, New York, NY, USA (2025). https://doi.org/10.1145/3746027.3755420, https://doi.org/10.1145/3746027.3755420

[24] Kikuchi, K., Simo-Serra, E., Otani, M., Yamaguchi, K.: Constrained graphic layout generation via latent optimization. In: Proceedings of the 29th ACM International Conference on Multimedia. p. 88–96. MM '21, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3474085.3475497, https://doi.org/10.1145/3474085.3475497

[25] Kong, X., Jiang, L., Chang, H., Zhang, H., Hao, Y., Gong, H., Essa, I.: Blt: Bidirectional layout transformer for controllable layout generation. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision – ECCV 2022. pp. 474–490. Springer Nature Switzerland, Cham (2022)

[26] Lee, H.Y., Jiang, L., Essa, I., Le, P.B., Gong, H., Yang, M.H., Yang, W.: Neural design network: Graphic layout generation with constraints. In: Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III. p. 491–506. Springer-Verlag, Berlin, Heidelberg (2020). https://doi.org/10.1007/978-3-030-58580-8_29, https://doi.org/10.1007/978-3-030-58580-8_29

[27] Li, J., Yang, J., Zhang, J., Liu, C., Wang, C., Xu, T.: Attribute-conditioned layout gan for automatic graphic design. IEEE Transactions on Visualization and Computer Graphics **27**(10), 4039–4048 (Oct 2021). https://doi.org/10.1109/TVCG.2020.2999335, https://doi.org/10.1109/TVCG.2020.2999335

[28] Li, Y., Chen, J., Bai, Y., Cheng, J., Lei, J.: Design element aware poster layout generation. In: Proceedings of the 33rd ACM International Conference on Information and Knowledge Management. p. 1296–1305. CIKM '24, Association for Computing Machinery, New York, NY, USA

(2024). https://doi.org/10.1145/3627673.3679557, https://doi.org/10.1145/3627673.3679557

[29] Li, Y., Chen, Y., Liu, G., Fei, Y., Bai, Q., Wu, J., Hongfa, W., Chu, R., Yang, Y.: Layoutdit: Exploring content-graphic balance in layout generation with diffusion transformer. arXiv preprint arXiv:2407.15233 (2024)

[30] Lin, J., Guo, J., Sun, S., Yang, Z.J., Lou, J.G., Zhang, D.: Layoutprompter: awaken the design ability of large language models. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. NIPS '23, Curran Associates Inc., Red Hook, NY, USA (2024)

[31] 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: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8128–8137. IEEE Computer Society, Los Alamitos, CA, USA (Jun 2025). https://doi.org/10.1109/CVPR52734.2025.00761, https://doi.ieeecomputersociety.org/10.1109/CVPR52734.2025.00761

[32] Melendez Abarca, P., Havas, C.: Spatial information integration in small language models for document layout generation and classification. In: Proceedings of the 40th ACM/SIGAPP Symposium on Applied Computing. p. 1164–1171. SAC '25, Association for Computing Machinery, New York, NY, USA (2025). https://doi.org/10.1145/3672608.3707832, https://doi.org/10.1145/3672608.3707832

[33] O'Donovan, P., Agarwala, A., Hertzmann, A.: Designscape: Design with interactive layout suggestions. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. p. 1221–1224. CHI '15, Association for Computing Machinery, New York, NY, USA (2015). https://doi.org/10.1145/2702123.2702149, https://doi.org/10.1145/2702123.2702149

[34] O'Donovan, P., Agarwala, A., Hertzmann, A.: Learning layouts for single-pagegraphic designs. IEEE Transactions on Visualization and Computer Graphics **20**(8), 1200–1213 (2014). https://doi.org/10.1109/TVCG.2014.48

[35] 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: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 23701–23711. IEEE Computer Society, Los Alamitos, CA, USA (Jun 2025). https://doi.org/10.1109/CVPR52734.2025.02207, https://doi.ieeecomputersociety.org/10.1109/CVPR52734.2025.02207

[36] Rahman, S., Sermuga Pandian, V.P., Jarke, M.: Ruite: Refining ui layout aesthetics using transformer encoder. In: Companion Proceedings of the 26th International Conference on Intelligent User Interfaces. p. 81–83. IUI '21 Companion, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3397482.3450716, https://doi.org/10.1145/3397482.3450716

[37] Rozière, B., Gehring, J., Gloeckle, F., Sootla, S., Gat, I., Tan, X., Adi, Y., Liu, J., Remez, T., Rapin, J., Kozhevnikov, A., Evtimov, I., Bitton, J.,

Bhatt, M.P., Ferrer, C.C., Grattafiori, A., Xiong, W., D'efossez, A., Copet, J., Azhar, F., Touvron, H., Martin, L., Usunier, N., Scialom, T., Synnaeve, G.: Code llama: Open foundation models for code. ArXiv **abs/2308.12950** (2023), https://api.semanticscholar.org/CorpusID:261100919

[38] Seol, J., Kim, S., Yoo, J.: Posterllama: Bridging design ability of language model to content-aware layout generation. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds.) Computer Vision – ECCV 2024. pp. 451–468. Springer Nature Switzerland, Cham (2025)

[39] Tang, Z., Wu, C., Li, J., Duan, N.: LayoutNUWA: Revealing the hidden layout expertise of large language models. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=qCUWVT0Ayy

[40] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G.: Llama: Open and efficient foundation language models. ArXiv **abs/2302.13971** (2023), https://api.semanticscholar.org/CorpusID:257219404

[41] Wei, C., Mangalam, K., Huang, P.Y.B., Li, Y., Fan, H., Xu, H., Wang, H., Xie, C., Yuille, A.L., Feichtenhofer, C.: Diffusion models as masked autoencoders. 2023 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 16238–16248 (2023), https://api.semanticscholar.org/CorpusID:257985028

[42] Weng, H., Huang, D., Qiao, Y., Hu, Z., Lin, C.Y., Zhang, T., Chen, C.L.P.: Desigen: A pipeline for controllable design template generation. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 12721–12732 (2024), https://api.semanticscholar.org/CorpusID:268385148

[43] Xu, C., Han, K., Xu, W.: Image-aware layout generation with user constraints for poster design. The Visual Computer **41**(6), 4239–4252 (2025). https://doi.org/10.1007/s00371-024-03657-z, https://doi.org/10.1007/s00371-024-03657-z

[44] Yamaguchi, K.: Canvasvae: Learning to generate vector graphic documents. 2021 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 5461–5469 (2021), https://api.semanticscholar.org/CorpusID:236881543

[45] Yang, T., Luo, Y., Qi, Z., Wu, Y., Shan, Y., Chen, C.W.: Posterllava: Constructing a unified multi-modal layout generator with llm (2024), https://arxiv.org/abs/2406.02884

[46] Yu, N., Chen, C.C., Chen, Z., Meng, R., Wu, G., Josel, P., Niebles, J.C., Xiong, C., Xu, R.: Layoutdetr: Detection transformer is anbsp;good multimodal layout designer. In: Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part XX. p. 169–187. Springer-Verlag, Berlin, Heidelberg (2024). https://doi.org/10.1007/978-3-031-72661-3_10, https://doi.org/10.1007/978-3-031-72661-3_10

[47] Zhang, J., Guo, J., Sun, S., Lou, J.G., Zhang, D.: Layoutdiffusion: Improving graphic layout generation by discrete diffusion probabilistic models. 2023 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 7192–7202 (2023), https://api.semanticscholar.org/CorpusID:257636725

[48] Zhang, P., Zhang, J., Cao, J., Li, H., Jin, L.: Smaller but better: Unifying layout generation with smaller large language models. International Journal of Computer Vision **133**(7), 3891–3917 (2025). https://doi.org/10.1007/s11263-025-02353-2, https://doi.org/10.1007/s11263-025-02353-2

[49] Zheng, C., Liu, S., Li, M., Chen, X.H., Yu, B., Gao, C., Dang, K., Liu, Y., Men, R., Yang, A., Zhou, J., Lin, J.: Group sequence policy optimization (2025), https://arxiv.org/abs/2507.18071

[50] Zheng, X., Qiao, X., Cao, Y., Lau, R.W.H.: Content-aware generative modeling of graphic design layouts. ACM Trans. Graph. **38**(4) (Jul 2019). https://doi.org/10.1145/3306346.3322971, https://doi.org/10.1145/3306346.3322971

[51] Zhou, M., Xu, C., Ma, Y., Ge, T., Jiang, Y., Xu, W.: Composition-aware graphic layout gan for visual-textual presentation designs. ArXiv **abs/2205.00303** (2022), https://api.semanticscholar.org/CorpusID:248496330

[52] Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., Gao, Z., Cui, E., Wang, X., Cao, Y., Liu, Y., Wei, X., Zhang, H., Wang, H., Xu, W., Li, H., Wang, J., Deng, N., Li, S., He, Y., Jiang, T., Luo, J., Wang, Y., He, C., Shi, B., Zhang, X., Shao, W., He, J., Xiong, Y., Qu, W., Sun, P., Jiao, P., Lv, H., Wu, L., Zhang, K., Deng, H., Ge, J., Chen, K., Wang, L., Dou, M., Lu, L., Zhu, X., Lu, T., Lin, D., Qiao, Y., Dai, J., Wang, W.: Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models (2025), https://arxiv.org/abs/2504.10479