CreatiDesign: A Unified Multi-Conditional Diffusion Transformer for Creative Graphic Design

Anonymous authors
Paper under double-blind review

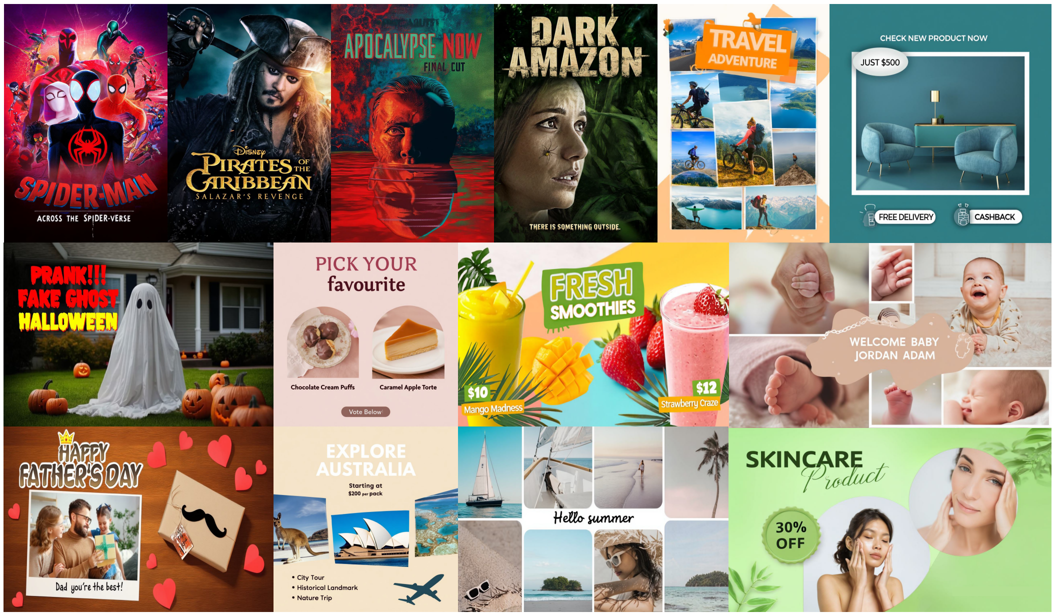

Figure 1: CreatiDesign generates high-quality graphic designs based on user-provided image assets and semantic layouts, covering a wide range of categories such as movie posters, brand promotions, product advertisements, and social media content.

Abstract

Graphic design plays a vital role in visual communication across advertising, marketing, and multimedia entertainment. Prior work has explored automated graphic design generation using diffusion models, aiming to streamline creative workflows and democratize design capabilities. However, complex graphic design scenarios require accurately adhering to design intent specified by multiple heterogeneous user-provided elements (*e.g.* images, layouts, and texts), which pose multi-condition control challenges for existing methods. Specifically, previous single-condition control models demonstrate effectiveness only within their specialized domains but fail to generalize to other conditions, while existing multi-condition methods often lack fine-grained control over each sub-condition and compromise overall compositional harmony. To address these limitations, we introduce CreatiDesign, a systematic solution for automated graphic design covering both model architecture and dataset construction. First, we design a unified multi-condition driven architecture that enables flexible and precise integration of heterogeneous design elements with minimal architectural modifications to the base diffusion model. Furthermore, to ensure that each condition precisely controls its designated image region and to avoid interference between conditions, we propose a multimodal attention mask mechanism. Additionally, we develop a fully automated pipeline for constructing graphic design datasets, and introduce a new dataset with 400K samples featuring multi-condition annotations, along with a comprehensive benchmark. Experimental results show that CreatiDesign outperforms existing models by a clear margin in faithfully adhering to user intent.

1 Introduction

Graphic design (Jobling & Crowley, 1996) is a fundamental vehicle for visual communication, affective perception, and brand identity across advertising, marketing, and multimedia entertainment.

Recently, diffusion models (Ho et al., 2020; Dhariwal & Nichol, 2021) have achieved remarkable advances, especially in text-to-image generation (stability.ai, 2024; Labs, 2024), which can produce visually compelling and semantically rich images. Leveraging these models to automate graphic design—thereby streamlining creative workflows and democratizing design capabilities—has attracted increasing attention (Gao et al., 2025a; Chen et al., 2025; Peng et al., 2025).

However, as illustrated in Figure 2, graphic design generation poses unique challenges because it requires the precise control and harmonious arrangement of multiple heterogeneous elements, typically comprising three categories: I) **Primary visual elements**, which act as visual focal points and convey the central theme (*e.g.* product subjects, provided in image format); II) **Secondary visual elements**, which offer contextual support and enrich the composition (*e.g.* decorative objects, specified by semantic description and position in a layout); and III) **Textual elements**, which directly convey essential information (*e.g.* slogans or product names, also provided as layout). This multi-element nature introduces multi-condition control requirements for diffusion models, as it demands both semantic and spatial fidelity to users' design intent.

While several works have explored unleashing the potential of diffusion models for automatic graphic design generation, three major challenges remain unresolved: I) **How to integrate multiple heterogeneous conditions in a unified manner.** Previous expert models are typically tailored for only a single type of condition, and often fail to follow other conditions. As illustrated in Figure 2, image-driven models Wang et al. (2024b); Wu et al. (2025); Labs (2025) focus exclusively on aligning with primary visual elements, whereas layout-driven models Peng et al. (2025); Ma et al. (2025b); Zhang et al. (2024) are limited to following the semantic descriptions and spatial arrangements of secondary visual or textual elements. Such biased capability often leads to reduced fidelity to user intent, as highlighted by the red and purple masks. II) **How to preserve fine-grained controllability for each condition while achieving harmonious compositions.** Existing multi-condition approaches (Xiao et al., 2024; Gao et al., 2025a; goo, 2025; ope, 2025) lack accurate control over each sub-condition and fail to effectively coordinate all elements, resulting in outputs that do not faithfully reflect user design intent. III) **How to construct large-scale, multi-element graphic design datasets in an automated manner.** Ready-to-use graphic design datasets with fine-grained, multi-condition annotations remain scarce, which naturally prevents models from learning design capabilities.

To this end, we propose CreatiDesign, a systematic solution for intelligent graphic design generation that addresses the aforementioned challenges through the following components: I) **Unified multi-condition driven architecture.** CreatiDesign preserves the strong generative capabilities of text-to-image diffusion models while unlocking their potential for graphic design with minimal architectural modifications. Specifically, the native image encoder embeds the multi-subject image condition into the latent space, while the semantic layout is processed by extracting textual features with the text encoder and fusing them with positional information. After encoding all modalities into a unified feature space, native multimodal attention (MM-Attention) is applied to enable deep integration and interaction across modalities. This allows for unified and flexible multi-condition control over the generated content. II) **Efficient Multi-Condition Coordination.** To ensure that each heterogeneous condition precisely controls its designated image regions and to avoid mutual interference that could compromise the unique characteristics of each condition, we introduce carefully designed attention masks to regulate the interaction scope of each modality within the multimodal attention mechanism. This design enables each condition to independently and efficiently control its target region, while maintaining high overall compositional harmony. III) **Automated Dataset Construction Pipeline.** We develop a fully automated pipeline for constructing graphic design datasets. This pipeline consists of design theme generation and rendering, conditional image generation, and multi-element annotation and filtering. As a result, we construct a training dataset containing 400K design samples with multi-condition annotations, along with a comprehensive benchmark for rigorous evaluation.

2 Related Work

2.1 Text-to-Image Generation

Text-to-image (T2I) generation (Rombach et al., 2022; Podell et al., 2024; Saharia et al., 2022; Chen et al., 2024c; Li et al., 2024) aims to generate visual content from textual descriptions, and has

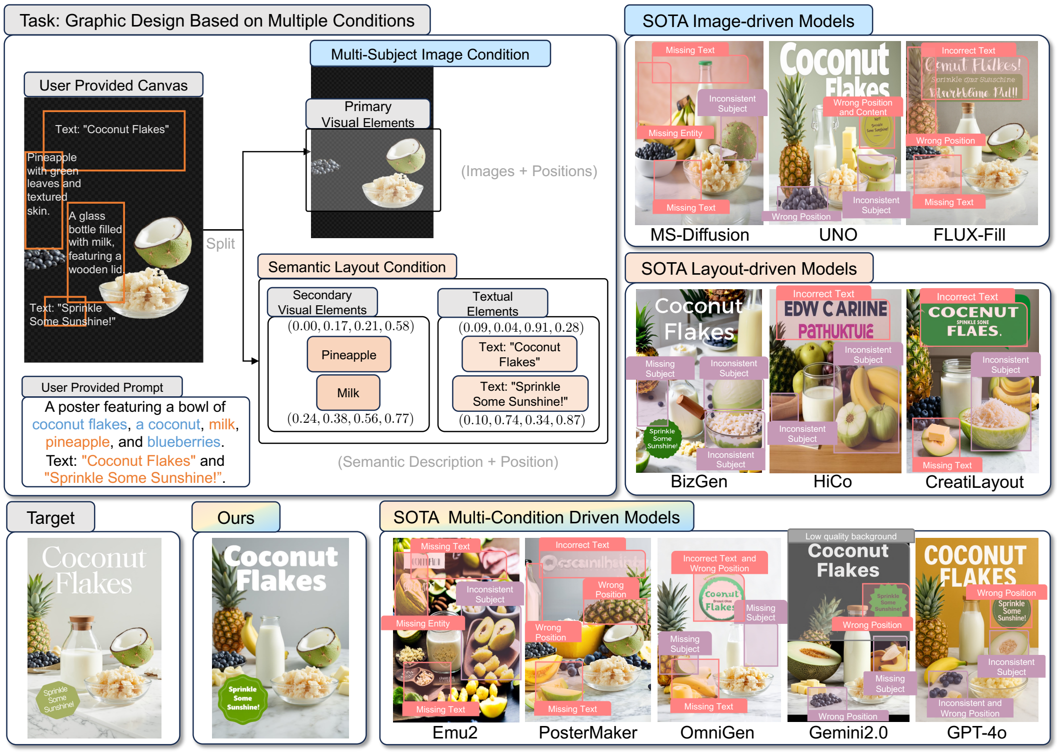
Task: Graphic Design Based on Multiple Conditions
User Provided Canvas
Text: "Coconut Flakes"
Pineapple with green leaves and textured skin.
A glass bottle filled with milk, featuring a wooden lid
Text: "Sprinkle Some Sunshine!"
Split
Multi-Subject Image Condition
Primary Visual Elements
(Images + Positions)
Semantic Layout Condition
Secondary Visual Elements
(0.00, 0.17, 0.21, 0.58)
Pineapple
Milk
(0.24, 0.38, 0.56, 0.77)
Textual Elements
(0.09, 0.04, 0.91, 0.28)
Text: "Coconut Flakes"
Text: "Sprinkle Some Sunshine!"
(0.10, 0.74, 0.34, 0.87)
(Semantic Description + Position)
User Provided Prompt
A poster featuring a bowl of coconut flakes, a coconut, milk, pineapple, and blueberries. Text: "Coconut Flakes" and "Sprinkle Some Sunshine!".
SOTA Image-driven Models
Missing Text
Inconsistent Subject
Missing Entity
Wrong Position and Content
Wrong Position
Incorrect Text
MS-Diffusion
UNO
FLUX-Fill
SOTA Layout-driven Models
Missing Subject
Inconsistent Subject
BizGen
HiCo
CreatiLayout
Target
Ours
SOTA Multi-Condition Driven Models
Incorrect Text and Wrong Position
Low quality background
Inconsistent and Wrong Position
Emu2
PosterMaker
OmniGen
Gemini2.0
GPT-4o

Figure 2: **An overview of our motivation**. Graphic design is a multi-condition driven generation task that requires the precise and harmonious arrangement of heterogeneous elements, including primary visual elements (provided as images with positions), as well as secondary visual and textual elements (both specified by semantic descriptions and positions). Previous methods either support only a single type of condition (*e.g.* image-driven or layout-driven models) or lack accurate control over each sub-condition(*e.g.* multi-condition driven models), resulting in failure to strictly adhere to user design intent, as highlighted by the red and purple masks.

achieved remarkable progress in both visual quality and semantic alignment. Recent advances, such as SD3 series (Esser et al., 2024; stability.ai, 2024), CogView4 (THU, 2025), FLUX.1 (Labs, 2024), HiDream (HiD, 2025), and Seedream series (Gong et al., 2025; Gao et al., 2025b), have pushed the frontier further by leveraging Multimodal Diffusion Transformer architectures (MM-DiT). Despite these advances, existing T2I models still struggle with fine-grained controllability, particularly in scenarios where users wish to specify precise subject identities or detailed compositional layouts.

2.2 Controllable Image Generation

To achieve precise control, a variety of conditional image generation paradigms have been proposed, including subject-driven (Ruiz et al., 2023; Cai et al., 2024; Tan et al., 2024; Shin et al., 2024; Zhu et al., 2025; Labs, 2025; Wu et al., 2025; Wang et al., 2024b), layout-driven (Li et al., 2023; Wang et al., 2024c; Zhou et al., 2024; Feng et al., 2024; Zhang et al., 2024; Peng et al., 2025; Zhou et al., 2025; Ma et al., 2025b), and so on. These expert models excel at controlling specific conditions—such as preserving the visual characteristics of the provided subjects or adhering to layout specifications—but often fail to follow other conditions. In response, multi-condition driven frameworks (Sun et al., 2024; Xiao et al., 2024; goo, 2025; ope, 2025; Wang et al., 2025a; Qin et al., 2023; Hu et al., 2023; Zhao et al., 2023; Ran et al., 2024) have been introduced to jointly handle heterogeneous user-provided conditions. However, these unified approaches often lack accurate control over each sub-condition.

2.3 Automatic Graphic Design

Several works (Gao et al., 2025a; Wang et al., 2025b; Chen et al., 2025; Pu et al., 2025; Ma et al., 2025a; Wang et al., 2024a; Liu et al., 2024; Chen et al., 2024b; Tuo et al., 2024) have attempted to automate graphic design generation, aiming to streamline creative workflows and democratize design capabilities. However, automatic graphic design introduces distinct challenges beyond general text-to-image or controllable image generation, requiring models to precisely preserve user-specified

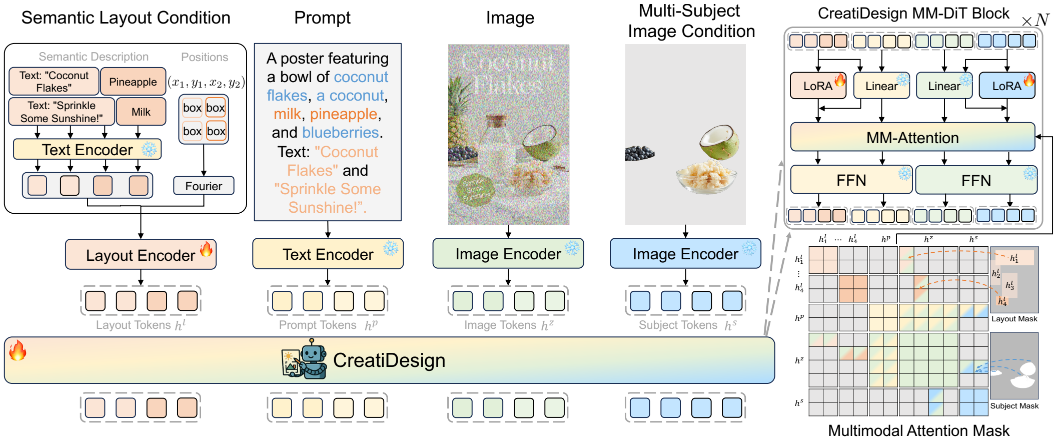

Figure 3: **An overview of the architecture.** CreatiDesign integrates subjects and semantic layout conditions through native multimodal attention. Multimodal attention mask ensures that each condition precisely controls its designated image regions while preventing leakage between conditions.

subjects, align secondary visual and textual elements with detailed semantic and spatial constraints, and maintain overall visual coherence. Despite recent progress, most existing methods struggle to meet all these demands simultaneously. This underscores the need for a unified, highly controllable, and harmonious solution, which is exactly the goal of this paper.

3 Method

3.1 Problem Formulation

This paper focuses on the task of graphic design generation, where each design typically comprises multiple heterogeneous elements provided by the user, such as primary visual elements, secondary visual elements, and textual elements, as illustrated in Figure 2. The key challenge is to accurately and harmoniously integrate these user-specified elements—each representing distinct aspects of user intent—into the generated image. Formally, the task can be defined as: $I_g = f(P, I_s, L)$, where I_g denotes the generated image, P is the global prompt describing the overall image, I_s represents the multi-subject image condition (*i.e.*, a set of primary visual elements). L denotes the semantic layout condition, which consists of n elements, partitioned into two categories: secondary visual elements and textual elements. Each layout element is defined by a pair (d_i, b_i), where d_i is the semantic description and b_i is the spatial position (bounding box), formally expressed as:

$$L = \{l_i = (d_i, b_i)\}_{i=0}^{n}, \quad l_i \in \{\text{secondary visual element, textual element}\}. \tag{1}$$

In the following sections, we will introduce the key parts of CreatiDesign in detail.

3.2 Unified Multi-Condition Driven Architecture

In MM-DiT-based text-to-image models (*e.g.* FLUX.1 (Labs, 2024)), a text encoder (*e.g.* T5 (Raffel et al., 2020)) is employed to tokenize and encode the input prompt into a sequence of text tokens, denoted as h_p. Concurrently, an image encoder (*e.g.* VAE (Kingma, 2013)) is utilized to encode the ground-truth image into a latent representation z, which is subsequently partitioned into patches to obtain image tokens, denoted as h_z. These text and image tokens are then fed into MM-Attention, which facilitates rich interactions between the textual and visual modalities, thereby enabling precise control over the image content. Our approach aims to retain the strong capabilities of T2I models while unlocking their potential for graphic design with minimal architectural modifications, as illustrated in Figure 3.

Tokenize Multi-Subject Image Condition. We first pad the multi-subject image condition with a background color (*e.g.* gray) and encode it using the native VAE. The encoded latent representation is then partitioned into patches to obtain the subject tokens h_s.

Tokenize Semantic Layout Condition. For each element $l_i = (d_i, b_i)$ in the semantic layout condition, we utilize the native T5 text encoder to extract the semantic feature h_i^d from d_i. For the bounding box b_i, we apply Fourier positional encoding (Mildenhall et al., 2021; Li et al., 2023) to obtain the spatial feature h_i^b. The final layout token h_i^l is obtained by concatenating h_i^d and h_i^b along the feature dimension, followed by a layout encoder (*i.e.* MLP): $h_i^l = \mathrm{MLP}(\mathrm{Concat}(h_i^d, h_i^b))$. In this way, layout tokens integrate semantic and spatial information.

Integrate Multi-Condition. After encoding the prompt, noise image, multi-subject image condition, and semantic layout condition into tokens, denoted as h^p, h^z, h^s, h^l, we concatenate them along the token dimension and feed the token sequence into a stack of MM-DiT Blocks. Each Block consists of linear projection layers (for Q, K, V), multimodal attention (MM-Attention), and feed-forward networks (FFN). Each type of tokens is linearly projected into its corresponding query, key, and value spaces: $Q^*, K^*, V^* = \mathrm{Linear}(h^*)$, where $*$ denotes the modality (layout, prompt, image, or subject). For the layout tokens h^l and subject tokens h^s, we further adapt their representations using LoRA modules deployed on the linear layer and adaptive layer normalization (AdaLN), enabling efficient fine-tuning and alignment. The multimodal attention is then computed as:

$$h^{l\prime}, h^{p\prime}, h^{z\prime}, h^{s\prime} = \mathrm{Attention}([\mathbf{Q}^l, \mathbf{Q}^p, \mathbf{Q}^z, \mathbf{Q}^s], [\mathbf{K}^l, \mathbf{K}^p, \mathbf{K}^z, \mathbf{K}^s], [\mathbf{V}^l, \mathbf{V}^p, \mathbf{V}^z, \mathbf{V}^s]). \quad (2)$$

This design enables multiple conditions to control the image content. To avoid positional embedding conflicts, such as between the noise image and image condition, or between the prompt and layout condition, we adopt positional encoding shifts to the image and layout condition tokens (Tan et al., 2024) to ensure clear separation in the token space. Overall, this architecture empowers the text-to-image model with multi-condition control capabilities through minimal architectural modifications.

3.3 Collaborative Multi-Condition Control

Multi-condition driven methods may suffer from degraded controllability over each sub-condition. We attribute this to the fact that the sub-condition is not precisely bound to its corresponding image region and that there is semantic leakage among sub-conditions. To address this, we introduce a multimodal attention mask within our architecture, consisting of a layout mask and a subject mask.

Layout Attention Mask. Given the user-specified bounding box b_i for each semantic description d_i, we can precisely locate the target image region. Inspired by (Chen et al., 2024a), we construct a layout mask such that each layout token h_i^l is only allowed to attend to and be attended by the image tokens h_i^z within its corresponding bounding box. This explicit attention modulation enhances the spatial controllability. Furthermore, we block interactions among layout tokens themselves, between layout tokens and subject tokens, and between layout tokens and prompt tokens, to prevent semantic leakage and to ensure that each layout token retains its unique characteristics.

Subject Attention Mask. Based on the user-provided multi-subject image, we extract the spatial location of each subject to form a subject mask. Each subject token h_i^s is only permitted to interact bidirectionally with the image tokens h_i^z within its own mask region, thereby achieving precise subject injection. In addition, to preserve the integrity and distinctive features of the subject token h^s, we block its interactions with all irrelevant tokens, including layout tokens h^l, prompt tokens h^p, and image tokens outside the target region of h^s.

With the proposed multimodal attention masks, CreatiDesign allows each condition to precisely and independently control its targeted image region without semantic leakage, thereby producing controllable and harmonious graphic designs that closely match user intent.

4 Graphic Design Datasets and Benchmark

4.1 Graphic Design Datasets

We propose a fully automatic dataset construction pipeline, as shown in Figure 4, to address the scarcity of graphic design datasets with fine-grained, multi-condition annotations.

Design Theme Generation. Based on a design keywords bank covering common graphic design elements (*e.g.* furniture, food, clothing *etc.*), we prompt a large language model (LLM, *e.g.* GPT-4) to act as a professional designer and generate design themes that include descriptions of primary visual elements, secondary visual elements, and textual elements.

Text Layer Rendering. Based on the design theme, we follow the Hierarchical Layout Generation (Cheng et al., 2025) (HLG) paradigm to generate a layout protocol of textual elements and a detailed background description. A rendering engine then converts the layout protocol into an RGBA image with accurately positioned foreground text.

Foreground-based Image Generation. To generate a visually coherent graphic design image, we draw inspiration from LayerDiffuse (Zhang & Agrawala, 2024) and develop a foreground-conditioned image generation model. Here, the RGBA text layer serves as the foreground, while the background is generated based on the aforementioned description. Specifically, we incorporate foreground-LoRA

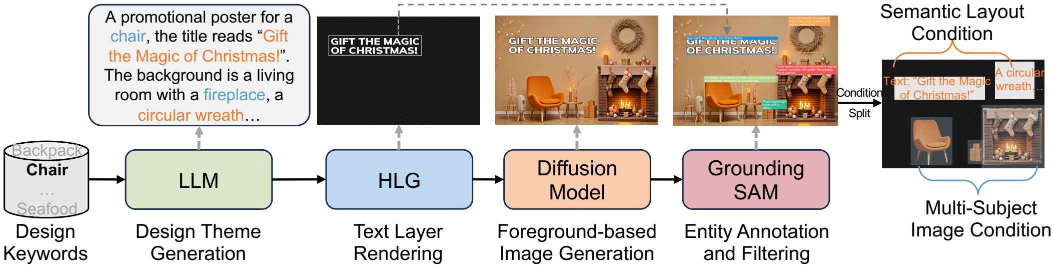
A promotional poster for a chair, the title reads "Gift the Magic of Christmas!". The background is a living room with a fireplace, a circular wreath…
GIFT THE MAGIC OF CHRISTMAS!
Semantic Layout Condition
Text: "Gift the Magic of Christmas!"
A circular wreath…
Condition
Split
Backpack
Chair
…
Seafood
LLM
HLG
Diffusion Model
Grounding SAM
Design Keywords
Design Theme Generation
Text Layer Rendering
Foreground-based Image Generation
Entity Annotation and Filtering
Multi-Subject Image Condition

Figure 4: Automated pipeline for graphic design dataset construction.

and background-LoRA modules into FLUX.1-dev and employ attention sharing to ensure seamless integration of foreground and background elements.

Entity Annotation. We use GroundingSAM2 (ide, 2024) to obtain bounding boxes and segmentation masks for all entities in the generated image. A vision-language model (OpenBMB, 2024) (VLM) is then employed to generate fine-grained descriptions for each entity. Entities are categorized as either primary or secondary visual elements. All primary visual elements are aggregated to form the multi-subject image condition, while secondary visual elements, together with the textual elements from the layout protocol, constitute the semantic layout condition.

Based on this automatic data construction pipeline, we synthesize graphic design samples at scale, alleviating the data bottleneck for model training. As a result, we construct a new dataset of 400K samples with annotations for various conditions.

4.2 Graphic Design Benchmark

To comprehensively evaluate graphic design generation under multiple conditions, we further construct a rigorous benchmark consisting of 1,000 carefully curated samples. This benchmark is designed to assess whether the generated results faithfully align with user intent—a critical requirement in practical graphic design scenarios.

The evaluation focuses on two key aspects, each with dedicated metrics: I) **Multi-Subject Preservation.** When given multi-subject image conditions (*i.e.* primary visual elements), it is crucial to strictly preserve the unique characteristics of each subject in the generated image. To quantify this, we measure the similarity between each subject and its corresponding region (obtained via bounding box priors or detected by GroundingDINO (Liu et al., 2023)) in the generated image using both CLIP (Radford et al., 2021) similarity (CLIP-I) and DINO (Oquab et al., 2023) similarity (DINO-I) scores. We further aggregate the DINO scores of all subjects by multiplication, denoted as M-DINO (Wang et al., 2024b). Unlike averaging, M-DINO is more sensitive to the failure of any single subject, providing a stricter assessment of subject preservation. II) **Semantic Layout Alignment.** For the semantic layout condition, specifying the positions and attributes of secondary visual elements and textual content, we assess alignment at spatial and semantic levels. For secondary visual elements, we employ a vision-language model in a Visual Question Answering manner to assess the spatial, color, textual, and shape attributes of each entity in the generated image (Zhang et al., 2024; Wu et al., 2024). For textual elements, we use PaddleOCR (pad, 2025) to detect text and calculate sentence accuracy (Sen. Acc), normalized edit distance (NED; *i.e.* 1 minus the edit distance) (Gao et al., 2025a), and IoU (spatial score) between detected and ground-truth texts.

5 Experiments

5.1 Experimental Setup

Dataset. We train our models on the 400K synthetic graphic design samples described in Section 4.1. The validation set contains 1,000 samples, covering diverse numbers of primary visual subjects and semantic layout annotations, enabling thorough evaluation of multi-condition controllability.

Evaluation Metrics. As described in Section 4.2, we evaluate model performance from two perspectives—multi-subject preservation and semantic layout alignment—to assess whether the generated designs accurately fulfill user intent. Additionally, to evaluate overall image quality, we report IR Score (Xu et al., 2023) and PickScore (Kirstain et al., 2023), which jointly capture prompt adherence, visual appeal, and compositional harmony across the entire image.

Implementation Details. We fine-tune FLUX.1-dev using LoRA with 256 rank, introducing 491.5M extra parameters (4.1% of FLUX's 12B). We employ the AdamW optimizer with a fixed

Table 1: **Quantitative Results**. We compare CreatiDesign with three types of previous SOTA models: multi-subject image-driven models, semantic layout-driven models, and multi-condition driven models. The best results are shown in **bold**, and the top-3 results are highlighted. Our proposed method significantly enhances the graphic design capabilities of the baseline, achieves top-tier performance across all metrics, and shows a clear lead in average score.

	Multi-Subject Preservation			**Semantic Layout Alignment**							**Image Quality**		**Avg.**
	Primary Visual Elements			Secondary Visual Elements				Textual Elements					
	CLIP-I	DINO-I	M-DINO	Spatial	Color	Textual	Shape	Spatial	Sen. Acc	NED	IR	PickScore	
UNO	77.97	47.88	20.79	53.10	43.44	42.62	41.30	11.47	40.87	74.51	**61.06**	**21.67**	44.72
MS-Diffusion	84.75	74.13	44.34	49.54	33.37	34.89	34.79	1.01	0.00	10.21	46.64	21.03	36.23
FLUX.1-Fill	**90.79**	87.32	69.05	67.55	57.48	56.26	55.75	12.48	12.07	56.69	40.74	20.71	52.24
CreatiLayout	78.41	55.54	25.31	77.42	63.07	62.67	60.21	18.59	12.27	74.21	59.92	21.04	50.72
HiCo	72.45	34.47	11.59	79.69	61.48	61.17	60.17	1.01	0.00	14.98	-36.16	19.58	31.70
BizGen	79.86	53.08	22.93	**79.84**	62.96	62.86	61.01	50.44	75.89	94.61	37.43	21.48	58.53
AnyText2	74.68	34.86	12.22	36.63	27.58	27.16	26.53	53.95	9.56	48.26	-25.95	20.24	28.81
Emu2	73.96	45.17	19.92	60.81	45.37	46.20	44.06	0.20	0.00	13.81	-1.84	20.18	30.65
PosterMaker	90.45	**87.72**	**69.56**	56.36	45.37	44.25	41.61	28.42	0.70	28.62	31.62	20.31	45.42
OmniGen	82.15	58.83	30.86	53.92	44.35	44.46	41.40	8.24	6.72	49.12	22.94	20.62	38.63
Gemini2.0	81.46	57.23	29.68	59.41	52.29	52.49	50.36	16.60	71.38	88.71	28.52	21.23	50.78
FLUX.1-dev	75.93	44.59	17.76	60.02	47.1	46.19	44.76	13.25	57.95	81.52	59.45	21.48	47.50
CreatiDesign	89.39	86.48	65.75	78.94	**66.02**	**66.94**	**65.82**	**56.90**	**78.30**	**94.68**	60.02	21.49	**69.28**
vs. Baseline	**+13.46**	**+41.89**	**+47.99**	**+18.92**	**+18.92**	**+20.75**	**+21.06**	**+43.65**	**+20.35**	**+13.16**	**+1.17**	**+0.01**	**+21.78**

learning rate of 1e-4, training for 100,000 steps with a batch size of 8 on 8 H20-96G GPUs over 4 days. We adopt a resolution bucketing strategy during training to support variable image sizes. The image condition is set to half the target image size; each layout description is capped at 30 tokens, with up to 10 layouts per image.

5.2 Comparison with Prior Works

Baseline Methods. We compare CreatiDesign with three types of previous SOTA models: multi-subject image-driven models (Wu et al., 2025; Wang et al., 2024b; Labs, 2025), semantic layout-driven models (Zhang et al., 2024; Ma et al., 2025b; Peng et al., 2025; Tuo et al., 2024), and multi-condition driven models (Sun et al., 2024; goo, 2025; Xiao et al., 2024; Gao et al., 2025a).

Quantitative Comparison. As shown in Table 1, specialist models excel primarily on their targeted control conditions—image-driven models can preserve subjects, while layout-driven models can follow semantic layout control—but perform poorly when handling other conditions. Conversely, previous multi-condition models often lack fine-grained control over each sub-condition, resulting in lower subject preservation and semantic alignment. In contrast, CreatiDesign achieves precise, balanced control across all conditions, as reflected in its top-tier performance across every sub-condition and clear lead in average scores. Remarkably, this advanced graphic design capability is achieved with minor architectural modifications to the base model FLUX.1-dev and only 4.1% extra parameters were introduced, demonstrating both effectiveness and efficiency.

Qualitative Comparison. To further illustrate the advantages of CreatiDesign, Figure 7 presents qualitative comparisons on challenging cases with multiple subjects and complex layouts. Existing SOTA methods—including multi-condition driven models and single-condition experts—consistently fall short in faithfully fulfilling user intent. Previous multi-condition models exhibit limited precision in controlling sub-conditions, resulting in misplaced or inconsistent subjects (highlighted by purple masks), as well as content or spatial misalignment in the layout (highlighted by red masks). Layout-driven models like BizGen (Peng et al., 2025) can follow the layout but struggle with subject consistency. Image-driven models such as FLUX.1-Fill (Labs, 2025) can preserve primary elements but often misplace or incorrectly render textual elements. In contrast, CreatiDesign consistently preserves the identity and position of all primary subjects, precisely aligns secondary and textual elements within the layout, and ensures overall compositional harmony.

User Study. To comprehensively assess the practical effectiveness of CreatiDesign, we conducted a user study involving feedback from both professional designers and general users. Specifically, we solicited 50 evaluation reports on 30 diverse graphic design samples, comparing our method with several state-of-the-art baselines. As illustrated in Figure 5, participants rated the generated designs on a scale of 1 to 10 across multiple criteria, including adherence to multi-subject image conditions (position accuracy and subject preservation), alignment with semantic layout conditions (position accuracy, attribute accuracy and text accuracy), and overall perceptual quality (prompt following and visual quality). The statistical results demonstrate that CreatiDesign outperforms previous methods in fine-grained controllability and overall visual appeal, delivering superior user satisfaction in real-world graphic design scenarios.

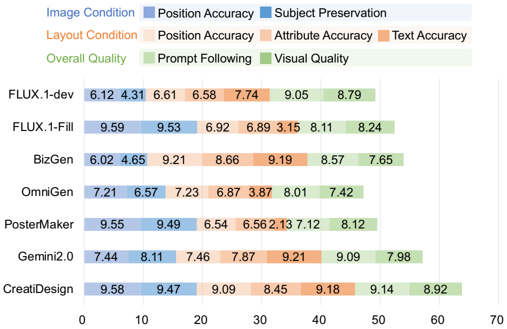

Figure 5: **User Study.** CreatiDesign demonstrates top-tier performance in preserving multi-subject characteristics, strictly following semantic layout conditions, and achieving high overall image quality.

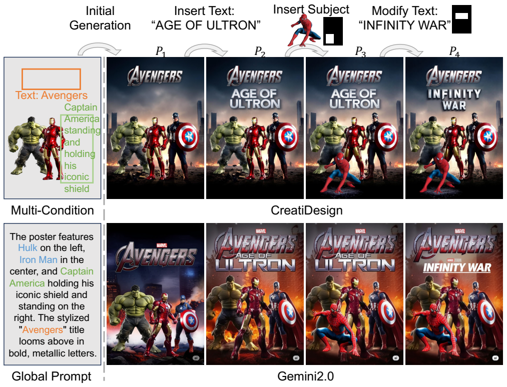
Initial Generation
Insert Text: "AGE OF ULTRON"
Insert Subject
Modify Text: "INFINITY WAR"
P_1
P_2
P_3
P_4
Text: Avengers
Captain America standing and holding his iconic shield
Multi-Condition
CreatiDesign
The poster features Hulk on the left, Iron Man in the center, and Captain America holding his iconic shield and standing on the right. The stylized "Avengers" title looms above in bold, metallic letters.
Global Prompt
Gemini2.0
AVENGERS
AGE OF ULTRON
INFINITY WAR

Figure 6: **Comparison on Loop Editing.** CreatiDesign precisely follows editing commands and maintains high consistency in non-edited areas. In contrast, Gemini2.0 frequently introduces unwanted attribute changes to subjects or text.

5.3 Free Lunch: Expanding to Editing Tasks

As illustrated in Figure 6, CreatiDesign naturally extends beyond graphic design to a wide range of editing tasks without extra retraining. We demonstrate this capability via editing a series of movie posters. Initially, the user provides a global prompt, a multi-subject image condition (*e.g*. Hulk and Iron Man), and a semantic layout specifying elements and their spatial positions (*e.g*. Captain America and the "Avengers" title). CreatiDesign generates a high-quality poster P_1 that precisely adheres to these controls. Subsequently, a sequence of editing operations is performed: first, by treating the previously generated poster P_1 as the new image condition and introducing a new text element "AGE OF ULTRON" with its desired position, CreatiDesign seamlessly inserts this subtitle to produce P_2; Next, by combining the Spider-Man image and its insertion mask with P_2 as the image condition, CreatiDesign generates P_3, achieving seamless integration of the new subject while preserving character fidelity and overall visual harmony; finally, by combining P_3 with the mask of the edited region as the image condition, the subtitle is modified to "INFINITY WAR" (P_4). Throughout these editing processes, CreatiDesign consistently maintains subject identity, achieves accuracy layout control and overall visual harmony. In contrast, strong baselines such as Gemini2.0 frequently fail to preserve non-edited regions during sequential edits, often resulting in unwanted attribute changes to subjects or text, highlighting a lack of strict adherence to user intent.

5.4 Ablation Study

Table 2 and Figure 8 evaluate the contributions of the three key components—Layout Encoder (LE), Layout Attention Mask (LAM), and Subject Attention Mask (SAM)—to the performance of CreativeDesign from quantitative and qualitative perspectives, respectively. The LE fuses the semantic features of the textual description with Fourier-encoded positional features and further aligns them into layout tokens; removing LE leads to a clear drop in the accuracy of generated text. The layout attention mask enables fine-grained spatial control by explicitly restricting each layout element to modulate only its designated image region and preventing semantic leakage across layout elements; removing LAM leads to imprecise placement of elements and increased confusion across different layout regions, as demonstrated by the decrease in spatial alignment and attribute accuracy. Similarly, the subject attention mask ensures that each subject token only interacts with its corresponding image region and blocks interference from global prompts and layout conditions. Without SAM, we observe

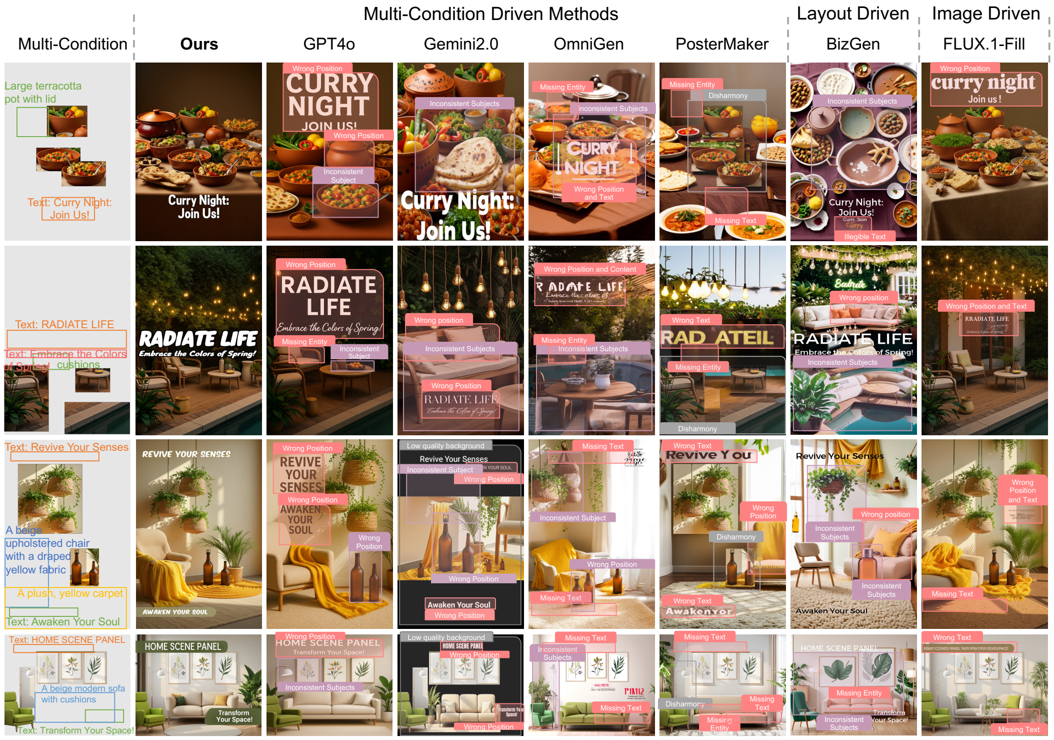

Figure 7: **Quantitative Results.** Compared with previous multi-condition or single-condition models, CreatiDesign demonstrates stricter adherence to user intent, including high subject preservation and precise layout alignment. Purple masks: inconsistent or mispositioned subjects. Red mask: entities with incorrect semantics or locations. Gray mask: disharmonious background or foreground regions.

the degradation in subject consistency, such as the incorrect digits on clocks or altered popcorn color. These results validate the effectiveness of each component in achieving faithful and controllable graphic design generation.

Table 2: Ablation study: quantitative analysis of key components in CreatiDesign.

	Subject Preservation		Semantic Layout Alignment			
	Visual Elements		Visual Elements		Textual Elements	
	DINO	M-DINO	Spatial	Attribute	Spatial	Sen. Acc
CreatiDesign	**86.48**	**65.75**	78.94	66.26	**56.90**	**78.30**
w/o LE	85.10	62.96	**80.99**	**69.24**	52.42	12.13
w/o LAM	85.79	64.28	66.94	56.19	20.16	68.41
w/o SAM	85.70	64.14	75.99	64.90	56.92	76.84

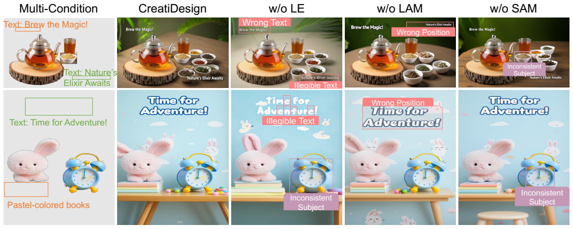

Figure 8: Qualitative Results of Ablation Study.

6 Conclusion

In this paper, we presented CreatiDesign, a systematic solution that empowers diffusion transformers for intelligent and highly controllable graphic design generation. We designed a unified multi-condition driven architecture that seamlessly integrates heterogeneous design elements. Furthermore, we proposed a multimodal attention mask mechanism to ensure that each condition precisely controls its designated image region and to prevent interference between conditions. In addition, we introduced a fully automated pipeline for constructing large-scale, richly annotated graphic design datasets. Extensive experiments demonstrated that CreatiDesign outperforms previous methods in subject preservation, semantic layout alignment, and overall visual quality.

Limitation and Future Work. CreatiDesign faces challenges in accurately preserving facial details and generating dense text, as our current dataset is not tailored for these scenarios. Improving performance in such cases, either through dataset enhancement or model-level advances, represents an important direction for future research.

References

Grounded-sam-2. https://github.com/IDEA-Research/Grounded-SAM-2, 2024.

Hidream-i1. https://github.com/HiDream-ai/HiDream-I1, 2025.

Cogview4. https://github.com/THUDM/CogView4, 2025.

Gemini-2.0-flash. https://aistudio.google.com/prompts/new_chat?model=gemini-2.0-flash-exp, 2025.

4o image generation. https://openai.com/index/introducing-4o-image-generation/, 2025.

Paddleocr-v4. https://github.com/PaddlePaddle/PaddleOCR, 2025.

Shengqu Cai, Eric Chan, Yunzhi Zhang, Leonidas Guibas, Jiajun Wu, and Gordon Wetzstein. Diffusion self-distillation for zero-shot customized image generation. *arXiv preprint arXiv:2411.18616*, 2024.

Anthony Chen, Jianjin Xu, Wenzhao Zheng, Gaole Dai, Yida Wang, Renrui Zhang, Haofan Wang, and Shanghang Zhang. Training-free regional prompting for diffusion transformers. *arXiv preprint arXiv:2411.02395*, 2024a.

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

Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. Textdiffuser-2: Unleashing the power of language models for text rendering. In *ECCV*, 2024b.

Junsong Chen, YU Jincheng, GE Chongjian, Lewei Yao, Enze Xie, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *ICLR*, 2024c.

Yutao Cheng, Zhao Zhang, Maoke Yang, Hui Nie, Chunyuan Li, Xinglong Wu, and Jie Shao. Graphic design with large multimodal model. In *AAAI*, 2025.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, 2021.

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

Yutong Feng, Biao Gong, Di Chen, Yujun Shen, Yu Liu, and Jingren Zhou. Ranni: Taming text-to-image diffusion for accurate instruction following. In *CVPR*, 2024.

Yifan Gao, Zihang Lin, Chuanbin Liu, Min Zhou, Tiezheng Ge, Bo Zheng, and Hongtao Xie. Postermaker: Towards high-quality product poster generation with accurate text rendering. In *CVPR*, 2025a.

Yu Gao, Lixue Gong, Qiushan Guo, Xiaoxia Hou, Zhichao Lai, Fanshi Li, Liang Li, Xiaochen Lian, Chao Liao, Liyang Liu, et al. Seedream 3.0 technical report. *arXiv preprint arXiv:2504.11346*, 2025b.

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

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020.

Minghui Hu, Jianbin Zheng, Daqing Liu, Chuanxia Zheng, Chaoyue Wang, Dacheng Tao, and Tat-Jen Cham. Cocktail: Mixing multi-modality control for text-conditional image generation. In *NeurIPS*, 2023.

Paul Jobling and David Crowley. Graphic design: reproduction and representation since 1800. 1996.

Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *arXiv preprint arXiv:2305.01569*, 2023.

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

Black Forest Labs. Flux.1-fill-dev. https://huggingface.co/black-forest-labs/FLUX.1-Fill-dev, 2025.

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

Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xinchi Deng, Yingfang Zhang, Xingchao Liu, Minbin Huang, Zedong Xiao, et al. Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding. *arXiv e-prints*, 2024.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.

Zeyu Liu, Weicong Liang, Zhanhao Liang, Chong Luo, Ji Li, Gao Huang, and Yuhui Yuan. Glyph-byt5: A customized text encoder for accurate visual text rendering. In *ECCV*, 2024.

Jian Ma, Yonglin Deng, Chen Chen, Nanyang Du, Haonan Lu, and Zhenyu Yang. Glyphdraw2: Automatic generation of complex glyph posters with diffusion models and large language models. In *AAAI*, 2025a.

Yuhang Ma, Shanyuan Liu, Ao Ma, Xiaoyu Wu, Dawei Leng, and Yuhui Yin. Hico: Hierarchical controllable diffusion model for layout-to-image generation. In *NeurIPS*, 2025b.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 2021.

OpenBMB. Minicpm-v-2.6. https://huggingface.co/openbmb/MiniCPM-V-2_6, 2024.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Yuyang Peng, Shishi Xiao, Keming Wu, Qisheng Liao, Bohan Chen, Kevin Lin, Danqing Huang, Ji Li, and Yuhui Yuan. Bizgen: Advancing article-level visual text rendering for infographics generation. In *CVPR*, 2025.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *ICLR*, 2024.

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

Can Qin, Shu Zhang, Ning Yu, Yihao Feng, Xinyi Yang, Yingbo Zhou, Huan Wang, Juan Carlos Niebles, Caiming Xiong, Silvio Savarese, et al. Unicontrol: A unified diffusion model for controllable visual generation in the wild. *arXiv preprint arXiv:2305.11147*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICLM*, 2021.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 2020.

Lingmin Ran, Xiaodong Cun, Jia-Wei Liu, Rui Zhao, Song Zijie, Xintao Wang, Jussi Keppo, and Mike Zheng Shou. X-adapter: Adding universal compatibility of plugins for upgraded diffusion model. In *CVPR*, pp. 8775–8784, 2024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, 2023.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022.

Chaehun Shin, Jooyoung Choi, Heeseung Kim, and Sungroh Yoon. Large-scale text-to-image model with inpainting is a zero-shot subject-driven image generator. *arXiv preprint arXiv:2411.15466*, 2024.

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

Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. In *CVPR*, 2024.

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

Yuxiang Tuo, Yifeng Geng, and Liefeng Bo. Anytext2: Visual text generation and editing with customizable attributes. *arXiv preprint arXiv:2411.15245*, 2024.

Haoxuan Wang, Jinlong Peng, Qingdong He, Hao Yang, Ying Jin, Jiafu Wu, Xiaobin Hu, Yanjie Pan, Zhenye Gan, Mingmin Chi, et al. Unicombine: Unified multi-conditional combination with diffusion transformer. *arXiv preprint arXiv:2503.09277*, 2025a.

Shaodong Wang, Yunyang Ge, Liuhan Chen, Haiyang Zhou, Qian Wang, Xinhua Cheng, and Li Yuan. Prompt2poster: Automatically artistic chinese poster creation from prompt only. In *ACMMM*, 2024a.

Xierui Wang, Siming Fu, Qihan Huang, Wanggui He, and Hao Jiang. Ms-diffusion: Multi-subject zero-shot image personalization with layout guidance. *arXiv preprint arXiv:2406.07209*, 2024b.

Xudong Wang, Trevor Darrell, Sai Saketh Rambhatla, Rohit Girdhar, and Ishan Misra. Instancediffusion: Instance-level control for image generation. In *CVPR*, 2024c.

Zhendong Wang, Jianmin Bao, Shuyang Gu, Dong Chen, Wengang Zhou, and Houqiang Li. Designdiffusion: High-quality text-to-design image generation with diffusion models. In *CVPR*, 2025b.

Shaojin Wu, Mengqi Huang, Wenxu Wu, Yufeng Cheng, Fei Ding, and Qian He. Less-to-more generalization: Unlocking more controllability by in-context generation. *arXiv preprint arXiv:2504.02160*, 2025.

Yinwei Wu, Xianpan Zhou, Bing Ma, Xuefeng Su, Kai Ma, and Xinchao Wang. Ifadapter: Instance feature control for grounded text-to-image generation. *arXiv preprint arXiv:2409.08240*, 2024.

Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. *arXiv preprint arXiv:2409.11340*, 2024.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. In *NeurIPS*, 2023.

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

Lvmin Zhang and Maneesh Agrawala. Transparent image layer diffusion using latent transparency. *arXiv preprint arXiv:2402.17113*, 2024.

Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K Wong. Uni-controlnet: All-in-one control to text-to-image diffusion models. *NeurIPS*, 2023.

Dewei Zhou, You Li, Fan Ma, Xiaoting Zhang, and Yi Yang. Migc: Multi-instance generation controller for text-to-image synthesis. In *CVPR*, 2024.

Dewei Zhou, Mingwei Li, Zongxin Yang, and Yi Yang. Dreamrenderer: Taming multi-instance attribute control in large-scale text-to-image models. *arXiv preprint arXiv:2503.12885*, 2025.

Chenyang Zhu, Kai Li, Yue Ma, Chunming He, and Xiu Li. Multibooth: Towards generating all your concepts in an image from text. In *AAAI*, 2025.

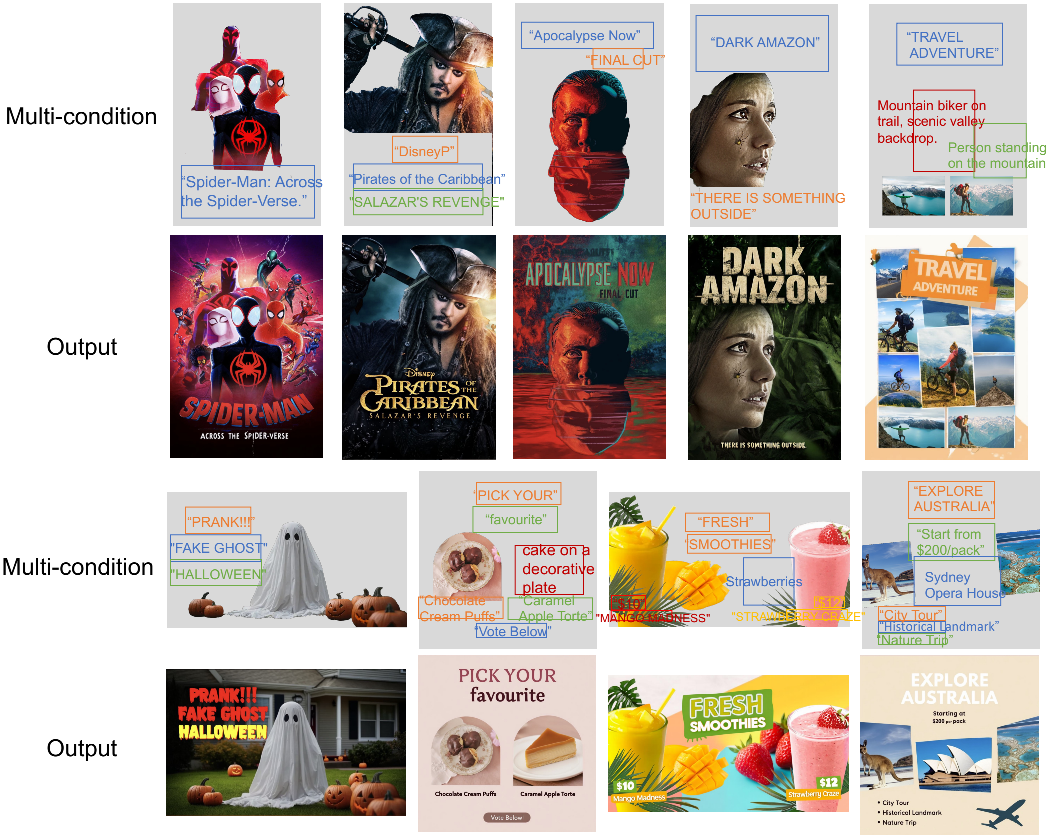

Figure 9: Quantitative results under detailed multi-condition inputs.

A Appendix

A.1 Further Discussion on Conditional Flexibility

Although CreatiDesign supports multiple fine-grained conditional controls, if a user does not specify all types of conditions—only provides one type of control, such as the subject image or semantic layout—the model can still generate high-quality and coherent designs. These conditions collectively serve as an upper-bound interface: the more information provided, the greater the controllability. Even with minimal conditional input, the results remain reasonable, as the model can still follow the global prompt for overall guidance. Besides, in real applications, CreatiDesign offers an intuitive interactive canvas, allowing users to drag and drop primary subject images, optionally sketch bounding boxes and enter their descriptions. We observe that the average interaction time per design is less than 20 seconds, making the workflow both efficient for casual users and sufficiently powerful for those requiring pixel-level control.

A.2 More Quantitative Results of Multi-Condition Generation

Figure 9 showcases more quantitative results illustrating CreatiDesign's ability to generate high-quality designs conditioned on multiple input conditions. In each example, the "Multi-condition" row presents a combination of fine-grained controls, including subjects and semantic layouts. The "Output" row shows the graphical design automatically generated by our model based on these conditions. As demonstrated, CreatiDesign faithfully adheres to the provided multi-condition inputs, accurately placing both textual and visual elements according to user intent. These results highlight the model's potential for automated creative design across a wide range of design scenarios.

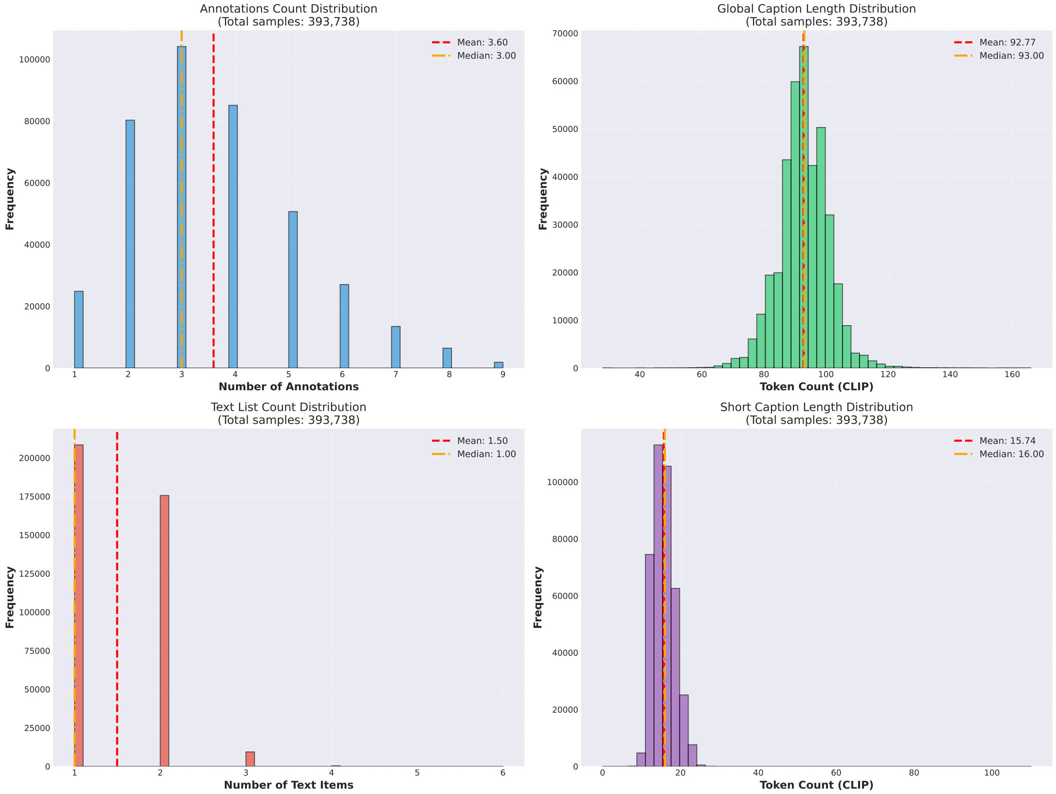

Figure 10: Statistical analysis of the CreatiDesign dataset.

A.3 Dataset Statistics

Figure 10 illustrates the dataset statistics. The dataset averages 3.60 annotations and 1.50 text instances per sample. Global prompts are provided in two formats for robustness. The long global prompts consist of 92.77 tokens on average (Median: 93.00), whereas the short prompts have a mean of 15.74 tokens.

A.4 LLM Usage Statement

Large Language Models (LLMs) were primarily used for language polishing, such as correcting grammatical errors and enhancing sentence clarity.