arXiv:2506.10890v1 [cs.CV] 12 Jun 2025

CreatiPoster: Towards Editable and Controllable Multi-Layer Graphic Design Generation

Zhao Zhang
zzhang@mail.nankai.edu.cn
ByteDance, Intelligent Creation

Yutao Cheng
taorebobi@gmail.com
ByteDance, Intelligent Creation

Dexiang Hong
hongdexiang@bytedance.com
ByteDance, Intelligent Creation

Maoke Yang
yangmaoke@bytedance.com
ByteDance, Intelligent Creation

Gonglei Shi
shigonglei@gmail.com
ByteDance, Intelligent Creation

Lei Ma
malei.luciano@bytedance.com
ByteDance, Intelligent Creation

Hui Zhang
hui_zhang23@m.fudan.edu.cn
ByteDance, Fudan University

Jie Shao
shaojie.mail@bytedance.com
ByteDance, Intelligent Creation

Xinglong Wu
wuxinglong@bytedance.com
ByteDance, Intelligent Creation

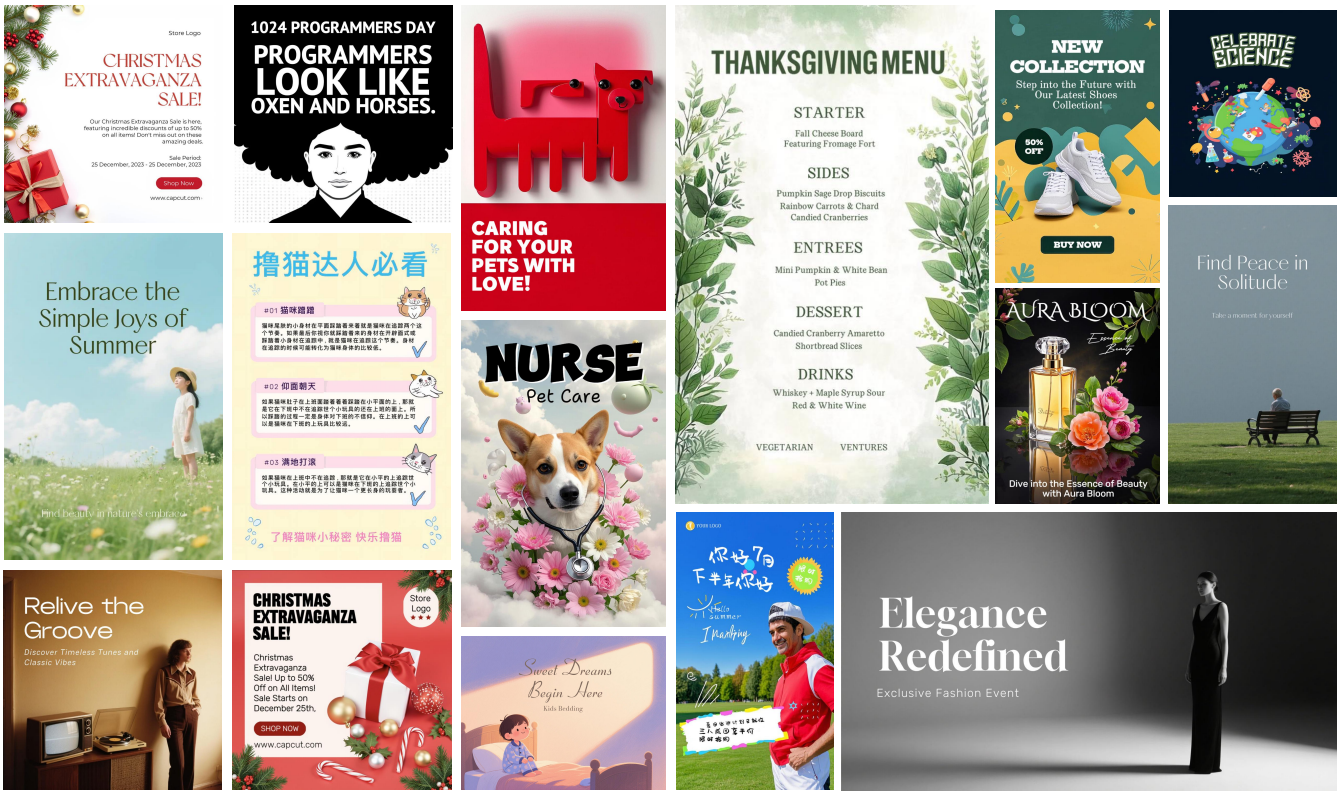


Figure 1: Multi-layer compositions produced by CreatiPoster. The protocol lists every text and upload asset layer, letting users freely edit content, placement, and style in the GUI editor.

Authors' addresses: Zhao Zhang, zzhang@mail.nankai.edu.cn, ByteDance, Intelligent Creation, ; Yutao Cheng, taorebobi@gmail.com, ByteDance, Intelligent Creation, ; Dexiang Hong, hongdexiang@bytedance.com, ByteDance, Intelligent Creation, ; Maoke Yang, yangmaoke@bytedance.com, ByteDance, Intelligent Creation, ; Gonglei Shi, shigonglei@gmail.com, ByteDance, Intelligent Creation, ; Lei Ma, malei.luciano@bytedance.com, ByteDance, Intelligent Creation, ; Hui Zhang, hui_zhang23@m.fudan.edu.cn, ByteDance, Fudan University, ; Jie Shao, shaojie.mail@bytedance.com, ByteDance, Intelligent Creation, ; Xinglong Wu, wuxinglong@bytedance.com, ByteDance, Intelligent Creation,

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.
0730-0301/2025/6-ART $15.00
https://doi.org/XXXXXXX.XXXXXXX

ABSTRACT

Graphic design plays a crucial role in both commercial and personal contexts, yet creating high-quality, editable, and aesthetically pleasing graphic compositions remains a time-consuming and skill-intensive task—especially for beginners. Current AI tools automate parts of the workflow, but struggle to accurately incorporate user-supplied assets, maintain editability, and achieve professional visual appeal. Commercial systems, like Canva Magic Design, rely on vast template libraries, which are impractical for replicate. In this paper, we introduce CreatiPoster, a framework that generates editable, multi-layer compositions from optional natural-language instructions or assets. A protocol model, an RGBA large multimodal model, first produces a JSON specification detailing every layer—text or asset—with precise layout, hierarchy, content and style, plus a concise background prompt. A conditional background model then synthesizes a coherent background conditioned on this rendered foreground layers. We construct a benchmark with automated metrics for graphic-design generation and show that CreatiPoster surpasses leading open-source approaches and proprietary commercial systems. To catalyze further research, we release a copyright-free corpus of 100,000 multi-layer designs. CreatiPoster supports diverse applications such as canvas editing, text overlay, responsive resizing, multilingual adaptation, and animated posters, advancing the democratization of AI-assisted graphic design. Project homepage: https://github.com/graphic-design-ai/creatiposter

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; • **Applied computing** → **Arts and humanities**.

KEYWORDS

Graphic design, poster generation, diffusion model, large multimodal model, animated poster

ACM Reference Format:
Zhao Zhang, Yutao Cheng, Dexiang Hong, Maoke Yang, Gonglei Shi, Lei Ma, Hui Zhang, Jie Shao, and Xinglong Wu. 2025. CreatiPoster: Towards Editable and Controllable Multi-Layer Graphic Design Generation. *ACM Trans. Graph.* 1, 1 (June 2025), 11 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Graphic design is a professional field that uses creative visual communication to deliver targeted messages. It combines text and images in forms such as posters, social media graphics, and business cards, playing a key role in both marketing and daily life. The discipline demands a high level of expertise, creativity, and innovation, with designers investing significant time to master complex tools and techniques. Creating antithetical graphic compositions is especially time-consuming and requires considerable skill and effort. For beginners, this process is even more challenging.

Recently, there has been a significant paradigm shift as Artificial intelligence (AI) is increasingly integrated into various stages of the design process, including layout generation [Tabata et al. 2019; Yamaguchi 2021], hierarchical layout generation [Cheng et al. 2025], logo creation [Wang et al. 2022], artistic text generation [Gao et al. 2019], and color harmony [Cohen-Or et al. 2006] Some recent methods [Inoue et al. 2024; Jia et al. 2023] can even generate complete graphic compositions based on user language instructions. However, these approaches still face limitations, such as difficulty incorporating user-provided assets and lacking sufficient aesthetic appeal. Commercial platforms—Canva Magic Design[1], Adobe Express[2], and Microsoft Designer[3]—offer strong usability by pairing vast template libraries with intelligent search and auto-copywriting. Building and licensing such libraries, however, is impractical for most individuals and small organizations.

A high-quality AI-generated graphic composition should satisfy four core criteria: Text accuracy: Posters rely on text to convey information, so accurate spelling and proper typography are crucial. asset fidelity: User-provided assets—such as product images, personal photos, or brand logos—must be correctly placed and preserved in the final design. Editability: Users may wish to modify the generated text or replace certain assets, so the composition should be easily editable. Aesthetic appeal: Beyond functionality, the overall design should be visually pleasing. Currently, no AI system fully achieves all these requirements. However, reaching this level would greatly enhance design efficiency and promote the democratization of graphic design.

To close this gap, we have developed an open-source, multi-layer, editable graphic design generation system, CreatiPoster, which not only meets the above requirements but also offers interactive flexibility and high expansiveness. As shown in Figure 2, CreatiPoster couples a protocol model with a background model: the protocol model—a large multimodal network [Cheng et al. 2025]—turns user instructions and optional assets into a JSON "protocol" that lists every layer (text or asset) with its z-order, position, font, size, and content, and also emits a concise prompt describing the desired backdrop; this JSON can be rendered immediately with engines such as Skia[4], yielding fully editable foreground layers as illustrated in Figure 4; the background model then uses the rendered foreground and the prompt to synthesize a matching background, after which foreground and background are composited into a complete, multi-layer graphic whose elements remain disentangled for later editing; evaluated on a new benchmark with automated metrics, CreatiPoster outperforms existing open-source and commercial systems, and we further release 100,000 copyright-free, multi-layer samples to catalyze research in AI-driven graphic design. The contributions of this paper can be summarized as follows:

- We propose CreatiPoster, a open system that produces visually compelling, multi-layer graphic compositions while preserving full editability of text and assets.
- CreatiPoster accommodates diverse user instructions, including prompt-only, asset-only, mixed input, as well as explicit specification of text/asset layout or attributes.
- Our method demonstrates high extensibility, supporting a wide range of interesting applications canvas operation, text overlay, responsive resizing, multilingual generation, and animated poster.

[1] https://www.canva.com/magic-design/
[2] https://www.adobe.com/express/
[3] https://designer.microsoft.com/
[4] https://skia.org/

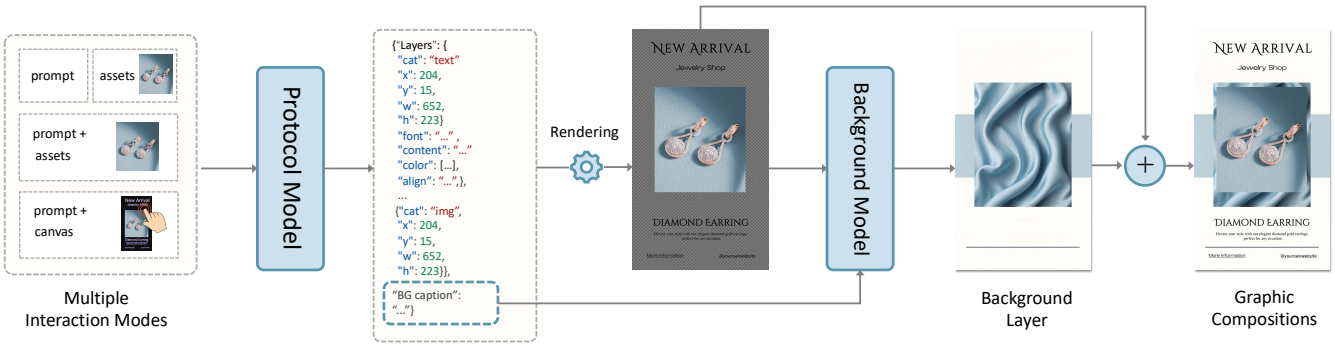


Figure 2: Overview of the CreatiPoster pipeline. User inputs are processed by the protocol model to generate editable design layers, while the background model creates a matching background. The final graphic composition combines both outputs.

- We release a copyright-free training dataset and a comprehensive benchmark to the community, aiming to advance research in multi-layer graphic design generation.

2 RELATED WORK

2.1 Large Multimodal Models

LMM [Li et al. 2023a; Yin et al. 2023; Zhang et al. 2024b] has widely used in autonomous driving [Cui et al. 2024; Ding et al. 2024], video understanding [Li et al. 2023b; Tang et al. 2023; Zhang et al. 2023b], operating system [Wang et al. 2024a; Yang et al. 2023], image/video generation [Wang et al. 2024b; Zhou et al. 2024], embodied intelligence [Brohan et al. 2023; Driess et al. 2023; Mu et al. 2024] and so on. In CreatiPoster, we use LMM to predict the coordinates of text and users' assets. LMMs have also made remarkable progress in coordinate representation [Peng et al. 2023; Wang et al. 2023]. Notable method is Shikra [Chen et al. 2023], which represents spatial positions using numerical values in natural language. Recently, Graphist [Cheng et al. 2025] introduce these coordinate representation in layout generation.

2.2 Image Generation

Text-to-image (T2I) generation has progressed rapidly in both visual fidelity and text–image alignment. Recent Multimodal Diffusion Transformer (MM-DiT) models—Stable Diffusion series [Esser et al. 2024; stability.ai 2024], FLUX.1 [Labs 2024], HiDream [HiD 2025], and Seedream series [Gao et al. 2025; Gong et al. 2025]—push the frontier by scaling parameters and training data. In parallel, conditional control of image generation—complementary to MM-DiT frameworks—has achieved notable progress, exemplified by techniques for subject control [Mou et al. 2025; Tan et al. 2024; Zhang et al. 2025], layout control [Zhang et al. 2024a] and other semantic-preserving manipulations. LayerDiffuse [Zhang and Agrawala 2024] represents the first approach to enable multi-layer conditional control for image generation.

2.3 Design System

Ideally, automated graphic design should be visually appealing, clearly present information, support user-defined assets (like product images and logos), and remain editable. Early automated design systems relied on aesthetic rules [Bauerly and Liu 2006; Yang et al. 2016] or restricts [Hurst et al. 2009; Jahanian et al. 2013], later works also addressing subproblems like layout [Chai et al. 2023; Cheng et al. 2025; Inoue et al. 2023; Jiang et al. 2023; Li et al. 2019; Yu et al. 2022; Zhang et al. 2023a] and color [Cohen-Or et al. 2006; Tokumaru et al. 2002; You et al. 2019] to improve automation in design. Recent methods have mainly sought to simplify the modeling of graphic design problems, often achieving a more unified and complete system at the cost of either editability or support for user-defined assets. For example, some layout-based approaches [Horita et al. 2024; Hsu et al. 2023; Li et al. 2023c; Seol et al. 2024; Zhou et al. 2022] only support designs where the main image fills the entire canvas, while certain multi-stage generation methods abstract and simplify layers—methods [Chen et al. 2025; Inoue et al. 2024; Jia et al. 2023; Wang et al. 2025] sacrifice support for user assets, and ART [Pu et al. 2025] compromises on text editability. Our approach also simplifies the modeling of graphic design, but crucially, it preserves both support for user-defined assets and text editability. This makes our system a more versatile and general-purpose solution for automated graphic design.

3 PROPOSED METHOD

Given a user prompt $p \in \mathcal{S}$, a canvas size $\mathbf{s} \in \mathbb{R}^2$, and an optional set of n RGBA assets $\mathcal{I} = \{I_i \in \mathbb{R}^{h_i \times w_i \times 4}\}_{i=1}^{n}$, our pipeline (Fig. 2) comprises a Protocol Model PM and a Background Model BM. The former predicts a layerwise protocol, while the latter synthesises a background consistent with both the protocol and the rendered foreground. Formally,

$$\mathrm{PM}(\mathcal{I}, p, \mathbf{s}) \longrightarrow \mathcal{P} = \{\texttt{caption} = c,\ \texttt{layers} = \mathcal{L}\}, \tag{1}$$

$$\mathrm{R}(\mathbf{s}, \mathcal{L}) \longrightarrow I_{\mathrm{fg}}, \tag{2}$$

$$\mathrm{BM}(I_{\mathrm{fg}}, c) \longrightarrow I_{\mathrm{bg}}, \tag{3}$$

where c is a concise background caption and $\mathcal{L} = [\ell_1, \ldots, \ell_m]$ is an ordered list of layer specifications (text or asset). It is represented in the form of JSON, containing the necessary fields to render its visual content. R is the rendering engine. The final output is the editable pair $\{\texttt{BG}: I_{\mathrm{bg}},\ \texttt{FG}: \mathcal{L}\}$. In Figure 3, we show the intermediate results of the CreatiPoster output under different modes.

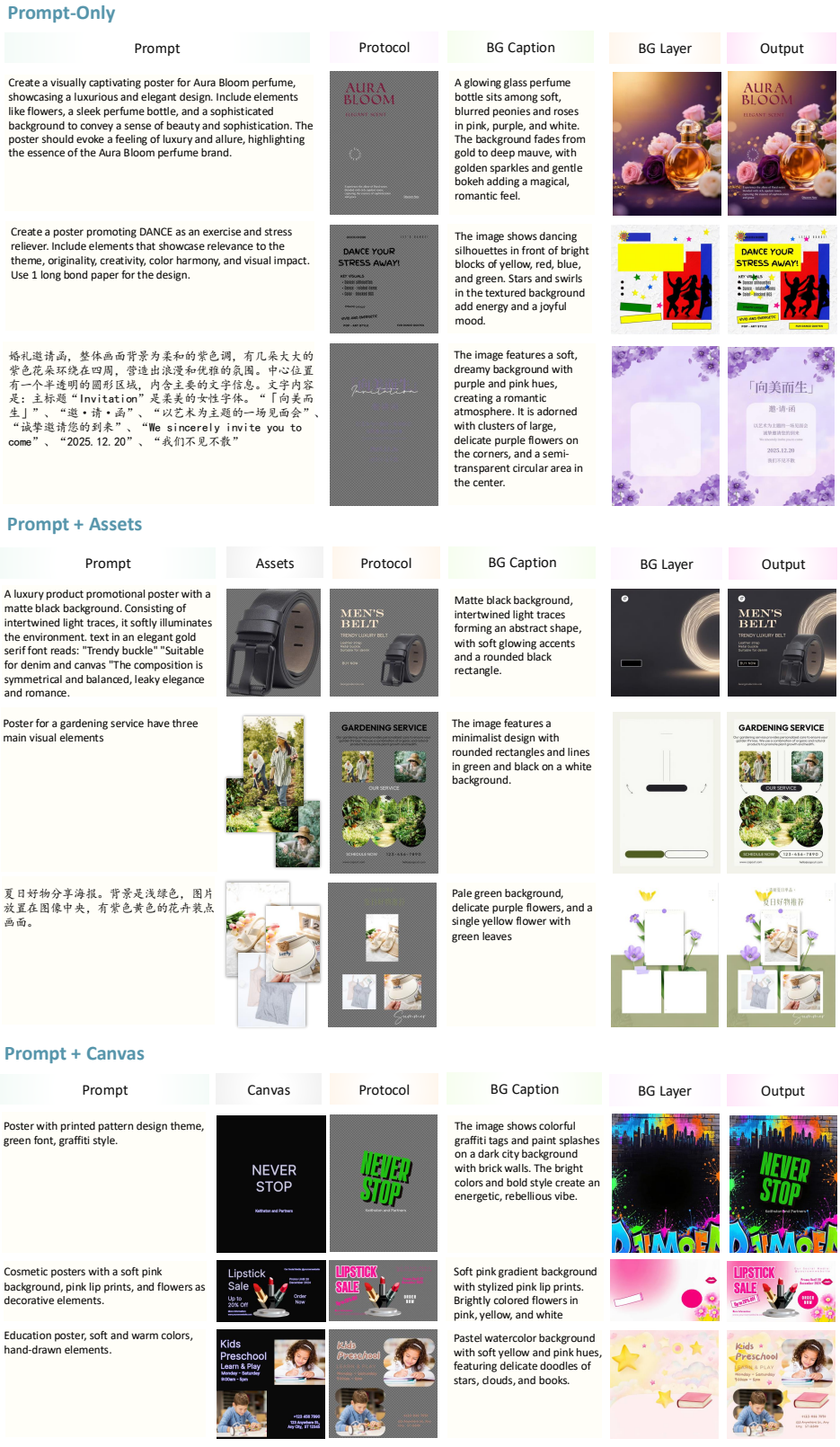


Figure 3: Graphic compositions generated by CreatiPoster with different interaction modes.

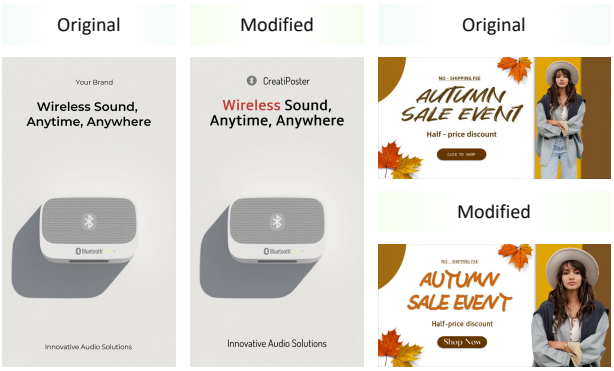


Figure 4: Example of editing a graphic composition generated by CreatiPoster in the editor. Users can modify text and asset content, layout, and style via JSON fields using an intuitive GUI, enabling professional-level customization.

3.1 Protocol Model

3.1.1 ***Architecture.*** Protocol Model adopts the same architecture as CreatiGraphist [Cheng et al. 2025], consisting of an RGBA-Encoder, a Visual Shrinker, and a Large Language Model (LLM). Each input image is processed by the RGBA-Encoder and the Vision Shrinker into 64 tokens. Unlike many existing Large Multimodal Models (LMMs) that employ multi-resolution strategies to enhance the description and localization of small objects, our approach deliberately compresses the image representation and standardizes the token count to 64. This design choice is motivated by our focus on capturing edge textures (from RGBA images) and global visual information, while also maintaining strict control over the total number of tokens for computational efficiency.

3.1.2 ***Protocol.*** The JSON protocol $\mathcal{L}$ predicted by CreatiPoster contains both text and asset layers. For text layers, the protocol specifies attributes such as content, font family, font size, position, color, stroke properties, rotation angle, bend, bold, italic, underline, alignment, line spacing, and character spacing—collectively defining the text's visual appearance. For asset layers, the protocol includes fields such as position, cropping, rotation, and mask type. The layer order, from bottom to top, determines the z-order in the composition and thus the stacking sequence in the final output. All these fields are fully editable within the editor, providing a high degree of flexibility and control over the generated results, as illustrated in the Figure 4.

3.1.3 ***Training Stragy.*** For the basic training schedule, we follow the multi-stage training approach described in [Cheng et al. 2025]. During the training phase, we construct both prompt-only and prompt-assets modes to support flexible input scenarios. Additionally, to better enable the canvas mode, we randomly sample a subset of layers $\mathcal{L}^{'}$ from the predicted $\mathcal{L}$ and incorporated it into the prompt. Additionally, for each layer in $\mathcal{L}^{'}$, we randomly drop out part of their attribute and position fields, prompting the model to infer the missing information based on the given context. In this way, we enable users to freely specify their desired content and attributes during inference time.

3.2 Background Model

3.2.1 ***Model Architecture.*** The background model takes as input the RGBA foreground image output by the protocol model and a caption describing the background, then generates a matching background that harmonizes with the foreground. To process the RGBA foreground, we first convert it into an RGB image with a gray background, which is then encoded using the native variational autoencoder (VAE). For the background prompt, we employ the corresponding text encoder to generate background prompt embeddings. After encoding the background prompt, noise image, and foreground image condition into tokens, denoted as h_b, h_z, and h_f respectively, we concatenate these tokens along the token dimension and feed the resulting token sequence into a stack of MM-DiT blocks. Each block comprises linear projection layers (for Q, K, V), multimodal attention (MM-Attention), and feed-forward networks (FFN). For each type of token, a linear projection is applied to map them into the corresponding query, key, and value spaces: $Q^*, K^*, V^* = \text{Linear}(h^*)$, where the symbol $*$ represents different modalities (background prompt, noisy image, or foreground condition). For the foreground image condition tokens h_f, we further adapt their representations through the deployment of LoRA modules on the linear layers and adaptive layer normalization (AdaLN). This approach enables efficient fine-tuning and feature alignment. To ensure that the positional slots of design elements in the generated background are strictly consistent with those in the foreground, we reuse the positional encoding of the noisy image and add it to the foreground conditional image representation. The multimodal attention is then computed as:

$$h^{b\prime}, h^{z\prime}, h^{f\prime} = \text{Att}([\mathbf{Q}^b, \mathbf{Q}^z, \mathbf{Q}^f], [\mathbf{K}^b, \mathbf{K}^z, \mathbf{K}^f], [\mathbf{V}^b, \mathbf{V}^z, \mathbf{V}^f]). \quad (4)$$

3.2.2 ***Training Strategy.*** We divide the training process into two stages. In both stages, the backbone network is frozen, and only the LoRA components are trained. **Pre-training:** During the pre-training phase, we employ 512-resolution images and use a lognormal noise schedule with a mean of 0.5 and a standard deviation of 1. This approach aims to enhance the model's perception of foreground locations by leveraging the characteristics of lognormal noise, which emphasizes spatial features in the low-resolution domain. **Post-training:** In the post-training stage, the training resolution is increased to 1024, and a uniform noise schedule is adopted. This configuration is designed to optimize the model's performance across all noise scales, ensuring comprehensive generalization in the high-resolution setting. This two-stage strategy combines low-resolution pre-training for improved foreground sensitivity and high-resolution post-training for multi-scale noise adaptability, thereby enhancing the overall performance of the model across different scenarios.

4 EXPERIMENTS

4.1 Implement Details

We use InternLM2.5 [Cai et al. 2024] as the LLM backbone for the Protocol Model, training it on a combination of in-house designer poster data, multimodal content understanding data, and conversational data. For the background model, we trained two versions: CreatiPoster-F, which uses FLUX-dev [Labs 2024] as the backbone, and CreatiPoster-S, which uses an Seedream3 [Gao et al. 2025]

model. Both versions employ a LoRA rank of 256. Training was conducted on 8 NVIDIA A100 GPUs. The Protocol Model was trained for approximately 5 days, while the Background Model required about 3 days.

4.2 Test Dataset

For the construction of the test set, we use 45 prompt-only test cases, 39 prompt-single-asset input cases, and 6 prompt-multi-asset input cases. For the prompt-only test cases, we collected 45 poster images from the internet and use a lmm model[5] to generate captions, which served as user prompts. For the cases with image inputs, we generated images using a text-to-image model[5], extracted the corresponding main subjects as foreground materials, and then use a lmm model[5] to simulate user input for prompt generation.

4.3 Metrics

There are many quantifiable methods for evaluation of aesthetics in the field of generic image generation, such as FID [Dowson and Landau 1982], PSNR, SSIM. However, these metrics are heavily influenced by the generative model, and can be highly variable in the use of models such as stable diffusion, flux, and others. And it is still a gap in the field of graphic design in terms of a general and rational evaluation system. We consulted experts in the field of graphic design, who summarized some of the more appropriate evaluation dimensions for current poster designs as layout, color, graphic style and compliance. Table 1 explains a more detailed description of each evaluation dimension. In addition to this, we gathered 10 volunteers to single-blindly rate the overall feeling on a scale of 1-5. In order to be more objective and comprehensive, we quantitatively scored the cases from 90 cases using GPT4.1 [6] based on the above dimensions and scored them on a scale of 1-5 for each dimension. Given the diversity of GPT results, we sampled each case 10 times and used a majority of votes to determine the final score for each example. The method of comparison consists of the results from open-source approaches (OpenCOLE) and closed-source commercial systems (Microsoft Designer, Adobe Express).

Dimension	Description
Layout	Focuses on layout and compositional appropriateness.
Color	Evaluates whether the color scheme aligns with the poster content and whether the colors are coordinated.
Graphic Style	Evaluate how well the fonts, decorative elements, assets, and backgrounds work together, as well as the overall style of the poster.
Compliance	Evaluate how well the poster generation results follow the prompt.

Table 1: Graphic Design Evaluation Dimensions

4.4 Results and Discussion

During testing, because some models do not support asset input, we divided the experiments into two scenarios: one without asset input (Table 2) and one with asset input (Table 3). Furthermore, since none of the other comparison methods support multiple asset inputs, only the remaining 84 cases were included in the comparative analysis.

Scores from GPT

Method	CreatiPoster-S	CreatiPoster-F	OpenCOLE	Microsoft Designer	Canva
Layout	**2.89**	2.71	1.60	2.61	2.85
Color	4.33	4.36	**4.57**	3.55	4.11
Graphic Style	**4.24**	3.97	2.33	3.64	3.68
Compliance	**3.73**	3.67	3.03	2.38	3.20

Scores from human

Method	CreatiPoster-S	CreatiPoster-F	OpenCOLE	Microsoft Designer	Canva
Overall	**2.80**	2.59	1.68	1.77	2.33

Table 2: Scores for each method from aspects of Layout, Color, Graphic Style and Compliance, without assets input.

Scores from GPT

Method	CreatiPoster-S	CreatiPoster-F	Microsoft Designer	Canva
Layout	2.41	2.25	2.54	**2.84**
Color	**4.28**	4.25	4.25	4.02
Graphic Style	**3.92**	**3.92**	3.23	3.76
Compliance	**3.59**	3.48	2.54	2.79

Scores from human

Method	CreatiPoster-S	CreatiPoster-F	Microsoft Designer	Canva
Overall	**2.82**	2.67	2.05	2.33

Table 3: Scores for each method from aspects of Layout, Color, Graphic Style and Compliance with assets input.

From Table 2 and Table 3 CreatiPoster-S or CreatiPoster-F get first or near-first score in almost all evaluation dimensions, which proves the leadership of our method.

Layout. Based on both the table scores and human's overall rating, CreatiPoster demonstrates satisfactory performance in certain straightforward graphic design scenarios. However, it is important to note that all evaluated methods still achieve relatively low scores in the layout dimension, with none surpassing a score of 3. Among the methods, CreatiPoster-S achieves the highest score, closely followed by Canva (2.85). These results highlight a persistent challenge for current automated design tools: they struggle to produce layouts that are truly design-oriented and comparable to those created by human experts. This is especially evident in complex scenarios where graphics and text are intricately interwoven and must interact seamlessly. The limitations suggest that further advancements are needed for AI-driven tools to reach human-level proficiency in layout design.

Color. Across all evaluated methods, the color dimension consistently achieved relatively high scores, reflecting the strengths of the underlying generative models in handling color. Notably, GPT exhibited a clear preference for the color schemes produced by OpenCOLE [Inoue et al. 2024] and CreatiPoster-S, underscoring the exceptional ability of these models to generate visually harmonious and appealing color palettes. These results indicate that OpenCOLE and method-S are particularly adept at color selection, which plays a crucial role in enhancing the overall aesthetic quality of the posters.

[5]https://www.doubao.com/chat/
[6]https://openai.com/index/gpt-4-1/

Graphic Style. In evaluating graphic style, particular emphasis is placed on the overall stylistic coherence of the poster, including the harmonious integration of text and images. Leveraging the combined strengths of LMMs and generative modeling techniques, CreatiPoster excels in this dimension, achieving the highest overall score. This synergy enables CreatiPoster to produce posters that are not only visually appealing but also exhibit a strong sense of unity and coordination between graphic elements and textual content, resulting in a more polished and professional appearance.

Compliance. In terms of compliance, CreatiPoster-S/F stand out as the leading methods, consistently responding well to user prompts in both scenarios with and without asset inputs. Their superior performance is evident in their ability to accurately interpret and fulfill user requirements, regardless of the presence or resolution of input assets. This robust handling of diverse input conditions underscores the effectiveness of CreatiPoster-S/F in adhering to user instructions and generating outputs that closely match the intended design specifications.

Human Evaluate. Human evaluators generally concurred that our approach achieved higher scores across the assessed criteria. During the evaluation process, however, they also observed a notable pattern in the outputs (which is generated by Canva Magic Design and Microsoft Designer). Specifically, these methods may rely on a set of generic poster templates as their guarantee mechanism, simply replacing the text and the provided assets for each new case. As a result, the posters produced by Canva Magic Design and Microsoft Designer often appeared repetitive and lacked diversity in their overall design. This reliance on template-based generation limited the visual variety and creativity of the outputs, making them less distinctive compared to those generated by our approach, which demonstrated greater originality and adaptability to different prompts and assets.

Failure Cases. Although CreatiPoster delivers strong overall results, as shown in Figure 7, we still observe two recurring failure modes: 1) small icons often appear distorted, and 2) text and asset layers sometimes drift out of alignment. Icon distortion stems from the inability of current generators to retain fine detail and crisp edges at tiny scales—especially when an icon's structure is intricate or rendered in a low-resolution region of a complex layout. Misalignment, on the other hand, reflects limits in the protocol model's spatial reasoning: it can mis-estimate coordinates or ignore overlap and padding between layers. Users can fix these issues manually, but that adds extra work. Eliminating icon distortion will require higher-fidelity generators—e.g., native 4K diffusion models—while mitigating misalignment calls for better layout modeling and font-aware reasoning in large multimodal models, both of which remain open research challenges.

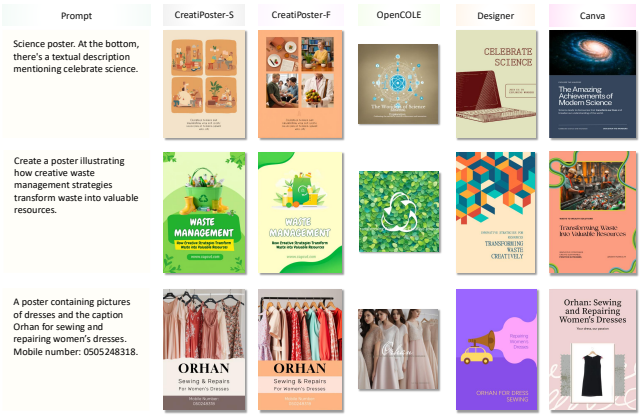


Figure 5: Comparison with methods in only text prompt input, including open-source approaches (OpenCOLE [Inoue et al. 2024] and closed-source commercial systems (Microsoft Designer, Canva Magic Design). Except for CreatiPoster, other methods can only output a few fixed sizes.

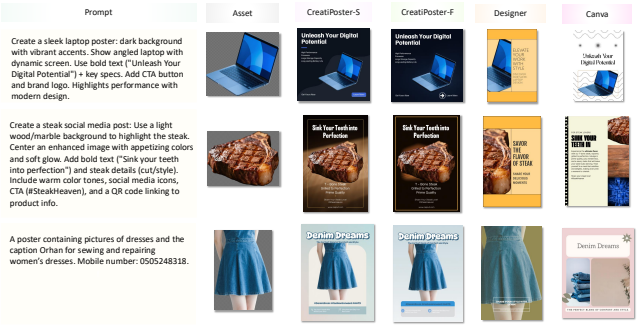


Figure 6: Comparison with methods in prompt-asset input setting, including closed-source commercial systems, Microsoft Designer and Canva Magic Design.

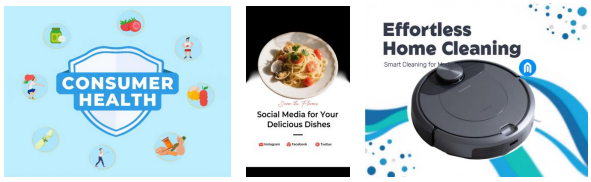


Figure 7: The main failure cases of CreatiPoster are due to distortion when generating small icons and occasional misalignment between text and assets.

5 APPLICATIONS

Thanks to its flexible architecture, CreatiPoster supports a diverse range of practical and creative applications. Below, we highlight several representative use cases.

Text Overlay. As illustrated in Figure 8, our protocol model allows users to overlay text directly to uploaded assets without invoking the background model. This is particularly useful for tasks such as adding titles to product images for e-commerce or overlaying text on photos for social media.

Poster Re-Layout. As shown in Figure 9, users can generate alternative graphic compositions of different sizes while preserving the original content and style. By reusing the rendered text and asset layers and specifying a new size, the protocol model predicts

the new foreground, and the background is regenerated using the same caption. This workflow is ideal for producing multiple video covers or posters tailored to various platforms.

Canvas Mode. The protocol model can accept fixed protocol fields, allowing for advanced canvas operations. For example, users can lock certain elements and only update new ones, or engage in multi-round editing. In this paper, we demonstrate scenarios where element positions are specified. The closed-source commercial system of the famous design platform Recraft [7] has similar canvas operation capabilities, but its generated posters are single-layer and not editable. As shown in the Figure 10, its IP retention ability and text accuracy are weaker than CreatiPoster.

Multilingual Generation. Since our protocol model was pretrained and trained on multilingual dialogue data, although it was not trained on multilingual protocol data, it demonstrates generalization to other languages. As shown in Figure 11, we present examples in Japanese, French, and Arabic.

Animated Poster. Because CreatiPoster outputs layered results, Text-to-Video models [Sand-AI 2025] can animate the background layer, which can then be merged with the protocol layers. This approach maintains text accuracy and editability while enabling dynamic, visually engaging posters. Examples are shown in Figure 12; more can be found in the demo video.

6 CONCLUSION

CreatiPoster couples a multimodal protocol model—drafting editable text and asset layers in JSON—with a conditional background generator, turning simple user prompts into polished designs. Benchmarks show clear gains over open- and closed-source rivals, and we release code, model, a 100,000 sample dataset, and benchmark to catalyze further work. By unifying accuracy, asset fidelity, editability, and aesthetics in one extensible design stack, CreatiPoster moves AI-assisted graphic creation closer to true, democratized co-design.

REFERENCES

2025. HiDream-I1. https://github.com/HiDream-ai/HiDream-I1.

Michael Bauerly and Yili Liu. 2006. Computational modeling and experimental investigation of effects of compositional elements on interface and design aesthetics. International journal of human-computer studies 64, 8 (2006), 670–682.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. 2023. Rt-2: Vision-language-action models transfer web knowledge to robotic control. arXiv preprint arXiv:2307.15818 (2023).

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, Aijia Guo, Qipeng Guo, Conghui He, Yingfan Hu, Ting Huang, Tao Jiang, Penglong Jiao, Zhenjiang Jin, Zhikai Lei, Jiaxing Li, Jingwen Li, Linyang Li, Shuaibin Li, Wei Li, Yining Li, Hongwei Liu, Jiangning Liu, Jiawei Hong, Kaiwen Liu, Kuikun Liu, Xiaoran Liu, Chengqi Lv, Haijun Lv, Kai Lv, Li Ma, Runyuan Ma, Zerun Ma, Wenchang Ning, Linke Ouyang, Jiantao Qiu, Yuan Qu, Fukai Shang, Yunfan Shao, Demin Song, Zifan Song, Zhihao Sui, Peng Sun, Yu Sun, Huanze Tang, Bin Wang, Guoteng Wang, Jiaqi Wang, Jiayu Wang, Rui Wang, Yudong Wang, Ziyi Wang, Xingjian Wei, Qizhen Weng, Fan Wu, Yingtong Xiong, Chao Xu, Ruiliang Xu, Hang Yan, Yirong Yan, Xiaogui Yang, Haochen Ye, Huaiyuan Ying, Jia Yu, Jing Yu, Yuhang Zang, Chuyu Zhang, Li Zhang, Pan Zhang, Peng Zhang, Ruijie Zhang, Shuo Zhang, Songyang Zhang, Wenjian Zhang, Wenwei Zhang, Xingcheng Zhang, Xinyue Zhang, Hui Zhao, Qian Zhao, Xiaomeng Zhao, Fengzhe Zhou, Zaida Zhou, Jingming Zhuo, Yicheng Zou, Xipeng Qiu, Yu Qiao, and Dahua Lin. 2024. InternLM2 Technical Report. arXiv:2403.17297 [cs.CL]

[7] https://www.recraft.ai/

Shang Chai, Liansheng Zhuang, Fengying Yan, and Zihan Zhou. 2023. Two-stage Content-Aware Layout Generation for Poster Designs. In ACM MM. 8415–8423.

Haoyu Chen, Xiaojie Xu, Wenbo Li, Jingjing Ren, Tian Ye, Songhua Liu, Ying-Cong Chen, Lei Zhu, and Xinchao Wang. 2025. Posta: A go-to framework for customized artistic poster generation. arXiv preprint arXiv:2503.14908 (2025).

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. 2023. Shikra: Unleashing Multimodal LLM's Referential Dialogue Magic. arXiv preprint arXiv:2306.15195 (2023).

Yutao Cheng, Zhao Zhang, Maoke Yang, Hui Nie, Chunyuan Li, Xinglong Wu, and Jie Shao. 2025. Graphic Design with Large Multimodal Model. In AAAI.

Daniel Cohen-Or, Olga Sorkine, Ran Gal, Tommer Leyvand, and Ying-Qing Xu. 2006. Color harmonization. In ACM SIGGRAPH 2006 Papers. 624–630.

Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, Yang Zhou, Kaizhao Liang, Jintai Chen, Juanwu Lu, Zichong Yang, Kuei-Da Liao, et al. 2024. A survey on multimodal large language models for autonomous driving. In WACV. 958–979.

Xinpeng Ding, Jinahua Han, Hang Xu, Xiaodan Liang, Wei Zhang, and Xiaomeng Li. 2024. Holistic Autonomous Driving Understanding by Bird's-Eye-View Injected Multi-Modal Large Models. arXiv preprint arXiv:2401.00988 (2024).

DC Dowson and BV666017 Landau. 1982. The Fréchet distance between multivariate normal distributions. Journal of multivariate analysis 12, 3 (1982), 450–455.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. 2023. PaLM-E: An embodied multimodal language model. arXiv preprint arXiv:2303.03378 (2023).

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. 2024. Scaling rectified flow transformers for high-resolution image synthesis. arXiv preprint arXiv:2403.03206 (2024).

Yu Gao, Lixue Gong, Qiushan Guo, Xiaoxia Hou, Zhichao Lai, Fanshi Li, Liang Li, Xiaochen Lian, Chao Liao, Liyang Liu, et al. 2025. Seedream 3.0 Technical Report. arXiv preprint arXiv:2504.11346 (2025).

Yue Gao, Yuan Guo, Zhouhui Lian, Yingmin Tang, and Jianguo Xiao. 2019. Artistic glyph image synthesis via one-stage few-shot learning. ACM Transactions on Graphics (ToG) 38, 6 (2019), 1–12.

Lixue Gong, Xiaoxia Hou, Fanshi Li, Liang Li, Xiaochen Lian, Fei Liu, Liyang Liu, Wei Liu, Wei Lu, Yichun Shi, et al. 2025. Seedream 2.0: A native chinese-english bilingual image generation foundation model. arXiv preprint arXiv:2503.07703 (2025).

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. 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 CVPR. 6018–6026.

Nathan Hurst, Wilmot Li, and Kim Marriott. 2009. Review of automatic document formatting. In Proceedings of the 9th ACM symposium on Document engineering. 99–108.

Naoto Inoue, Kotaro Kikuchi, Edgar Simo-Serra, Mayu Otani, and Kota Yamaguchi. 2023. Towards Flexible Multi-modal Document Models. In CVPR. 14287–14296.

Naoto Inoue, Kento Masui, Wataru Shimoda, and Kota Yamaguchi. 2024. OpenCOLE: Towards Reproducible Automatic Graphic Design Generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 8131–8135.

Ali Jahanian, Jerry Liu, Qian Lin, Daniel Tretter, Eamonn O'Brien-Strain, Seungyon Claire Lee, Nic Lyons, and Jan Allebach. 2013. Recommendation system for automatic design of magazine covers. In Proceedings of the 2013 international conference on Intelligent user interfaces. 95–106.

Peidong Jia, Chenxuan Li, Zeyu Liu, Yichao Shen, Xingru Chen, Yuhui Yuan, Yinglin Zheng, Dong Chen, Ji Li, Xiaodong Xie, et al. 2023. COLE: A Hierarchical Generation Framework for Graphic Design. arXiv preprint arXiv:2311.16974 (2023).

Zhaoyun Jiang, Jiaqi Guo, Shizhao Sun, Huayu Deng, Zhongkai Wu, Vuksan Mijovic, Zijiang James Yang, Jian-Guang Lou, and Dongmei Zhang. 2023. LayoutFormer++: Conditional Graphic Layout Generation via Constraint Serialization and Decoding Space Restriction. In CVPR. 18403–18412.

Black Forest Labs. 2024. FLUX.1-dev. https://blackforestlabs.ai/announcing-black-forest-labs.

Chunyuan Li, Zhe Gan, Zhengyuan Yang, Jianwei Yang, Linjie Li, Lijuan Wang, and Jianfeng Gao. 2023a. Multimodal foundation models: From specialists to general-purpose assistants. arXiv preprint arXiv:2309.10020 1, 2 (2023), 2.

Fengheng Li, An Liu, Wei Feng, Honghe Zhu, Yaoyu Li, Zheng Zhang, Jingjing Lv, Xin Zhu, Junjie Shen, and Zhangang Lin. 2023c. Relation-Aware Diffusion Model for Controllable Poster Layout Generation. In Proceedings of the 32nd ACM international conference on information & knowledge management. 1249–1258.

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

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2023b. Videochat: Chat-centric video understanding. arXiv preprint arXiv:2305.06355 (2023).

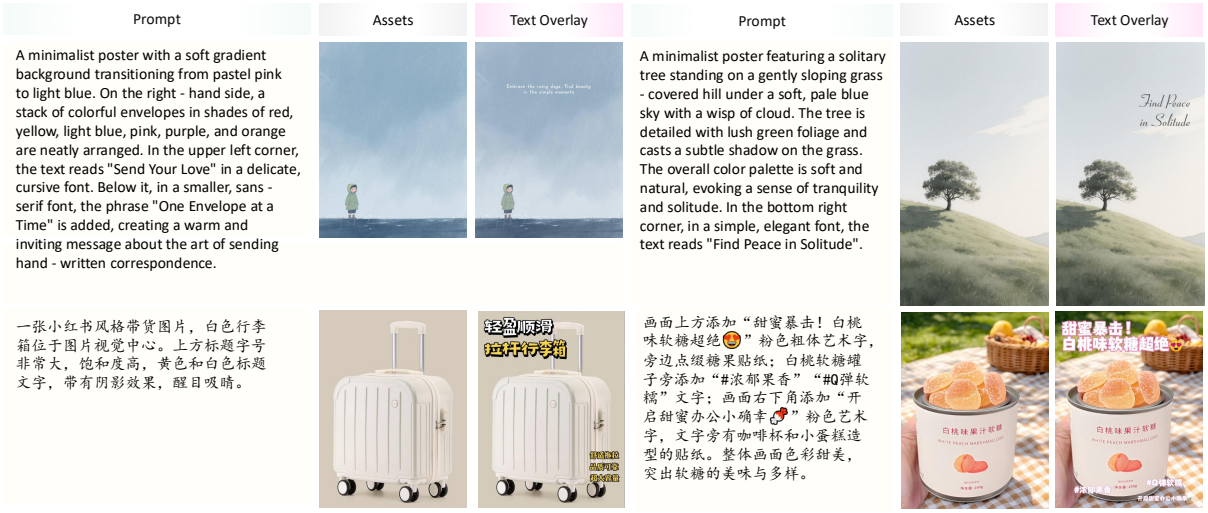


Figure 8: Graphic compositions created by overlaying text onto user-uploaded assets, which serve as the background layer (BG-Layer). The Protocol Model generates the necessary text layers based on the provided information, and the final result is produced using a rendering engine. In this mode, the background model is not required.

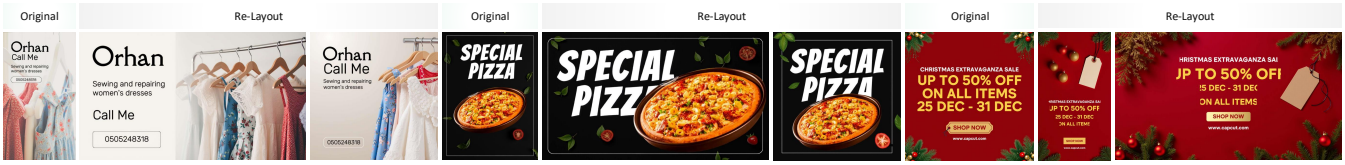


Figure 9: Given an original graphic design, CreatiPoster generates alternative layouts of various sizes while preserving content and style. By reusing rendered layers and predicting new foreground and background elements, the approach enables efficient adaptation of designs for different platforms.

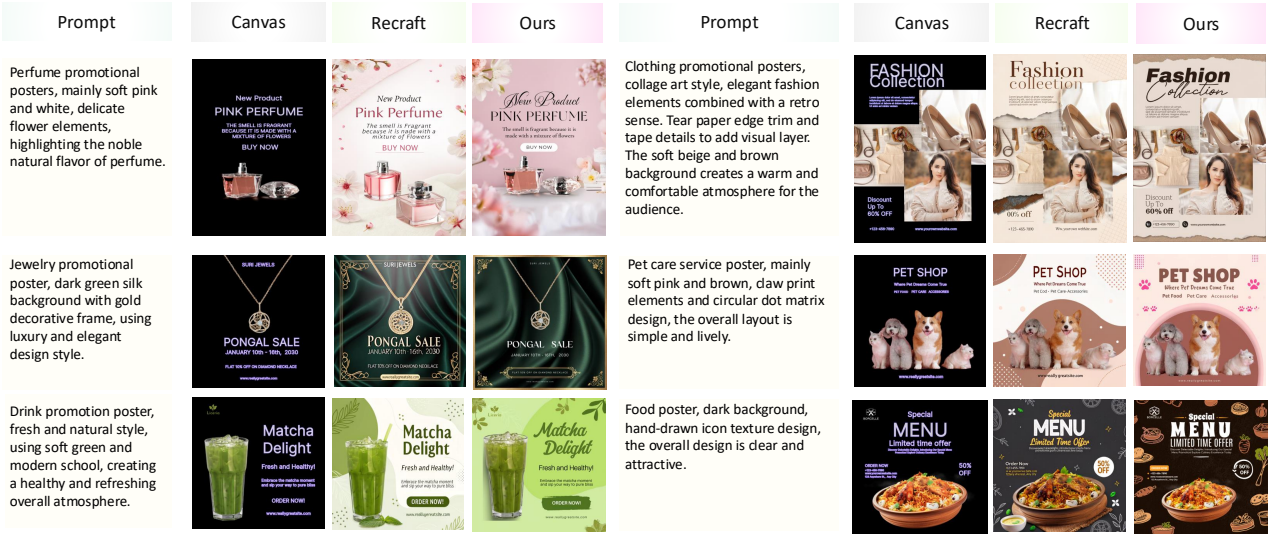


Figure 10: Graphic compositions are generated in canvas mode, allowing users to freely position text and elements and specify their attributes. The system then refines the design for improved aesthetics based on prompts. We compare our approach with Recraft, which generates single-layer posters from canvas images. In contrast to Recraft, our protocol-based method preserves text accuracy and material identity, whereas Recraft's results often show text errors and loss of material fidelity.

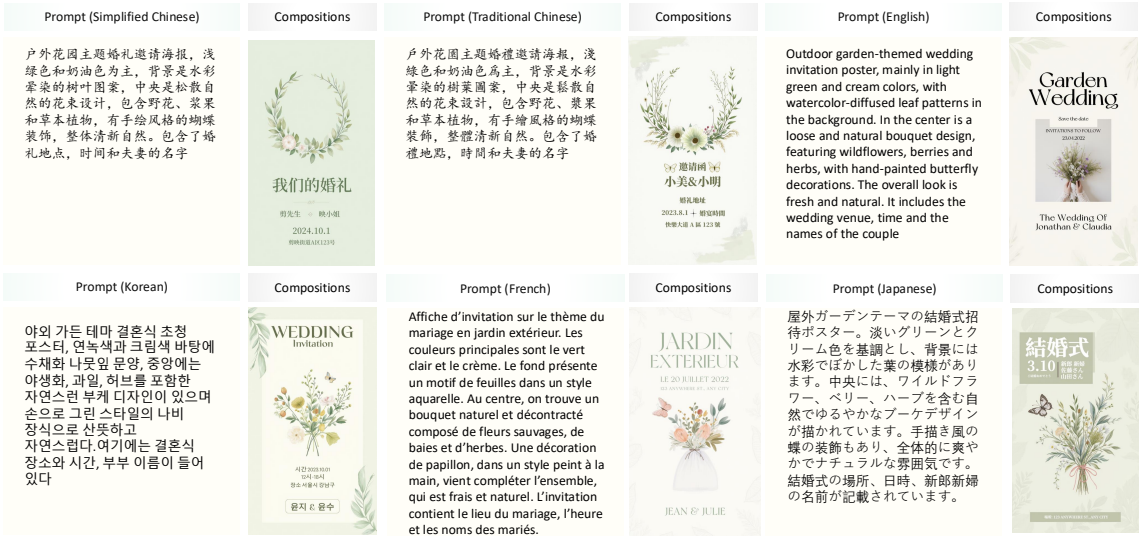


Figure 11: CreatiPoster's multilingual prediction results. Although the training data includes only Simplified Chinese and English graphic compositions, pre-training and multilingual fine-tuning enable the protocol model to generalize to other languages.



Figure 12: Extend graphic compositions generated by CreatiPoster to animated posters. The videos are generated according to background layers using image-to-video model.

Chong Mou, Yanze Wu, Wenxu Wu, Zinan Guo, Pengze Zhang, Yufeng Cheng, Yiming Luo, Fei Ding, Shiwen Zhang, Xinghui Li, et al. 2025. DreamO: A Unified Framework for Image Customization. arXiv preprint arXiv:2504.16915 (2025).

Yao Mu, Qinglong Zhang, Mengkang Hu, Wenhai Wang, Mingyu Ding, Jun Jin, Bin Wang, Jifeng Dai, Yu Qiao, and Ping Luo. 2024. Embodiedgpt: Vision-language pre-training via embodied chain of thought. NeurIPS 36 (2024).

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. 2023. Kosmos-2: Grounding Multimodal Large Language Models to the World. arXiv preprint arXiv:2306.14824 (2023).

Yifan Pu, Yiming Zhao, Zhicong Tang, Ruihong Yin, Haoxing Ye, Yuhui Yuan, Dong Chen, Jianmin Bao, Sirui Zhang, Yanbin Wang, et al. 2025. Art: Anonymous region transformer for variable multi-layer transparent image generation. arXiv preprint arXiv:2502.18364 (2025).

Sand-AI. 2025. MAGI-1: Autoregressive Video Generation at Scale. https://static.magi.world/static/files/MAGI_1.pdf

Jaejung Seol, Seojun Kim, and Jaejun Yoo. 2024. Posterllama: Bridging design ability of langauge model to contents-aware layout generation. arXiv preprint arXiv:2404.00995 (2024).

stability.ai. 2024. Stable Diffusion 3.5. https://stability.ai/news/introducing-stable-diffusion-3-5.

Sou Tabata, Hiroki Yoshihara, Haruka Maeda, and Kei Yokoyama. 2019. Automatic layout generation for graphical design magazines. In ACM SIGGRAPH 2019 Posters. 1–2.

Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. 2024. Ominicontrol: Minimal and universal control for diffusion transformer. arXiv preprint arXiv:2411.15098 (2024).

Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, et al. 2023. Video understanding with large language models: A survey. arXiv preprint arXiv:2312.17432 (2023).

Masataka Tokumaru, Noriaki Muranaka, and Shigeru Imanishi. 2002. Color design support system considering color harmony. In 2002 IEEE world congress on computational intelligence. 2002 IEEE international conference on fuzzy systems. FUZZ-IEEE'02. Proceedings (Cat. No. 02CH37291), Vol. 1. IEEE, 378–383.

Heng Wang, Yotaro Shimose, and Shingo Takamatsu. 2025. BannerAgency: Advertising Banner Design with Multimodal LLM Agents. arXiv preprint arXiv:2503.11060 (2025).

Junyang Wang, Haiyang Xu, Jiabo Ye, Ming Yan, Weizhou Shen, Ji Zhang, Fei Huang, and Jitao Sang. 2024a. Mobile-Agent: Autonomous multi-modal mobile device agent with visual perception. arXiv preprint arXiv:2401.16158 (2024).

Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. 2023. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. arXiv preprint arXiv:2305.11175 (2023).

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. 2024b. Emu3: Next-Token Prediction is All You Need. arXiv preprint arXiv:2409.18869 (2024).

Yizhi Wang, Guo Pu, Wenhan Luo, Yexin Wang, Pengfei Xiong, Hongwen Kang, and Zhouhui Lian. 2022. Aesthetic text logo synthesis via content-aware layout inferring. In CVPR. 2436–2445.

Kota Yamaguchi. 2021. Canvasvae: Learning to generate vector graphic documents. In ICCV. 5481–5489.

Xuyong Yang, Tao Mei, Ying-Qing Xu, Yong Rui, and Shipeng Li. 2016. Automatic generation of visual-textual presentation layout. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 12, 2 (2016), 1–22.

Zhao Yang, Jiaxuan Liu, Yucheng Han, Xin Chen, Zebiao Huang, Bin Fu, and Gang Yu. 2023. Appagent: Multimodal agents as smartphone users. arXiv preprint arXiv:2312.13771 (2023).

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023. A Survey on Multimodal Large Language Models. arXiv preprint arXiv:2306.13549 (2023).

Wei-Tao You, Ling-Yun Sun, Zhi-Yuan Yang, and Chang-Yuan Yang. 2019. Automatic advertising image color design incorporating a visual color analyzer. Journal of Computer Languages 55 (2019), 100910.

Ning Yu, Chia-Chih Chen, Zeyuan Chen, Rui Meng, Gang Wu, Paul Josel, Juan Carlos Niebles, Caiming Xiong, and Ran Xu. 2022. LayoutDETR: Detection Transformer Is a Good Multimodal Layout Designer. arXiv preprint arXiv:2212.09877 (2022).

Duzhen Zhang, Yahan Yu, Chenxing Li, Jiahua Dong, Dan Su, Chenhui Chu, and Dong Yu. 2024b. Mm-llms: Recent advances in multimodal large language models. arXiv preprint arXiv:2401.13601 (2024).

Hui Zhang, Dexiang Hong, Yitong Wang, Jie Shao, Xinglong Wu, Zuxuan Wu, and Yu-Gang Jiang. 2024a. Creatilayout: Siamese multimodal diffusion transformer for creative layout-to-image generation. arXiv preprint arXiv:2412.03859 (2024).

Hang Zhang, Xin Li, and Lidong Bing. 2023b. Video-llama: An instruction-tuned audio-visual language model for video understanding. arXiv preprint arXiv:2306.02858 (2023).

Junyi Zhang, Jiaqi Guo, Shizhao Sun, Jian-Guang Lou, and Dongmei Zhang. 2023a. LayoutDiffusion: Improving Graphic Layout Generation by Discrete Diffusion Probabilistic Models. In ICCV.

Lvmin Zhang and Maneesh Agrawala. 2024. Transparent Image Layer Diffusion using Latent Transparency. ACM Transactions on Graphics (TOG) 43, 4 (2024), 1–15.

Yuxuan Zhang, Yirui Yuan, Yiren Song, Haofan Wang, and Jiaming Liu. 2025. Easycontrol: Adding efficient and flexible control for diffusion transformer. arXiv preprint arXiv:2503.07027 (2025).

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. 2024. Transfusion: Predict the Next Token and Diffuse Images with One Multi-Modal Model. https://api.semanticscholar.org/CorpusID:271909855

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 (2022).