LAMIC: Layout-Aware Multi-Image Composition via Scalability of Multimodal Diffusion Transformer

Yuzhuo Chen[1], Zehua Ma[1*], Jianhua Wang[2], Kai Kang[3], Shunyu Yao[2*], Weiming Zhang[1]

[1]Anhui Province Key Laboratory of Digital Security, University of Science and Technology of China
[2]Onestory Team
[3]East China Normal University

Abstract

In controllable image synthesis, generating coherent and consistent images from multiple references with spatial layout awareness remains an open challenge. We propose **LAMIC**, a **L**ayout-**A**ware **M**ulti-**I**mage **C**omposition framework that, for the first time, extends single-reference diffusion models to multi-reference scenarios in a training-free manner. Built upon the MMDiT model, LAMIC introduces two plug-and-play attention mechanisms: 1) Group Isolation Attention (GIA) to enhance entity disentanglement; and 2) Region-Modulated Attention (RMA) to enable layout-aware generation. To comprehensively evaluate model capabilities, we further introduce three metrics: 1) Inclusion Ratio (IN-R) and Fill Ratio (FI-R) for assessing layout control; and 2) Background Similarity (BG-S) for measuring background consistency. Extensive experiments show that LAMIC achieves state-of-the-art performance across most major metrics: it consistently outperforms existing multi-reference baselines in ID-S, BG-S, IN-R and AVG scores across all settings, and achieves the best DPG in complex composition tasks. These results demonstrate LAMIC's superior abilities in identity keeping, background preservation, layout control, and prompt-following, all achieved without any training or fine-tuning, showcasing strong zero-shot generalization ability. By inheriting the strengths of advanced single-reference models and enabling seamless extension to multi-image scenarios, LAMIC establishes a new training-free paradigm for controllable multi-image composition. As foundation models continue to evolve, LAMIC's performance is expected to scale accordingly.

Code — https://github.com/Suchenl/LAMIC

1 Introduction

Creating consistent and controllable visual content is a core challenge in digital filmmaking, storyboarding, and narrative illustration (Zhou, Yang et al. 2024). In these domains, artists often need to construct scenes that involve multiple entities—such as characters and environments—while maintaining visual and stylistic consistency across varying perspectives and story beats. With the rapid progress of diffusion-based generative models, there is growing interest in leveraging such models to automate and accelerate the generation of entity-consistent, layout-controllable images guided by both textual and visual inputs (Xiao et al. 2025).

Recent advances in image generation have introduced impressive capabilities in text-to-image (T2I) and image-to-image (I2I) synthesis (Rombach et al. 2022a; Brooks, Holynski, and Efros 2023). More recently, multimodal text-and-image-to-image (T&I2I) models, especially FLUX.1-Kontext released on May 29, 2025 (Labs et al. 2025), have demonstrated the great potential of combining textual descriptions with visual references to generate semantically grounded and identity-consistent images. However, as a single-reference model, it still remains significantly limited in handling multiple reference images. To enable multi-image reference generation, some methods have introduced trainable extension modules into the fundamental T2I models (Chen et al. 2025), while others have retrained a slightly modified T2I architecture (Xiao et al. 2024b; Wu et al. 2025a). However, these training-based methods face challenges in generalization performance when combining more images, as large-scale datasets with multiple image references are difficult to collect (Chen et al. 2025).

In addition, many of these methods lack spatial layout capabilities, which limits their application in real scenarios. These limitations become particularly problematic in creative production workflows. For instance, in storyboard generation for films or animations, it is crucial to consistently generate the same multiple characters, objects and scenes which conform to explicit layout plans-such as character positioning, scene framing, or camera perspective determined by directors or artists. Previous studies have investigated layout control in T2I systems, which can be mainly divided into training-based (Zhang et al. 2025; Tan et al. 2025) and training-free (Yang et al. 2024; Chen et al. 2024b). However, the former requires the introduction of additional modules for specific tasks or fine-tuning with LoRA to force the generated image to conform to the layout input, which is still constrained by the dataset. The latter is mainly achieved by manipulating the area where the prompt word is injected or using the model's local predicted noise to replace the corresponding area in the global noise predicted by the model. However, such methods are prone to cross-image interference and semantic leakage, especially in the case of similar appearance of entities (for example, multiple humans or animals), resulting in reduced subject consistency.

*Co-corresponding authors.
Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

To address the above limitations, we propose **LAMIC (Layout-Aware Multi-Image Composition)**, a training-free framework built upon a pretrained single-image reference Multimodal Diffusion Transformer (MMDiT) (Esser et al. 2024) model. Leveraging our proposed attention mechanism and the strong scalability of the MMDiT architecture, LAMIC enables users to incorporate an arbitrary number of reference images along with region-level layout priors (e.g., masks or bounding boxes). Our framework achieves superior overall performance compared to prior approaches in multi-image composition, particularly excelling in tasks involving fine-grained layout control. The main contributions of this paper are as follows:

- We propose **LAMIC**, a novel framework for high-quality **layout-aware multi-image composition**, supporting flexible spatial layout control and seamless multi-reference integration.
- LAMIC is the first to **extend single-reference diffusion models to multi-reference scenarios** in a **training-free** manner. It inherits the consistency-preserving editing capabilities of the underlying model, while circumventing the generalization issues caused by the scarcity of large-scale multi-reference training datasets.
- To reduce semantic entanglement across entities, we introduce **Group Isolation Attention (GIA)**, which enforces localized attention within aligned visual-textual-spatial (VTS) triplets.
- Building on GIA, we further propose **Region-Modulated Attention (RMA)**, which defers inter-region fusion and cross-entity interaction (CEI) instruction injection to enhance layout controllability and prevent early-stage semantic leakage.

2 Related Work

Reference-Guided Image Generation. Recent advancements in multimodal image generation enable synthesizing images guided by both textual and visual references. UniAdapter (Wang, Hu et al. 2023) introduced early adapter-based methods for reference adaptation, though limited to single references without spatial disentanglement. EasyRef (Zong et al. 2024) proposed leveraging multiple reference images but relies on complex multimodal large language models (MLLMs), hindering practical usability. FLUX.1-Kontext (Labs et al. 2025), built on MMDiT architecture, demonstrates significant improvements in identity consistency using a single reference image. However, these methods remain inadequate for handling multiple image references effectively.

Layout-Aware Generation. Spatial layout control has been explored via supervised segmentation maps, bounding-box conditioning (Chen et al. 2024b), and region-aware prompts (Chen et al. 2024a; Hsiao et al. 2025). However, most approaches rely on either fine-tuning (Ruiz, Li et al. 2023), prompt heuristics (Yang et al. 2024), training-time supervision (He et al. 2025) or repeated inference (Chen et al. 2024b), making them less flexible for open-domain generation. In contrast, our method avoids parameter tuning, extra inference, and complex prompt engineering, offering a more practical solution for open-domain scenarios.

Multi-Image Composition. Effective compositional generation involves integrating multiple visual references into coherent images. MS-Diffusion (Wang et al. 2025) pioneered multi-modal inference with layout control but exhibits limitations in identity preservation and spatial accuracy. Methods like OmniGen (Xiao et al. 2024b) and OmniGen2 (Wu et al. 2025a) enhance identity disentanglement but generally require extensive retraining, restricting scalability. UNO (Wu et al. 2025b) and DreamO (Mou et al. 2025) provide consistent cross-reference synthesis but lack explicit layout control. Although XVerse (Chen et al. 2025) achieves fine-grained identity control, like most training-based methods, it relies on large-scale multi-reference datasets that are difficult to collect, leading to generalization limitations in practical scenarios.

3 Method

3.1 Preliminaries and Key Insights

Multimodal Diffusion Transformer. MMDiT (Esser et al. 2024), introduced in Stable Diffusion 3, extends DiT (Peebles and Xie 2023) by concatenating text and latent image tokens for multimodal conditioning within the LDM framework (Rombach et al. 2022b). This design has been adopted in models such as FLUX.1 (Labs 2024) and FLUX.1-Kontext (Labs et al. 2025), with the latter demonstrating strong identity preservation in single-reference generation. Notably, both Kontext and recent control frameworks (e.g., EasyControl (Zhang et al. 2025), OmniControl2 (Tan et al. 2025)) introduce external control signals—like reference images—via token concatenation. This design paradigm reveals a **key insight**: it is possible to introduce multiple-reference images into a unified representation space using only a pretrained single-reference network.

Attention Mechanism. In MMDiT, both self-attention and cross-attention layers are applied at each denoising step to model complex dependencies. The attention computation is defined as:

$$\mathrm{Att}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right) \cdot V \tag{1}$$

where Q (query), K (key), and V (value) are projections of either latent tokens or conditioning embeddings. While standard attention enables global interaction across all tokens, this becomes problematic in the multi-reference setting: cross-token mixing can lead to interference between unrelated entities—whether in textual descriptions, visual references, or layout controls.

3.2 Overview of LAMIC

As shown in Figure 1, LAMIC enables layout-aware multi-entity generation through the following three stages: 1) **Structured Input Definition**, where each reference is organized into a visual-textual-spatial (VTS) triplet, complemented by cross-entity interaction (CEI) instructions and uncontrolled regions; 2) **Unified Token Representation**,

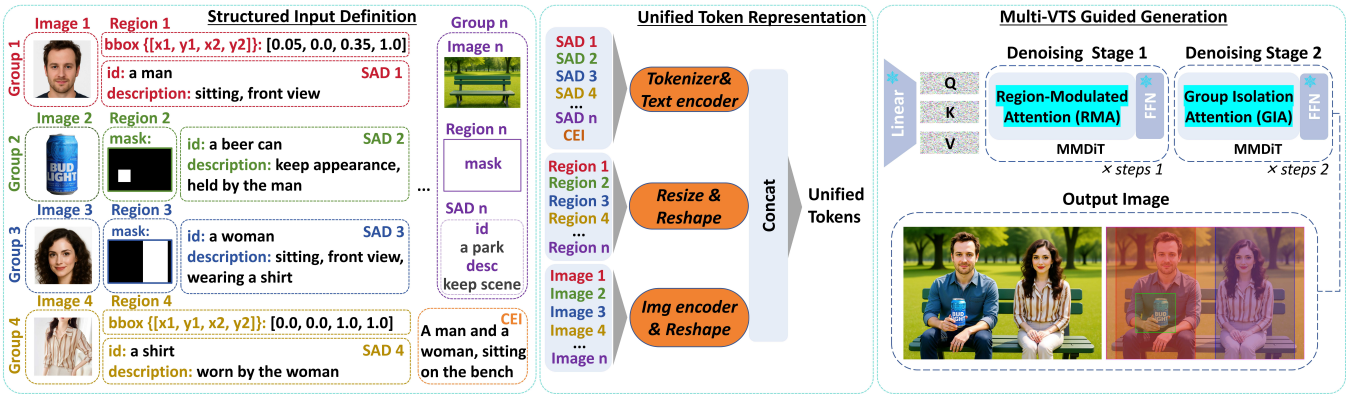


Figure 1: Framework of our proposed LAMIC. We illustrate the layout-aware multi-image composition process with 5 reference groups (n=5) provided as input.

where all components—VTS triplets, CEI, and uncontrolled regions—are encoded into a unified token sequence for joint representation in MMDiT; 3) **Multi-VTS Guided Generation**, where image synthesis is guided by all VTS tokens. Two attention mechanisms are introduced to support this stage: **Group Isolation Attention (GIA)** restricts cross-group interaction among textual, spatial, and visual tokens to prevent semantic entanglement; **Region-Modulated Attention (RMA)** defers inter-region fusion and CEI injection to enhance layout controllability and avoid early-stage semantic leakage.

3.3 Structured Input Definition

We structure each reference as a visual-textual-spatial (VTS) triplet group $G_i = (V_i, T_i, S_i)$, where V_i denotes the visual reference (image), T_i represents the textual condition—referred to as the self-attribute description (SAD), and S_i specifies the target spatial region (e.g., bounding box or mask). Each SAD consists of an **identifier** describing the entity (e.g., "a dragon", "a car"), and a **description** specifying appearance behavior (e.g., "keep the same appearance", "change the pose"). We further introduce a **Cross-Entity Interaction (CEI)** instruction C, which governs spatial or semantic relationships between entities (e.g., "A rides B"), and an **uncontrolled region** U, covering areas not assigned to any specific entity. Together, these components—$\{(V_i, T_i, S_i)\}_{i=1}^{N}$, C, and U—form a structured input that aligns visual, textual, and spatial guidance, and N denotes the number of references. This design supports multi-entity composition without relying on external reasoning modules such as MLLMs (Yang et al. 2024).

3.4 Unified Token Representation

We encode all components—$\{(V_i, T_i, S_i)\}_{i=1}^{N}$, C, and U—into a unified token sequence. Specifically, we use the pretrained VAE or AE from MMDiT to convert each V_i into latent tokens $L_i \in \mathbb{R}^{B\times(H_i\times W_i/4)\times 4C}$, where B, C, H_i, and W_i denote the batch size, channel count, and spatial dimensions. Textual inputs (T_i, C) are embedded via pretrained T5 (Raffel et al. 2020) or CLIP (Radford et al. 2021), and projected to match the latent token space. Spatial regions S_i are downsampled (typically $\times 8$) and reshaped to match the image token format. We then concatenate all tokens along the sequence dimension and record their positions for subsequent attention masking.

3.5 Multi-VTS Guided Generation

We design two attention mechanisms to support layout-aware generation guided by multi-VTS tokens: Group Isolation Attention (GIA) and Region-Modulated Attention (RMA), as illustrated in Figure 2.

Group Isolation Attention (GIA). GIA suppresses interference across VTS groups by restricting attention computations within each group:

$$GIA(Q_{G_i}, K_{G_j}, V_{G_j}) = \begin{cases} \mathrm{Att}(\cdot), & i = j \\ 0, & i \neq j \end{cases} \tag{2}$$

where $i, j \in \{1, \ldots, N\}$. For brevity, in multi-case equations, we use $\mathrm{Att}(\cdot)$ to denote the attention operation with the same (Q, K, V) inputs as those on the left-hand side of the equation.

To ensure structural coherence, we retain unrestricted attention between spatial regions on the condition of Eq. (2):

$$GIA(Q_{S_i}, K_{S_j}, V_{S_j}) = \mathrm{Att}(Q_{S_i}, K_{S_j}, V_{S_j}),\ \forall i, j \tag{3}$$

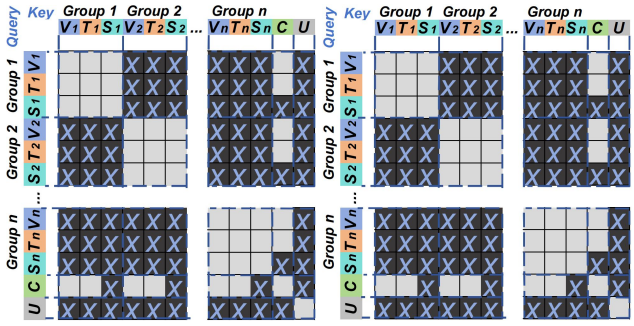
We also define cross-group interactions with CEI (C) and uncontrolled region (U): C is treated as a global prompt and interacts fully with all groups, while U follows the spatial attention pattern:

$$GIA(Q_{G_i}, K_C, V_C) = \mathrm{Att}(Q_{G_i}, K_C, V_C) \tag{4a}$$
$$GIA(Q_C, K_{G_i}, V_{G_i}) = \mathrm{Att}(Q_C, K_{G_i}, V_{G_i}) \tag{4b}$$

$$GIA(Q_{G_i^y}, K_U, V_U) = \begin{cases} \mathrm{Att}(\cdot), & y = S \\ 0, & y \neq S \end{cases} \tag{4c}$$

$$GIA(Q_U, K_{G_i^y}, V_{G_i^y}) = \begin{cases} \mathrm{Att}(\cdot), & y = S \\ 0, & y \neq S \end{cases} \tag{4d}$$

where $y \in \{V, T, S\}, \forall i \in N$.



(a) Group Isolation Attention (b) Region-Modulated Attention

Figure 2: Our proposed attention mechanisms.

Region-Modulated Attention (RMA). To promote precise spatial control and prevent early-stage semantic leakage, based on GIA, RMA further limits the inter-region cross-attention and CEI injection in the early denoising step:

$$RMA(Q_{S_i}, K_{S_j}, V_{S_j}) = \begin{cases} \text{Att}(\cdot), & i = j \\ 0, & i \neq j \end{cases} \quad (5)$$

and attention with U and C is disabled:

$$RMA(Q_{S_i}, K_U, V_U) = RMA(Q_U, K_{S_i}, V_{S_i}) = 0 \quad (6)$$
$$RMA(Q_{S_i}, K_C, V_C) = RMA(Q_C, K_{S_i}, V_{S_i}) = 0 \quad (7)$$
$$RMA(Q_U, K_C, V_C) = RMA(Q_C, K_U, V_U) = 0 \quad (8)$$

In practice, we implement these rules via attention masks for efficiency. The total denoising process is divided into two sub-stages: RMA is applied during the first stage, covering a predefined ratio of the total steps, followed by GIA in the remaining steps.

4 Experiments

4.1 Experimental Setting

Implementation Details. We implement LAMIC based on the open-source single-image reference MMDiT-based model (Flux.1 Kontext-dev). The inference process is configured with 20 denoising steps, a guidance scale of 2.5, and a first-stage step ratio of 0.05. To reduce memory consumption, both the Transformer and T5 modules are quantized to INT8 during inference. All experiments are conducted on a machine equipped with a single NVIDIA RTX 4090 and dual NVIDIA A6000 GPUs.

Benchmark Dataset. As a benchmark for multi-image composition, XVerseBench (Chen et al. 2025) originally includes 74 objects, 20 human faces, and 45 animals. However, we observed its limitations in subject diversity and visual quality. To address this, we augmented the dataset with 20 additional scenes, 17 clothing items, and 1 object sourced from DreamBench++ (Peng et al. 2025), MS-Bench (Wang et al. 2025), and GPT-4o generations. Moreover, due to the low resolution and noise present in some original samples, we regenerated 20 high-resolution human faces and replaced several degraded images using high-quality generation tools. Beyond improving image quality and type diversity, we constructed structured multi-image inputs with associated bounding boxes for each subject, enabling precise layout-aware generation and evaluation. Specifically, we created 60 inputs with two reference images, 40 inputs with three reference images, and 20 inputs with four reference images.

Evaluation Metrics. Following prior work (Chen et al. 2025), we adopt several established metrics to assess generation quality. To better evaluate background consistency and layout controllability, we further propose three novel metrics: **BG-S**, **IN-R**, and **FI-R**, which provide finer-grained analysis in multi-reference, layout-aware synthesis settings.

- **DPG Score** (Hu et al. 2024), measuring the text consistency editing ability of the model;
- **Face ID Similarity (ID-S)** (Deng et al. 2019), evaluating human identity preservation;
- **DINOv2 Similarity (IP-S)** (Oquab et al. 2023), capturing object appearance consistency;
- **Aesthetic Score (AES)** (discus0434 2024), judging overall aesthetic appeal;
- **Background Similarity (BG-S)**, a weighted combination of DINOv2, CLIP (Radford et al. 2021), SSIM, and color histogram (CH):

$$\textit{BG-S} = 0.4 \cdot \textit{DINO} + 0.25 \cdot \textit{CLIP} + 0.2 \cdot \textit{SSIM} + 0.15 \cdot \textit{CH} \quad (9)$$

We report the unweighted average of the above five metrics as **AVG** for overall generation quality.

For layout evaluation, we employ the Grounded SAM 2 pipeline, where Florence-2 (Xiao et al. 2024a) handles object detection and grounding, generating bounding boxes for target entities, which are then processed by SAM-2 to produce precise segmentation masks M_{gen}, which are compared to the ground-truth target region mask M_{trg}.

Specifically, **IN-R (Inclusion Ratio)** measures how much of the generated entity lies within the target region, while **FI-R (Fill Ratio)** evaluates how well the target region is covered by the generated entity. These two ratios jointly reflect the precision and completeness of layout control. In order to unify the value range with other metrics, we multiplied both ratios by 100.

$$\textit{IN-R} = \frac{\sum(M_{\text{gen}} \cap M_{\text{trg}})}{\sum M_{\text{gen}}} \times 100 \quad (10)$$

$$\textit{FI-R} = \frac{\sum(M_{\text{gen}} \cap M_{\text{trg}})}{\sum M_{\text{trg}}} \times 100 \quad (11)$$

To avoid artificially inflated layout scores caused by large targets (e.g., full-image regions), we discard samples where the target mask occupies more than 75% of the image area.

Baseline Methods. We first evaluate the performance of our method on multi-image composition tasks, comparing it against several state-of-the-art multi-reference generation approaches, including MS-Diffusion (Wang et al. 2025), MIP-Adapter (Huang et al. 2025), OmniGen (Xiao et al. 2024b), UNO (Wu et al. 2025b), OmniGen2 (Wu et al. 2025a), DreamO (Mou et al. 2025), and XVerse (Chen et al.

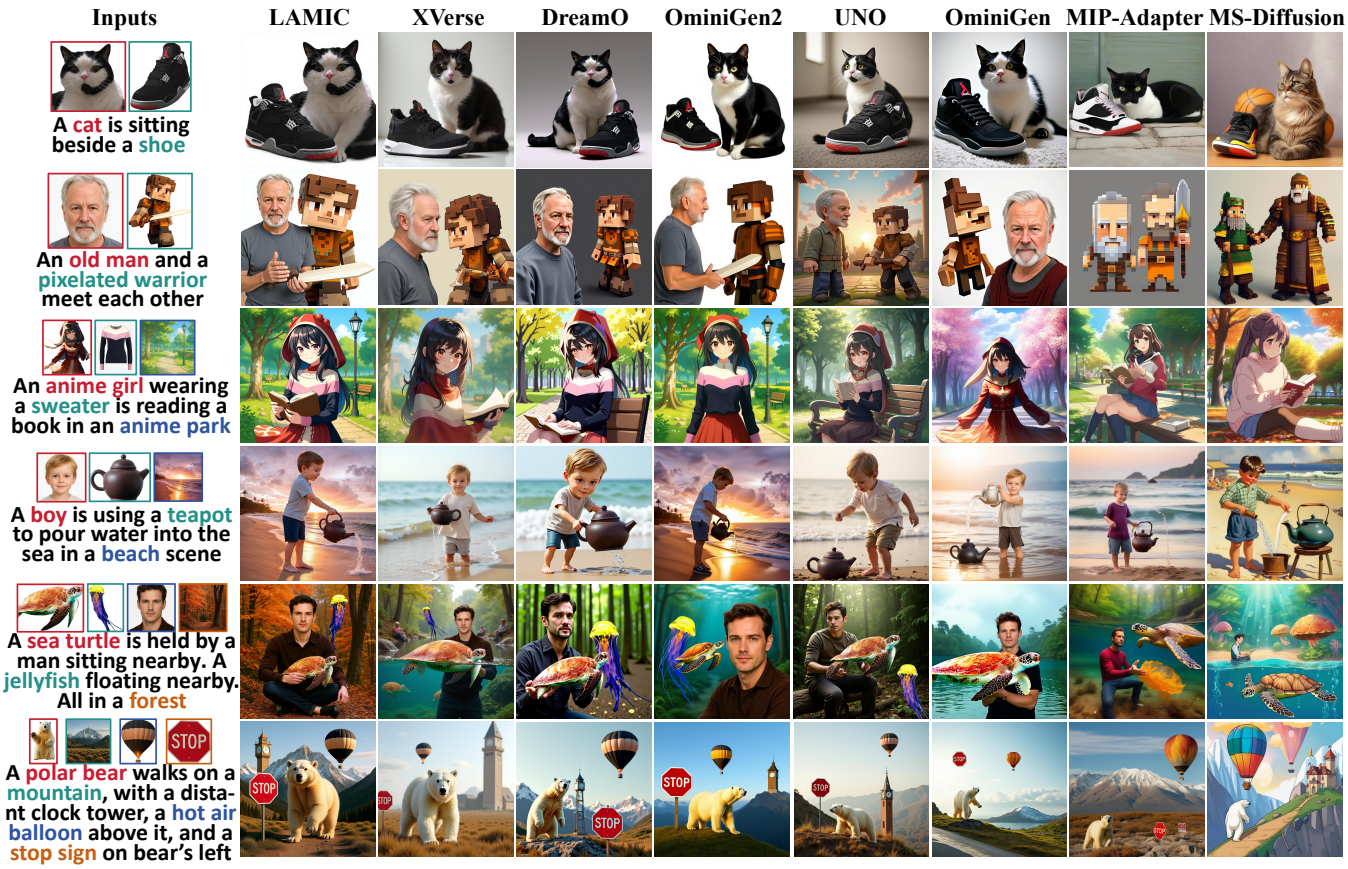


Figure 3: Visual comparison of different methods under different multi-reference images.

2025). These models are evaluated across various subject composition and background integration scenarios under three settings. After that, we further evaluate the layout control performance of the above methods.

4.2 Multi-Image Composition Performance

Quantitative Comparison. Table 1 quantitatively demonstrates that LAMIC consistently achieves the best overall performance across all settings. Notably, it obtains the highest ID-S, BG-S, and AVG scores in each reference configuration, indicating strong identity and background preservation, as well as balanced generation quality. In the three- and four-reference settings, LAMIC also achieves the highest DPG score, demonstrating its superior editing capability and prompt consistency. ***(1) Specifically, in the two-reference setting***, LAMIC achieves an ID-S of 78.04, surpassing the second-best OmniGen by nearly 9 points; a BG-S of 83.14, exceeding the second-best OmniGen2 by 2.55; and an AVG of 74.54, outperforming the runner-up OmniGen2 by 4.3. Furthermore, the IP-S of 72.33 is only 1.12 below the best-performing model, demonstrating excellent object preservation. ***(2) In the three-reference setting***, LAMIC achieves a DPG of 91.95, an ID-S of 65.63, and an AVG of 73.92, all outperforming the respective second-best models by approximately 1 point. Notably, it attains a BG-S of 86.06, which is 2.5 points higher than the second-best OmniGen2, further validating LAMIC's superior capability in background consistency. ***(3) In the more challenging four-reference setting***, LAMIC still leads with a DPG of 90.16 (tied with UNO), an ID-S of 70.25 (exceeding the second-best XVerse by 8.41), a BG-S of 87.02, and an AVG of 74.44, surpassing all competitors by a large margin, indicating LAMIC maintains strong performance as the reference number increases.

These improvements are achieved without any fine-tuning or model re-training, highlighting the zero-shot generalization capability of our method. It is also worth noting that UNO consistently achieves the best AES, reflecting its strong aesthetic appeal, while OmniGen2 performs second-best on BG-S, demonstrating good background synthesis ability. DreamO and XVerse also achieve solid overall performance, closely following LAMIC in terms of AVG.

Qualitative Comparison. Figure 3 presents qualitative comparisons under diverse multi-reference scenarios. LAMIC excels in preserving subject identity and structural fidelity, generating visually coherent and high-quality results. For example, in the "old man-pixelated warrior" composition (Row 2), LAMIC successfully maintains the subject's stylized structure and realistic blending, while other methods exhibit over-smoothing or distortions. In the "sea

Method	Two-Reference						Three-Reference						Four-Reference					
	DPG	ID-S	IP-S	BG-S	AES	AVG	DPG	ID-S	IP-S	BG-S	AES	AVG	DPG	ID-S	IP-S	BG-S	AES	AVG
MS-Diffusion	75.01	12.70	47.05	71.48	52.88	51.82	86.67	3.46	43.13	72.49	57.85	52.72	75.24	3.78	41.25	73.24	55.41	49.78
MIP-Adapter	82.64	22.28	59.59	75.15	55.77	59.09	84.97	20.40	61.58	78.93	60.61	61.30	83.49	13.31	57.73	77.82	58.60	58.19
OmniGen	82.06	69.32	66.63	73.25	56.54	69.56	72.79	61.24	64.52	78.30	59.60	67.29	74.60	54.88	61.60	77.21	59.39	65.54
UNO	**89.42**	38.71	72.92	72.32	**59.52**	66.58	91.00	43.52	69.03	78.72	**62.45**	68.94	**90.16**	39.98	**73.73**	78.34	**61.75**	68.79
OmniGen2	85.53	59.89	70.60	80.59	54.61	70.24	81.00	55.53	65.88	83.58	59.32	69.06	81.09	50.66	62.90	81.56	56.49	66.54
DreamO	88.54	58.71	**73.45**	73.99	55.28	69.99	90.60	63.74	71.49	80.73	57.01	72.71	90.04	57.29	70.19	80.08	55.95	70.71
XVerse	87.90	56.49	70.72	74.62	59.38	69.82	90.23	63.70	**73.19**	79.53	59.06	73.14	84.00	61.84	70.69	78.72	56.71	70.39
LAMIC(Ours)	85.61	**78.04**	72.33	**83.14**	53.59	**74.54**	**91.95**	**65.63**	67.54	**86.06**	59.24	**73.92**	**90.16**	**70.25**	66.67	**87.02**	58.10	**74.44**

Table 1: Quantitative results of multi-image combination. **Bold** indicates the best result, single underline indicates the second-best, and double underline indicates the third-best.

Method	Two-Ref		Three-Ref		Four-Ref	
	IN-R	FI-R	IN-R	FI-R	IN-R	FI-R
MS-Diffusion	56.83	23.17	72.84	19.06	58.55	20.32
MIP-Adapter	59.25	21.73	70.43	20.34	63.16	19.40
OmniGen	58.60	16.40	66.43	20.87	58.99	14.96
UNO	63.46	15.74	70.37	18.54	70.13	17.85
OmniGen2	67.49	27.87	70.72	**27.27**	58.50	**22.30**
DreamO	69.25	24.37	73.84	23.57	72.95	**23.71**
XVerse	62.65	20.30	75.42	22.83	62.24	19.62
LAMIC(Ours)	**92.39**	**32.75**	**91.90**	24.26	**89.81**	20.81

Table 2: Comparative results of our LAMIC and other methods under layout-aware multi-image composition.

turtle–jellyfish–man–forest" composition (Row 5), LAMIC respects spatial arrangements and visual semantics, accurately merging all referenced elements, whereas most baselines suffer from object mismatching or semantic drifting.

XVerse consistently delivers visually pleasing generations with relatively high ID-S, particularly in cases with prominent human references (Rows 2 and 4), but tends to oversimplify background compositions. DreamO achieves smoother image transitions and realistic style rendering, as observed in Row 3 ("anime girl-sweater-anime park") and Row 4 ("boy-teapot-beach"), but occasionally struggles with precise identity preservation and text following, especially in more complex scenes. In contrast, methods like MIP-Adapter and MS-Diffusion exhibit limitations in balancing layout, identity, and appearance, often leading to incomplete or mismatched object integration.

Overall, both quantitative and qualitative results validate that LAMIC achieves superior multi-image composition, maintaining generation quality and consistency. More comparative samples are illustrated in *Appendix A*.

4.3 Layout-Controlled Multi-Image Composition Performance

Although MS-Diffusion is currently the only baseline that explicitly supports layout-aware multi-image composition, we still compare our method against all aforementioned approaches to provide a comprehensive evaluation of our proposed layout control metrics.

The results are presented in Table 2. LAMIC achieves an IN-R of approximately 90 across all settings, significantly outperforming all other methods. This is expected, as most baselines lack explicit layout control capabilities. While MS-Diffusion claims to support layout-aware generation, its performance on both IN-R and FI-R is relatively subpar. We attribute this to the nature of our proposed IN-R and FI-R metrics, which evaluate the consistency of entity placement and spatial accuracy. These metrics rely on a model's ability to preserve entities and maintain compositional alignment—areas where MS-Diffusion underperforms (Sec. 4.2), thus leading to lower scores.

It is also worth noting that although LAMIC consistently achieves the best FI-R among all methods, the margin over layout-unaware baselines is not large. This suggests that while LAMIC demonstrates superior layout control, there is still considerable room for improvement in fully capturing and preserving target spatial configurations. Some visual results of LAMIC are illustrated in *Appendix B*.

4.4 Ablation Study

In our ablation study, we first analyze the impact of removing each proposed module, and then investigate the effect of varying the ratio of first-stage steps on generation quality.

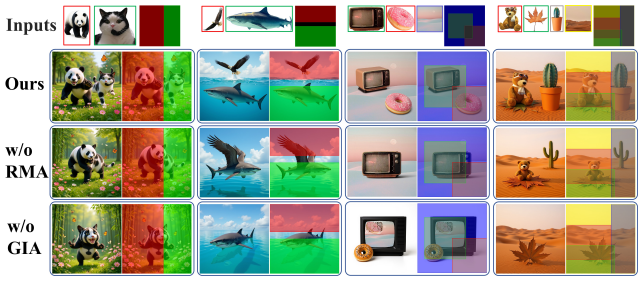


Figure 4: Visual comparison of different settings of LAMIC under layout-aware multi-image composition.

Settings		Two-Reference							Three-Reference							Four-Reference						
		DPG	ID-S	IP-S	BG-S	AES	IN-R	FI-R	DPG	ID-S	IP-S	BG-S	AES	IN-R	FI-R	DPG	ID-S	IP-S	BG-S	AES	IN-R	FI-R
LAMIC	**0.05**	85.61	78.04	72.33	83.14	53.59	92.39	32.75	**91.95**	65.63	**67.54**	86.06	59.24	91.90	24.26	90.16	**70.25**	**66.67**	**87.02**	58.10	89.81	**20.81**
	0.10	**86.76**	79.73	**72.99**	84.74	52.73	94.34	31.98	89.56	66.37	66.76	86.33	59.24	93.92	26.31	**92.71**	66.85	65.59	86.14	58.36	89.27	20.23
	0.15	86.26	81.21	72.90	**85.17**	52.39	**95.34**	32.94	88.84	**66.97**	67.23	86.52	59.09	**94.67**	27.33	89.50	66.81	65.57	86.10	58.17	90.25	20.54
	0.20	85.47	**81.37**	72.93	84.48	51.90	95.26	**33.43**	86.48	65.53	67.15	**86.58**	59.17	94.65	**27.49**	89.14	66.80	65.59	86.13	58.41	**90.72**	20.71
w/o RMA		84.18	67.96	67.25	82.95	**54.91**	81.77	28.93	90.92	64.58	64.51	85.79	**59.31**	87.45	23.81	91.25	65.81	64.84	86.30	**58.83**	88.88	19.39
w/o GIA		86.20	39.15	61.66	79.95	53.30	66.12	24.47	87.30	42.19	55.60	84.16	57.63	69.37	20.70	84.74	32.07	51.62	85.08	54.85	66.95	16.57

Table 3: Ablation results for different attentions and ratios of first-stage steps under layout-aware multi-image composition.

Impact of Proposed Modules. Table 3 quantitatively demonstrates the impact of individually removing the proposed RMA and GIA components. The full model configuration (LAMIC) achieves the best overall performance. Using RMA results in a slight drop in aesthetic quality, which is reasonable given the substantial improvement in layout control—IN-R increases by up to 10.62 points in the two-reference setting and 7.22 points in the three-reference setting compared to the version without RMA. In contrast, removing GIA causes a significant degradation across nearly all metrics, highlighting its critical role in ensuring high-quality multi-image composition.

Figure 4 presents qualitative comparisons across different settings. It clearly shows that removing RMA weakens layout control performance (e.g., in the "TV-donut" and "bear-maple leaf-cactus-desert" cases), and leads to partial fusion or entanglement of entities within target regions (e.g., in "panda-cat" and "eagle-shark"). The degradation becomes more pronounced when GIA is removed: layout control capabilities nearly vanish (consistent with the scores in Table 3), which are comparable to layout-unaware baselines Table 2, and multiple reference entities collapse into a single blended form in most situations or just keep a single entity.

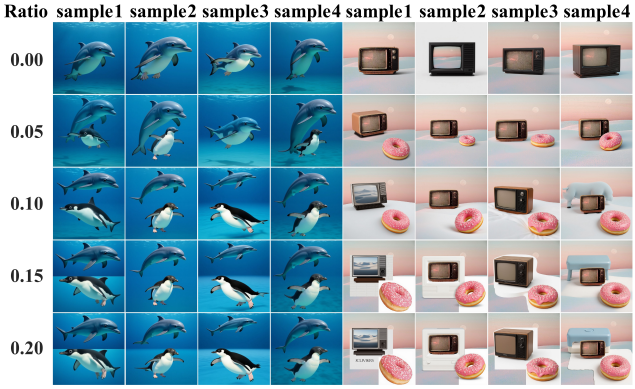


Figure 5: Visual comparison of different first-stage step ratios in LAMIC under layout-aware multi-image composition. For the left group, the target regions are the upper and lower halves (0.5 each); for the right group, the target regions correspond to those shown in Figure 4.

Impact of First Stage Steps. Table 3 also reveals the impact of varying the ratio of first-stage steps. As expected, increasing this ratio generally improves layout control: in both two-, three-, and four-reference settings, a ratio of 0.15 or 0.20 consistently achieves the best or second-best scores in IN-R and FI-R. An exception occurs in the four-reference case, where a lower ratio of 0.05 yields the highest FI-R.

However, this improvement in layout control comes at the cost of other aspects of generation quality. As the ratio increases, the AES tends to decline, with the best scores observed at lower ratios (0.05 or 0.00, the latter corresponding to the w/o RMA case). A similar trend is observed for DPG, suggesting that excessive first-stage emphasis may impair prompt-following fidelity and regional continuity.

To visualize these effects, we present two groups of image samples in Figure 5, where each sample in a single group shares identical inputs but uses different random seeds. These examples highlight the trade-off between layout precision and global coherence when adjusting the first-stage ratio. We further observe that when the ratio reaches 0.10 or higher, distinct boundaries begin to appear between adjacent regions, and visual consistency within individual regions degrades noticeably. In contrast, setting the ratio to 0.05 preserves better global coherence and intra-region consistency. Although the lower ratio may occasionally result in attribute blending between closely placed entities, its quantitative scores remain comparable and consistently yield superior visual quality in practice. Therefore, we fix the first-stage ratio to 0.05 in the main experiments.

5 Conclusion

In this work, we present LAMIC, a zero-shot framework for layout-aware multi-image composition that, for the first time, extends consistent single-reference generators to multi-reference and layout-controllable generation without any fine-tuning. To address semantic entanglement and enable accurate region-wise control, we introduce two plug-and-play modules—Group Isolation Attention and Region-Modulated Attention—and further propose three targeted evaluation metrics (IN-R, FI-R, BG-S) to more comprehensively assess layout controllability and background fidelity. LAMIC achieves state-of-the-art performance across most metrics, consistently ranking first in ID-S, BG-S, IN-R, and overall AVG under 2-, 3-, and 4-reference settings, demonstrating strong identity preservation, spatial control, and prompt adherence. Mild blending near region boundaries remains, and future work will explore more refined attention strategies and earlier prompt–reference binding for improved separation and controllability.

Acknowledgements

This work was supported in part by the Natural Science Foundation of China under Grant 62121002, 62402469, 62472398, U2336206, by Fundamental Research Funds for the Central Universities under Grant WK2100000041.

References

Brooks, T.; Holynski, A.; and Efros, A. A. 2023. InstructPix2Pix: Learning to Follow Image Editing Instructions. In *CVPR*.

Chen, A.; Xu, J.; Zheng, W.; Dai, G.; Wang, Y.; Zhang, R.; Wang, H.; and Zhang, S. 2024a. Training-free regional prompting for diffusion transformers. *arXiv preprint arXiv:2411.02395*.

Chen, B.; Zhao, M.; Sun, H.; Chen, L.; Wang, X.; Du, K.; and Wu, X. 2025. XVerse: Consistent Multi-Subject Control of Identity and Semantic Attributes via DiT Modulation. *arXiv preprint arXiv:2506.21416*.

Chen, Z.; Li, Y.; Wang, H.; Chen, Z.; Jiang, Z.; Li, J.; Wang, Q.; Yang, J.; and Tai, Y. 2024b. Region-Aware Text-to-Image Generation via Hard Binding and Soft Refinement. *arXiv preprint arXiv:2411.06558*.

Deng, J.; Guo, J.; Xue, N.; and Zafeiriou, S. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4690–4699.

discus0434. 2024. Aesthetic Predictor v2.5: SIGLIP-Based Aesthetic Score Predictor [Source code]. GitHub repository. Accessed: 2025-07-01.

Esser, P.; Kulal, S.; Blattmann, A.; Entezari, R.; Müller, J.; Saini, H.; Levi, Y.; Lorenz, D.; Sauer, A.; Boesel, F.; et al. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*.

He, J.; Yang, X.; Zeng, A.; et al. 2025. EliGen: Entity-Level Controlled Image Generation with Regional Attention. arXiv:2501.01097.

Hsiao, T.-F.; Ruan, B.-K.; Wu, Y.-L.; Lin, T.-L.; and Shuai, H.-H. 2025. Tf-ti2i: Training-free text-and-image-to-image generation via multi-modal implicit-context learning in text-to-image models. *arXiv preprint arXiv:2503.15283*.

Hu, X.; Wang, R.; Fang, Y.; Fu, B.; Cheng, P.; and Yu, G. 2024. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*.

Huang, Q.; Fu, S.; Liu, J.; Jiang, H.; Yu, Y.; and Song, J. 2025. Resolving multi-condition confusion for finetuning-free personalized image generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 3707–3714.

Labs, B. F. 2024. FLUX. https://github.com/black-forest-labs/flux.

Labs, B. F.; Batifol, S.; Blattmann, A.; Boesel, F.; Consul, S.; Diagne, C.; Dockhorn, T.; English, J.; English, Z.; Esser, P.; Kulal, S.; Lacey, K.; Levi, Y.; Li, C.; Lorenz, D.; Müller, J.; Podell, D.; Rombach, R.; Saini, H.; Sauer, A.; and Smith, L. 2025. FLUX.1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space. arXiv:2506.15742.

Mou, C.; Wu, Y.; Wu, W.; Guo, Z.; Zhang, P.; Cheng, Y.; Luo, Y.; Ding, F.; Zhang, S.; Li, X.; et al. 2025. Dreamo: A unified framework for image customization. *arXiv preprint arXiv:2504.16915*.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Peebles, W.; and Xie, S. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4195–4205.

Peng, Y.; Cui, Y.; Tang, H.; Qi, Z.; Dong, R.; Bai, J.; Han, C.; Ge, Z.; Zhang, X.; and Xia, S.-T. 2025. DreamBench++: A Human-Aligned Benchmark for Personalized Image Generation. In *The Thirteenth International Conference on Learning Representations*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140): 1–67.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022a. High-Resolution Image Synthesis with Latent Diffusion Models. In *CVPR*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022b. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Ruiz, N.; Li, Y.; et al. 2023. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. In *CVPR*.

Tan, Z.; Xue, Q.; Yang, X.; Liu, S.; and Wang, X. 2025. Ominicontrol2: Efficient conditioning for diffusion transformers. *arXiv preprint arXiv:2503.08280*.

Wang, X.; Fu, S.; Huang, Q.; He, W.; and Jiang, H. 2025. MS-Diffusion: Multi-subject Zero-shot Image Personalization with Layout Guidance. In *The Thirteenth International Conference on Learning Representations*.

Wang, X.; Hu, S.; et al. 2023. UniAdapter: Parameter-Efficient Tuning for Text-to-Image Diffusion Models. arXiv:2311.01845.

Wu, C.; Zheng, P.; Yan, R.; Xiao, S.; Luo, X.; Wang, Y.; Li, W.; Jiang, X.; Liu, Y.; Zhou, J.; Liu, Z.; Xia, Z.; Li, C.; Deng, H.; Wang, J.; Luo, K.; Zhang, B.; Lian, D.; Wang, X.; Wang, Z.; Huang, T.; and Liu, Z. 2025a. OmniGen2: Exploration to Advanced Multimodal Generation. *arXiv preprint arXiv:2506.18871*.

Wu, S.; Huang, M.; Wu, W.; Cheng, Y.; Ding, F.; and He, Q. 2025b. Less-to-More Generalization: Unlocking More Controllability by In-Context Generation. *arXiv preprint arXiv:2504.02160*.

Xiao, B.; Wu, H.; Xu, W.; Dai, X.; Hu, H.; Lu, Y.; et al. 2024a. Florence-2: Advancing a Unified Representation for a Variety of Vision Tasks. In *CVPR*.

Xiao, J.; Yang, C.; Zhang, L.; Cai, S.; Zhao, Y.; Guo, Y.; Wetzstein, G.; Agrawala, M.; Yuille, A.; and Jiang, L. 2025. Captain Cinema: Towards Short Movie Generation. *arXiv preprint arXiv:2507.18634*.

Xiao, S.; Wang, Y.; Zhou, J.; Yuan, H.; Xing, X.; Yan, R.; Wang, S.; Huang, T.; and Liu, Z. 2024b. Omnigen: Unified image generation. *arXiv preprint arXiv:2409.11340*.

Yang, L.; Yu, Z.; Meng, C.; Xu, M.; Ermon, S.; and Cui, B. 2024. Mastering Text-to-Image Diffusion: Recaptioning, Planning, and Generating with Multimodal LLMs. In *International Conference on Machine Learning*.

Zhang, Y.; Yuan, Y.; Song, Y.; Wang, H.; and Liu, J. 2025. Easycontrol: Adding efficient and flexible control for diffusion transformer. *arXiv preprint arXiv:2503.07027*.

Zhou, X.; Yang, J.; et al. 2024. StoryDiffusion: Story-aware Visual Synthesis with Character Consistency. arXiv:2403.10023.

Zong, M.; Wu, Z.; Li, C.; et al. 2024. EasyRef: A Reference-Tuning-Free Framework for Multi-Image Composition. arXiv:2403.01887.