From Pixels to Policies: Reinforcing Spatial Reasoning in Language Models for Content-Aware Layout Design

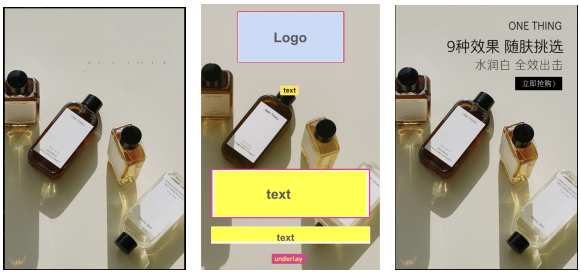
Sha Li[1], Stefano Petrangeli[2], Yu Shen[2], Xiang Chen[2]
[1]Virginia Tech
[2]Adobe Research

Abstract

We introduce *LaySPA*, a reinforcement learning framework that equips large language models (LLMs) with explicit and interpretable spatial reasoning for content-aware graphic layout design. *LaySPA* addresses two key challenges: LLMs' limited spatial reasoning and the lack of transparency in design decision making. Instead of operating at the pixel level, we reformulate layout design as a policy learning problem over a structured textual spatial environment that explicitly encodes canvas geometry, element attributes, and inter-element relationships. *LaySPA* produces dual-level outputs comprising interpretable reasoning traces and structured layout specifications, enabling transparent and controllable design decision making. Layout design policy is optimized via a multi-objective spatial critique that decomposes layout quality into geometric validity, relational coherence, and aesthetic consistency, and is trained using relative group optimization to stabilize learning in open-ended design spaces. Experiments demonstrate that *LaySPA* improves structural validity and visual quality, outperforming larger proprietary LLMs and achieving performance comparable to specialized state-of-the-art layout generators while requiring fewer annotated samples.

1 Introduction

Large Language Models (LLMs) have demonstrated strong capabilities in structured reasoning (Zhang et al., 2025), planning (Ferrag et al., 2025) and decision making (Zhai et al., 2025). These advances extend beyond natural language understanding and generation to domains requiring multi-step reasoning (Plaat et al., 2025), interaction (Chen et al., 2025), and symbolic manipulation, positioning LLMs as a promising substrate for design-oriented tasks that demand explicit relational and structural reasoning (Wu et al., 2025b). Within this context, *automatic content-aware graphic layout design* (Zheng et al., 2019) emerges as a compelling testbed for LLM-centric spatial reasoning, with applications spanning advertising (Hsu et al., 2023), web interfaces (Wang et al., 2025), and poster design (Chai et al., 2023b), all of which require coordinated placement of multiple elements under intertwined semantic, geometric, and aesthetic constraints.



(a) Input canvas (b) GPT-5 output (c) Human-designed

Figure 1: Comparison between a human-designed poster and a GPT-5–generated layout, illustrating common spatial reasoning failures in LLM-based layout generation.

Prior approaches typically formulate layout design as coordinate regression or image-conditioned synthesis, using diffusion models (Chai et al., 2023a; Zhang et al., 2023) or autoregressive generators (Horita et al., 2024). While effective at producing visually plausible layouts, these methods rely heavily on large annotated datasets and offer limited interpretability: *spatial decisions are entangled with perception in dense latent representations, often prioritizing appearance over structural principles*. In contrast, layout design requires explicit global spatial coherence, multi-object relational reasoning and semantic–functional coupling (*e.g.*, hierarchy, alignment, decoration), and generalizability to novel element configurations or constraints, which distinguishes it from pixel-level im-

age generation and expose the limitations of purely perceptual modeling.

LLMs provide a complementary perspective as they encode prior knowledge of design principles (Feng et al., 2023), support symbolic reasoning (Wu et al., 2023) and produce interpretable outputs. Early attempts apply LLMs to layout generation via structured serializations using in-context learning (Horita et al., 2024), chain-of-thought reasoning (Wei et al., 2022), or prompt engineering (Lin et al., 2023). However, these approaches treat spatial reasoning as an *emergent byproduct* instead of an explicit modeling objective: they neither represent spatial constraints explicitly nor enforce geometric or relational consistency at the element or canvas level. Multimodal extensions using visual–language models (Ranasinghe et al., 2024; Cheng et al., 2024) improve visual grounding but still struggle with structural consistency and geometric plausibility. As illustrated in Fig. 1, GPT-5 exhibits systematic layout errors including misalignment, inconsistent sizing and violated semantic relations (*e.g.*, underlay–text), underscoring that effective layout design requires explicit reasoning over spatial constraints and relational geometry, not merely visual grounding. Moreover, multimodal architectures often rely on inductive biases to implicitly encode geometry. A text-only formulation removes these shortcuts, making spatial reasoning an explicit and learnable capability.

These observations motivate a new research question: ***How can an LLM learn robust and generalizable spatial reasoning for layout design from textual serialization without pixel-level supervision?*** Addressing this question requires moving beyond prompt-based generation and multimodal shortcuts. However, LLMs are known to struggle with spatial reasoning, particularly in tasks involving relative geometry and multi-object relationships (Bang et al., 2023; Zha et al., 2025). In layout design, these limitations are exacerbated by (1) *the lack of intrinsic geometric reasoning capability* for alignment, containment and hierarchy, and (2) *open-ended design spaces with sparse supervision*, where multiple layouts may be equally valid and paired annotations are scarce.

To address these challenges, we propose *LaySPA*, a framework that equips LLMs with explicit and task-grounded spatial reasoning for content-aware layout design. We formulate layout generation as a *policy learning* problem in a spatially grounded environment, where an LLM optimizes design decisions through multi-objective spatial critique. By decoupling spatial reasoning from pixel-level generation, *LaySPA* enables interpretable decision making, geometry-driven generalization, and direct optimization of spatial competence via reward design.

Our contributions are threefold:

- We reformulate content-aware layout design as a policy learning problem, enabling explicit and interpretable spatial decision-making beyond pixel-level generation.
- We reinforce LLMs' spatial reasoning over serialized canvases via multi-objective critiques that encode geometric, relational, and aesthetic constraints as learnable objectives.
- We empirically demonstrate that *LaySPA* learns robust spatial reasoning, outperforming larger proprietary LLMs and most visual-based layout generators, while elucidating the complementary strengths and limitations of language-only and vision-based approaches.

2 Method

2.1 Problem Formulation

We formulate content-aware layout design as a *policy learning* problem, where an LLM learns a design policy π_θ that performs explicit spatial reasoning over a structured textual canvas specification $\mathcal{S}$, generating layouts $\mathcal{L} = \pi_\theta(\cdot \mid \mathcal{S})$ that are structurally coherent and visually appealing, without pixel-level supervision and with spatial reasoning as the primary learning signal.

2.2 *LaySPA*

We propose *LaySPA* (Fig. 2), a reinforcement learning (§2.3) framework that grounds the canvas as a structured spatial environment (§2.2.1), produces interpretable layout actions (§2.2.2), and optimizes the LLM's design policy via a multi-objective critique (§2.2.3).

2.2.1 Spatially-Grounded Environment

LaySPA represents each canvas as a spatially grounded environment that encodes normalized coordinates, semantic elements (*e.g.*, text, logos, underlays), and salient regions to be avoided. This structured abstraction makes geometric and relational constraints explicit to the LLM. Formally, a layout $\mathcal{L}$ comprises a set of elements E and saliency regions S, where each item is defined by top-left coordinates (x_i, y_i), width and height (w_i, h_i), and category c_i. Conditioned on element

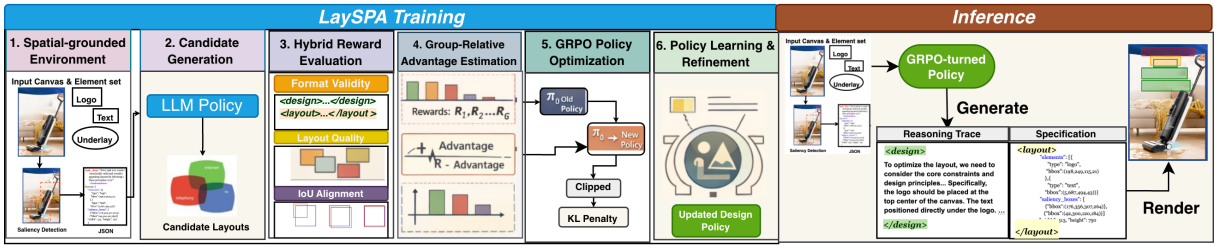


Figure 2: The framework and workflow of *LaySPA*.

semantics and the saliency context, the LLM infers the geometric attributes (x_i, y_i, w_i, h_i) for all elements. Layouts are serialized in JSON.

2.2.2 Dual-Level Design Artifacts

Given a canvas specification $\mathcal{S} = \{E, S\}$, the LLM policy π_θ generates layouts $\ell_{g=1}^{G}$, each comprising: (i) an explicit *design reasoning trace* that articulates the model's spatial reasoning and design logic, and (ii) a *structured layout specification* that encodes precise geometric configurations of all elements. These outputs are specialized in $\langle design \rangle$ and $\langle layout \rangle$ blocks, respectively. This dual-level representation decouples relational reasoning from geometric instantiation, enabling interpretability, error diagnosis, and reward computation for downstream evaluation and optimization.

2.2.3 Multi-Objective Spatial Critique

To guide policy learning, we introduce a multi-objective spatial critique that evaluates candidate layouts and assigns normalized scalar rewards capturing geometric validity, structural coherence and visual aesthetics. This structured feedback enables the policy to differentiate fine-grained spatial trade-offs and converge toward globally coherent and visually balanced layouts.

1) Format Correctness Reward R_{format} enforces strict compliance with the required *design–layout* output schema, ensuring that each output is structurally complete and parsable. Specifically, outputs must contain a $\langle design \rangle$ block that articulates design rationale and a $\langle layout \rangle$ block that specifies element geometries in valid JSON. The reward is assigned hierarchically: 0.1 if either block is missing, 0.2 if the JSON is unparsable, 0.5 if the JSON is valid but element types or counts mismatch the input, and 1.0 if both blocks are present, the JSON is valid, and all element types and counts match the specification. R_{format} mitigates hallucinated or malformed outputs and guarantees compatibility with downstream geometric evaluation.

2) Layout Quality Reward R_{quality} evaluates layouts based on geometric constraints and design principles using five normalized sub-metrics.

2.1) Inverse Collision Rate S_{icr} penalizes unintended overlaps between incompatible elements (*e.g.*, text, logos, embellishments and salient regions), while permitting intended overlaps between underlays and their associated text. Each element is represented as an axis-aligned bounding box $b_i = (x_i, y_i, w_i, h_i)$. For any pair (b_i, b_j), the pairwise inverse collision score is defined as:

$$S_{\text{icr}}(b_i, b_j) = 1 - \frac{\text{Area}(b_i \cap b_j)}{\text{Area}(b_i) + \text{Area}(b_j) - \text{Area}(b_i \cap b_j)} \tag{1}$$

where $\text{Area}(\cdot)$ denotes the area of a bounding box. Higher scores correspond to layouts with minimal unintended overlap.

2.2) Alignment Score S_{align} quantifies layout regularity by measuring (i) global centering relative to the canvas, and (ii) mutual alignment among elements. Well-aligned layouts exhibit small deviations from the canvas center and low dispersion of element centers. Consider a layout with N elements, the center of element e_i $\boldsymbol{\mu}_i = (x_i + \frac{w_i}{2},\ y_i + \frac{h_i}{2})$, and the canvas center $\boldsymbol{\mu}_{canvas} = (\frac{w_{canvas}}{2},\ \frac{h_{canvas}}{2})$. Then we measure:
(i) Element-to-canvas alignment: valuating global centering by the normalized average Euclidean distance between element and canvas center:

$$A_{\text{ele-canvas}} = 1 - \frac{1}{N}\sum_{i=1}^{N} \frac{\|\mu_i - \mu_{\text{canvas}}\|_2}{D_{\max}} \tag{2}$$

where $D_{\max} = \sqrt{w_c^2 + h_c^2}$ ensures scale invariance.
(ii) Element-to-element alignment: capturing mutual alignment by penalizing the variance of element centers along the horizontal and vertical axes:

$$A_{\text{ele-ele}} = 1 - \frac{1}{2}\Big(\text{Var}(\mu_{i,x}) + \text{Var}(\mu_{i,y})\Big) \tag{3}$$

where $\mathrm{Var}(\cdot)$ is computed across all element centers along each axis. Lower variance corresponds to higher alignment.
The final alignment score is computed as:

$$S_{\text{align}} = \alpha \cdot A_{\text{ele-canvas}} + (1-\alpha) \cdot A_{\text{ele-ele}} \quad (4)$$

where $\alpha \in [0,1]$ balances between global centering and inter-element alignment.

2.3) Distribution Score S_{dist} measures how evenly elements are arranged on the canvas by jointly evaluating spatial dispersion and area coverage. It combines (i) a normalized spread variance that encourages broad spatial distribution, and (ii) a grid coverage score that promotes utilization of the canvas while discouraging element clustering.

The normalized spread variance is defined as:

$$V_{\text{spread}} = \frac{1}{N}\sum_{i=1}^{N}\frac{\|\boldsymbol{\mu}_i - \bar{\boldsymbol{\mu}}\|_2^2}{D_{\max}^2} \quad (5)$$

where $\bar{\boldsymbol{\mu}} = (\bar{x}, \bar{y}) = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\mu}_i$ is the average center of all elements. Higher values indicate broader spatial dispersion.

Grid coverage measures area utilization by partitioning the canvas into a 3×3 grid. Let $G_{ij} \in \{0,1\}$ indicate whether grid cell (i,j) contains at least one element center. The coverage score is:

$$V_{\text{grid}} = \frac{1}{9}\sum_{i=1}^{3}\sum_{j=1}^{3} G_{ij} \quad (6)$$

The final distribution score is computed by:

$$S_{\text{dist}} = \frac{1}{2}\left(V_{\text{spread}} + V_{\text{grid}}\right) \quad (7)$$

2.4) Spacing Consistency Score S_{spacing} quantifies vertical rhythm by computing the normalized variance of distances between adjacent element centers. Elements are first sorted by vertical positions, then the distances between adjacent centers are computed. The variance of these spacings is normalized by the mean spacing and canvas height. Higher scores correspond to more even spacing. Formally, let $c_i^y = y_i + h_i/2$ denote the vertical center of element e_i. After sorting elements by c_i^y, the adjacent spacings are $s_i = c_{(i+1)}^y - c_{(i)}^y$. The spacing consistency is defined as:

$$S_{\text{spacing}} = 1 - \frac{\mathrm{Var}(s_1, \ldots, s_{N-1})}{\bar{s}^2} \quad (8)$$

where $\bar{s}$ is the average spacing. Larger values indicate more uniform vertical spacing.

2.5) Underlay-Text Constraint Score S_{underlay} enforces semantic consistency by aligning each underlay with exactly one corresponding text element. Layouts are penalized if multiple text elements overlap a single underlay or if the underlay's position and size deviate from its associated text. Exact alignment receives a score of 1, partial matches receives intermediate scores, and incorrect pairings are scored 0. This mechanism promotes readability and preserves intended functional relationships between decorative and content elements.

Finally, these sub-rewards are aggregated into a weighted layout quality score. Appendix. A Figure 6 illustrates how these scores guide the model in balancing structural feasibility and visual appeal in content-aware layouts.

3) IoU Matching Reward R_{IoU} provides external supervision by measuring the geometric agreement between generated layouts and human-designed references. For each element, IoU is computed as the ratio of the intersection to the union area of the predicted and ground-truth bounding boxes. Unlike intrinsic metrics such as collision or alignment, which assess internal structural consistency, R_{IoU} anchors learning to human layout exemplars that implicitly encode balance, proportion, and functional grouping. This signal guides the policy toward realistic spatial organization and visually coherent designs.

4) Final Hybrid Reward for each generated layout is a weighted combination of:

$$R(L_i) = \lambda_f R_{\text{format}} + \lambda_q R_{\text{quality}} + \lambda_u R_{\text{IoU}} \quad (9)$$

2.3 Training Procedure

We update the design policy using Group Relative Policy Optimization (GRPO) (Shao et al., 2024), which uses group-wise comparisons as relative baselines. At each step, the old policy $\pi_{\theta_{\text{old}}}$ samples a group of candidate layouts G, each scored $R(\ell_i)$ by the hybrid reward in Eq. 9. Relative advantages are computed by normalizing rewards within the group $A(\ell_i) = \frac{R(\ell_i) - mean_{\ell_i \in G} R(\ell_i)}{std_{\ell_i \in G} R(\ell_i)}$, providing stable learning signals without an explicit critic. The policy π_θ is then optimized using the clipped surrogate objective:

$$\mathcal{L}(\theta) = \mathbb{E}_{\ell_i \sim \pi_\theta}\Big[\min\Big(r_\theta(\ell_i)A(\ell_i), \text{clip}\big(r_\theta(\ell_i), 1-\epsilon, 1+\epsilon\big)A(\ell_i)\Big) - \beta\,\mathrm{KL}(\pi_\theta \,\|\, \pi_{\theta_{\text{ref}}})\Big] \tag{10}$$

where

$$r_\theta(\ell_i) = \frac{\pi_\theta(\ell_i)}{\pi_{\theta_{\text{old}}}(\ell_i)}.$$

This group-relative update prioritizes higher-reward layouts while stabilizing learning across diverse candidates.

3 Experiments

Our experiments address three questions: (Q1) Does explicit reward-driven policy optimization enable *LaySPA* to learn spatial reasoning beyond prompt-based LLM baselines? (Q2) How does *LaySPA* compare with autoregressive and multimodal layout generators in structural validity and visual quality? (Q3) What performance gap remains between text-based, policy-optimized layout design and specialized task-specific generators?

Datasets We evaluate on CGL (Zhou et al., 2022) and PKU-PosterLayout (Hsu et al., 2023). PKU contains text, logo, and underlay elements, while CGL additionally includes embellishments (More details are on Appendix B). We train on 3,000 randomly sampled layouts per dataset and evaluate on full 1000 CGL and 905 PKU test instances.

Experimental Setup *LaySPA* uses Qwen-2.5-7B-Instruct (Team, 2024) as the backbone, fine-tuned via LoRA (Hu et al., 2022) (rank=8 and lora_alpha=16) for 200 steps. For each input, we generate 4 candidate rollouts. The initial learning rate is set to 0.0005. Salient regions are detected using off-the-shelf models (Qin et al., 2019) and converted to bounding boxes. Element geometries are masked with a `[MASK]` token, prompting the LLM to predict their positions (x_i, y_i) and sizes (w_i, h_i).

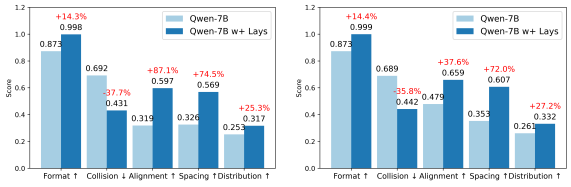
Baselines We evaluate *LaySPA* against: (i) autoregressive DS-GAN (Hsu et al., 2023), (ii) VLM-based PosterLlama (Seol et al., 2024), (iii) Qwen-2.5-7B-Instruct in a zero-shot setting, and (iv) proprietary GPT-4o (Hurst et al., 2024). (More details in Appendix.C)

Evaluation Metrics We evaluate *LaySPA* along two dimensions: (i) improvement over the base LLM, using structural metrics including collision, alignment, spacing, and distribution consistency to assess the effectiveness of reward-guided policy learning; and (ii) performance against SOTA layout methods, using graphic- and content-focused metrics (Hsu et al., 2023; Horita et al., 2024): Overlay (Ove↓) measures overlap among non-underlay elements, Underlay Effectiveness (Und ↑) assesses how well underlays support content, and Occlusion (Occ ↓) quantifies overlap with salient regions.

4 Results and Analysis

4.1 Quantitative Results

Learning Effectiveness Figure 3 compares the base Qwen model with its *LaySPA*-fine-tuned counterpart on the CGL and PKU benchmarks. *LaySPA* delivers consistent and substantial gains across all metrics, including alignment (+63%), spacing consistency (+73%), format correctness (+14%) and distribution consistency (+26%), together with a reduction in collision rate (−36%). These gains indicate that reward-driven optimization enables the model to internalize explicit geometric and relational constraints, yielding layouts that better satisfy structural requirements. The largest gains obtained in alignment and spacing suggest that *LaySPA* particularly strengthens global spatial organization and inter-element uniformity.



(a) CGL Dataset (b) PKU Dataset

Figure 3: Comparison of Qwen-7B w/wo *LaySPA* on CGL and PKU. Red highlights improvements.

Compared with SOTA Table 1 reports results on CGL and PKU by different methods. PosterLlama achieves the strongest overall performance, benefiting from its visual encoders and task-specific architectural priors tailored to layout generation. *LaySPA* ranks second, substantially improving over its base model, with pronounced reductions in overlap and occlusion and improved underlay effectiveness, indicating more accurate functional grouping and spatial allocation. Despite larger scale, GPT-4o

performs moderately, suggesting that model size alone does not guarantee fine-grained spatial reasoning in layout tasks. Notably, *LaySPA* attains competitive performance using only 3000 training samples per dataset, far fewer than vision-centric methods, demonstrating the strong spatial inductive bias and sample efficiency induced by reward-driven policy optimization.

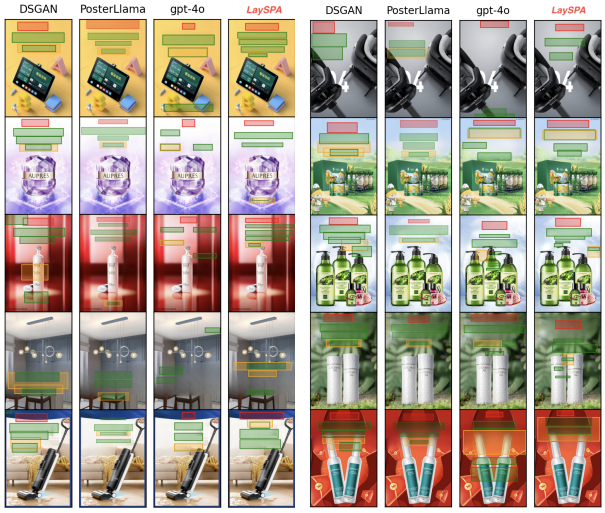
In addition, *LaySPA* enables efficient inference via a single-pass policy rollout. Unlike diffusion-based or iterative refinement pipelines, it generates complete layouts in a single forward pass without beam search or repeated optimization. Although no explicit multi-step refinement is performed at inference time, the policy implicitly encodes multi-step planning, enabling it to optimize global layout quality under multiple constraints. Finally, as a modular framework, *LaySPA* remains compatible with different saliency detection methods.

Table 1: Comparison of overlap (Ove↓), underlay effectiveness (Und↑) and occlusion (Occ↓) on CGL and PKU datasets. "Real Data" denotes the annotated layout. Best results are **bolded**; second-best are underlined.

Model	#Params	CGL Dataset			PKU Dataset		
		Ove ↓	Und ↑	Occ ↓	Ove ↓	Und ↑	Occ ↓
Real Data	-	0.0003	0.9926	0.1379	0.0013	0.9974	0.1828
DS-GAN	30M	0.0361	0.6309	0.1521	0.0336	0.7613	0.2574
PosterLLama	7B	**0.0024**	**0.9918**	**0.1476**	**0.0032**	**0.9998**	0.2087
GPT-4o	200B	0.0365	0.5873	0.1591	0.0371	0.6384	0.2743
Qwen-7B	7B	0.0474	0.5729	0.1615	0.0479	0.6059	0.2384
Qwen-7B+*LaySPA*	7B	0.0257	0.6989	0.1487	0.0260	0.7688	**0.2072**

4.2 Qualitative Results

Figure 4 presents qualitative comparisons across methods. DS-GAN achieves reasonable canvas coverage and consistent element sizing but often exhibits misalignment, unintended overlaps, and interference with salient regions, reflecting limited global spatial coordination. PosterLlama produces the most visually coherent layouts, benefiting from vision-specific encoders that impose strong structural priors and balanced compositions. GPT-4o underperforms particularly on CGL, with frequent misalignment and weak relational organization, revealing limitations in fine-grained spatial control despite its scale. In contrast, *LaySPA* consistently generates visually harmonious and well-proportioned layouts, demonstrating that explicitly learned spatial reasoning and policy optimization improves overall design quality.



(a) CGL (b) PKU

Figure 4: Layouts generated by different methods.

4.3 Ablation Study

To dissect the contributions of each component in the hybrid reward (Eq. 9), we vary $(\lambda_f, \lambda_q, \lambda_u)$ and evaluate *LaySPA* on the CGL benchmark. We consider four configurations: (1) Format-focused $(0.5, 0.4, 0.1)$ emphasizes schema compliance, (2) Quality-focused $(0.1, 0.8, 0.1)$ prioritizes geometric and visual design principles, (3) IoU-focused $(0.1, 0.1, 0.8)$ emphasizes alignment with reference layouts, and (4) Balanced-hybrid $(0.1, 0.45, 0.45)$ distributes weight across all objectives.

Figure 5 shows performance across collision, alignment, spacing consistency, and distribution metrics. Quality-focused training yields the best overall results, highlighting the importance of explicit geometric and relational objectives for coherent, functionally valid layouts. The balanced-hybrid setting offers a pragmatic trade-off, maintaining reasonable reference alignment while preserving structural fidelity. In contrast, IoU-focused training performs worse, indicating that optimizing for reference matching alone leads to brittle spatial reasoning and insufficient attention to intrinsic geometric constraints. This suggests that the IoU reward functions primarily as auxiliary anchoring rather than as the main supervision signal. Instead, spatial reasoning is predominantly learned from intrinsic geometric objectives, enabling effective training even without paired layout annotations.

5 Related Works

Spatial Reasoning with LLMs Vision–Language Models (VLMs) (Fan et al.,

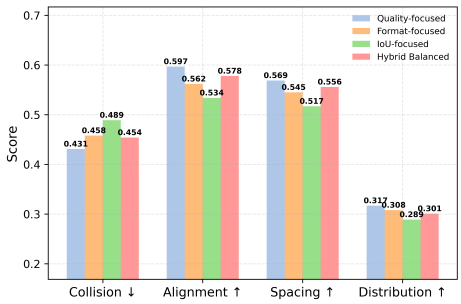


Figure 5: Ablation study on varying reward weights

2025) and Multimodal LLMs (MLLMs) (Wu et al., 2025a) enable spatial reasoning by grounding language in visual inputs, using multi-view images (Qi12 et al.; Fu et al., 2025; Gu et al., 2024), depth cues (Ma et al., 2024; Cheng et al., 2024; Cai et al., 2025) or cross-modal embeddings (Lüddecke and Ecker, 2022; Chen et al., 2024). While effective, these methods are computationally intensive, encode geometry implicitly, and offer limited interpretability.

Content-Aware Layout Generation Traditional layout generation relies on deep generative models such as VAEs (Jyothi et al., 2019), GANs (Zhou et al., 2022; Hsu et al., 2023; Li et al., 2019), transformers (Kong et al., 2022), autoregressive (Horita et al., 2024), and diffusion-based models (Chai et al., 2023b,a), supplemented by retrieval (Horita et al., 2024) or post-hoc refiners (Shen et al., 2025) to reduce overlap and misalignment. These approaches rely heavily on annotated data and thus generalizing poorly beyond training distributions. *LaySPA* differs by formulating layout design as reinforcement learning over a structured textual environment, enabling diverse exploration of geometrically valid layouts guided by spatial objectives.

LLM-based Layout Generation Recent works leverage LLMs' design priors by serializing layouts into text formats such as HTML (Tang et al., 2023; Seol et al., 2024), CSS (Feng et al., 2023), or SVG (Hsu and Peng, 2025) and applying recursive reasoning to capture spatial dependencies (Tian et al., 2025). Retrieval-based pipelines (Lin et al., 2023; Forouzandehmehr et al., 2025) ground outputs in exemplar layouts and iteratively refine them using LLMs and vision–language feedback. These methods often rely on emergent capabilities and exemplar retrieval, whereas *LaySPA* is explicitly guided by spatial and visual rewards and optimized through peer-relative policy learning.

6 Conclusion

We introduce *LaySPA*, a GRPO-based framework that reframes content-aware layout design as a policy learning problem, guiding LLMs to perform explicit spatial reasoning over structured textual environments. Optimized with multi-objective spatial critiques, *LaySPA* learns robust and generalizable spatial behaviors for layout design. Empirically, *LaySPA* substantially improves layout quality, outperforms larger proprietary LLMs, and approaches specialized multimodal generators without visual inputs and fewer training samples. These results demonstrate that spatial reasoning can be explicitly optimized in text, positioning language-grounded, reward-driven policy learning as a scalable alternative to vision-centric layout design. Limitations include reliance on precomputed saliency and single-step generation. Future work will explore richer semantic grounding, multi-turn reinforcement learning, and hierarchical design policies.

Acknowledgments

Original images from PKU-PosterLayout dataset by Hsu et al. available at https://huggingface.co/datasets/creative-graphic-design/PKU-PosterLayout, licensed under CC BY-SA 4.0. Original images from CGL dataset by Zhou et al. available at https://huggingface.co/datasets/creative-graphic-design/CGL-Dataset, licensed under CC BY-NC-SA 4.0.

References

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–718, Nusa Dua, Bali. Association for Computational Linguistics.

Wenxiao Cai, Iaroslav Ponomarenko, Jianhao Yuan, Xiaoqi Li, Wankou Yang, Hao Dong, and Bo Zhao. 2025. Spatialbot: Precise spatial understanding with vision language models. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9490–9498. IEEE.

Shang Chai, Liansheng Zhuang, and Fengying Yan. 2023a. Layoutdm: Transformer-based diffusion

model for layout generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18349–18358.

Shang Chai, Liansheng Zhuang, Fengying Yan, and Zihan Zhou. 2023b. Two-stage content-aware layout generation for poster designs. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 8415–8423.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. 2024. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14455–14465.

Junlong Chen, Jens Grubert, and Per Ola Kristensson. 2025. Analyzing multimodal interaction strategies for llm-assisted manipulation of 3d scenes. In *2025 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*, pages 206–216. IEEE.

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. 2024. Spatialrgpt: Grounded spatial reasoning in vision-language models. *Advances in Neural Information Processing Systems*, 37:135062–135093.

Zhiwen Fan, Jian Zhang, Renjie Li, Junge Zhang, Runjin Chen, Hezhen Hu, Kevin Wang, Huaizhi Qu, Dilin Wang, Zhicheng Yan, and 1 others. 2025. Vlm-3r: Vision-language models augmented with instruction-aligned 3d reconstruction. *arXiv preprint arXiv:2505.20279*.

Weixi Feng, Wanrong Zhu, Tsu-jui Fu, Varun Jampani, Arjun Akula, Xuehai He, Sugato Basu, Xin Eric Wang, and William Yang Wang. 2023. Layoutgpt: Compositional visual planning and generation with large language models. *Advances in Neural Information Processing Systems*, 36:18225–18250.

Mohamed Amine Ferrag, Norbert Tihanyi, and Merouane Debbah. 2025. From llm reasoning to autonomous ai agents: A comprehensive review. *arXiv preprint arXiv:2504.19678*.

Najmeh Forouzandehmehr, Reza Yousefi Maragheh, Sriram Kollipara, Kai Zhao, Topojoy Biswas, Evren Korpeoglu, and Kannan Achan. 2025. Cal-rag: Retrieval-augmented multi-agent generation for content-aware layout design. *arXiv preprint arXiv:2506.21934*.

Rao Fu, Jingyu Liu, Xilun Chen, Yixin Nie, and Wenhan Xiong. 2025. Scene-llm: Extending language model for 3d visual reasoning. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2195–2206.

Qiao Gu, Ali Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, Bipasha Sen, Aditya Agarwal, Corban Rivera, William Paul, Kirsty Ellis, Rama Chellappa, and 1 others. 2024. Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5021–5028. IEEE.

Daichi Horita, Naoto Inoue, Kotaro Kikuchi, Kota Yamaguchi, and Kiyoharu Aizawa. 2024. Retrieval-augmented layout transformer for content-aware layout generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 67–76.

Hsiao Yuan Hsu, Xiangteng He, Yuxin Peng, Hao Kong, and Qing Zhang. 2023. Posterlayout: A new benchmark and approach for content-aware visual-textual presentation layout. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6018–6026.

HsiaoYuan Hsu and Yuxin Peng. 2025. Postero: Structuring layout trees to enable language models in generalized content-aware layout generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8117–8127.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Akash Abdu Jyothi, Thibaut Durand, Jiawei He, Leonid Sigal, and Greg Mori. 2019. Layoutvae: Stochastic scene layout generation from a label set. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9895–9904.

Xiang Kong, Lu Jiang, Huiwen Chang, Han Zhang, Yuan Hao, Haifeng Gong, and Irfan Essa. 2022. Blt: Bidirectional layout transformer for controllable layout generation. In *European Conference on Computer Vision*, pages 474–490. Springer.

Jianan Li, Jimei Yang, Aaron Hertzmann, Jianming Zhang, and Tingfa Xu. 2019. Layoutgan: Generating graphic layouts with wireframe discriminators. *arXiv preprint arXiv:1901.06767*.

Jiawei Lin, Jiaqi Guo, Shizhao Sun, Zijiang Yang, Jian-Guang Lou, and Dongmei Zhang. 2023. Layoutprompter: Awaken the design ability of large language models. *Advances in Neural Information Processing Systems*, 36:43852–43879.

Timo Lüddecke and Alexander Ecker. 2022. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7086–7096.

Chenyang Ma, Kai Lu, Ta-Ying Cheng, Niki Trigoni, and Andrew Markham. 2024. Spatialpin: Enhancing spatial reasoning capabilities of vision-language models through prompting and interacting 3d priors.

Advances in neural information processing systems, 37:68803–68832.

Aske Plaat, Annie Wong, Suzan Verberne, Joost Broekens, Niki Van Stein, and Thomas Bäck. 2025. Multi-step reasoning with large language models, a survey. *ACM Computing Surveys*, 58(6):1–35.

Zekun Qi12, Runpei Dong12, Shaochen Zhang, Haoran Geng, Chunrui Han, Zheng Ge, Li Yi567, and Kaisheng Ma. Shapellm: Universal 3d object understanding for embodied interaction-supplementary material.

Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. 2019. Basnet: Boundary-aware salient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7479–7489.

Kanchana Ranasinghe, Satya Narayan Shukla, Omid Poursaeed, Michael S Ryoo, and Tsung-Yu Lin. 2024. Learning to localize objects improves spatial reasoning in visual-llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12977–12987.

Jaejung Seol, Seojun Kim, and Jaejun Yoo. 2024. Posterllama: Bridging design ability of language model to content-aware layout generation. In *European Conference on Computer Vision*, pages 451–468. Springer.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

I Shen, Ariel Shamir, Takeo Igarashi, and 1 others. 2025. Layoutrectifier: An optimization-based post-processing for graphic design layout generation. *arXiv preprint arXiv:2508.11177*.

Zecheng Tang, Chenfei Wu, Juntao Li, and Nan Duan. 2023. Layoutnuwa: Revealing the hidden layout expertise of large language models. *arXiv preprint arXiv:2309.09506*.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Jiaxu Tian, Xuehui Yu, Yaoxing Wang, Pan Wang, Guangqian Guo, and Shan Gao. 2025. Relayout: Integrating relation reasoning for content-aware layout generation with multi-modal large language models. *arXiv preprint arXiv:2507.05568*.

Heng Wang, Yotaro Shimose, and Shingo Takamatsu. 2025. BannerAgency: Advertising banner design with multimodal LLM agents. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 4304–4329, Suzhou, China. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. 2025a. Spatial-mllm: Boosting mllm capabilities in visual-based spatial intelligence. *arXiv preprint arXiv:2505.23747*.

Fang Wu, Vijay Prakash Dwivedi, and Jure Leskovec. 2025b. Large language models are good relational learners. *arXiv preprint arXiv:2506.05725*.

Xiaoqian Wu, Yong-Lu Li, Jianhua Sun, and Cewu Lu. 2023. Symbol-llm: leverage language models for symbolic system in visual human activity reasoning. *Advances in neural information processing systems*, 36:29680–29691.

Jirong Zha, Yuxuan Fan, Xiao Yang, Chen Gao, and Xinlei Chen. 2025. How to enable llm with 3d capacity? a survey of spatial reasoning in llm. *arXiv preprint arXiv:2504.05786*.

Yuanzhao Zhai, Tingkai Yang, Kele Xu, Dawei Feng, Cheng Yang, Bo Ding, and Huaimin Wang. 2025. Enhancing decision-making for llm agents via step-level q-value models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 27161–27169.

Jinghan Zhang, Xiting Wang, Weijieying Ren, Lu Jiang, Dongjie Wang, and Kunpeng Liu. 2025. Ratt: A thought structure for coherent and correct llm reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 26733–26741.

Junyi Zhang, Jiaqi Guo, Shizhao Sun, Jian-Guang Lou, and Dongmei Zhang. 2023. Layoutdiffusion: Improving graphic layout generation by discrete diffusion probabilistic models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7226–7236.

Xinru Zheng, Xiaotian Qiao, Ying Cao, and Rynson WH Lau. 2019. Content-aware generative modeling of graphic design layouts. *ACM Transactions on Graphics (TOG)*, 38(4):1–15.

Min Zhou, Chenchen Xu, Ye Ma, Tiezheng Ge, Yuning Jiang, and Weiwei Xu. 2022. Composition-aware graphic layout gan for visual-textual presentation designs. *arXiv preprint arXiv:2205.00303*.

A Visualization of Rewards

Figure 6 illustrates how individual components of the layout quality score guide the agent toward content-aware designs that jointly satisfy structural feasibility and visual appeal.

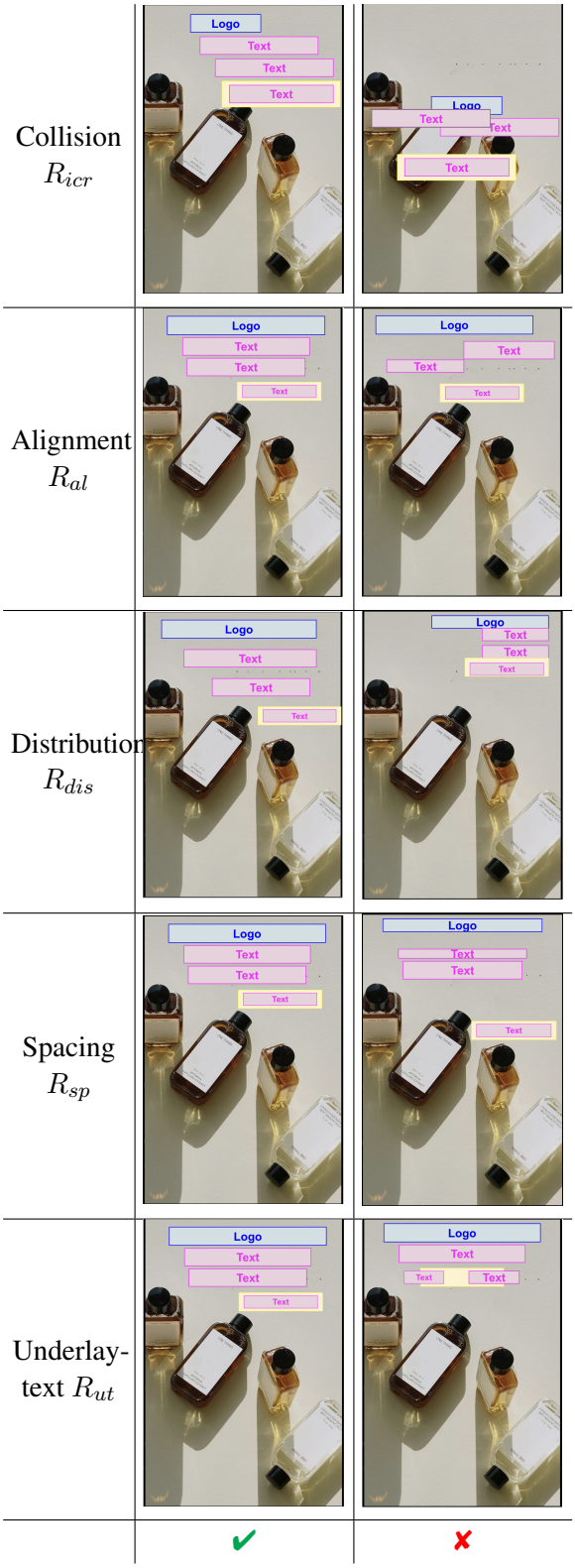


Figure 6: An illustration of how each layout quality score function guides the agent toward human-preferred designs (✔) while discouraging (✘) undesirable spatial arrangements.

B Dataset

Table 2 lists the details of the CGL and PKU benchmark.

Table 2: Layout generation datasets.

Dataset	CGL	PKU
# Layouts	60,548	9,974
# Canvases	1,000	905
Elements	Text, Logo, Underlay, Emb.	Text, Logo, Underlay
Canvas	Non-empty	Non-empty
Categories	Cosmetics, Elec., Clothing, etc.	Cosmetics, Elec., Clothing, Food, Toys, etc.

C Baseline Methods

- **DS-GAN** (Hsu et al., 2023): DS-GAN is a CNN-LSTM-based conditional generative adversarial network that learns design sequences conditioned on a given canvas to automatically generate content-aware visual-textual presentation layouts.
- **PosterLlama** (Seol et al., 2024): a multimodal layout generation system that uses CodeLLaMA-7B as the language backbone with a DINOv2 visual encoder. Layouts are reformulated as HTML sequences. The model is trained via supervised instruction tuning with depth-based data augmentation.
- **Qwen-2.5-7B-Instruct**: A dense transformer language model with 7 billion parameters. Given the same inputs as *LaySPA*, while used in a training-free, prompt-based setting to generate geometric attributes for all layout elements.
- **GPT-4o**: A large-scale proprietary multimodal foundation model that takes the canvas as visual input and is prompted to generate the geometric specifications of all layout elements.