# ZERO-SHOT MULTI-CONCEPT IMAGE GENERATION VIA DECOUPLED CROSS AND SELF-ATTENTION IN MMDIT

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Generating high-fidelity images with multiple reference concepts is an important but challenging task. Existing methods either rely on additional training or suffer from identity blending, attribute leakage, and structural inconsistency in complex multi-reference scenarios. We propose **Decoupled Cross- and Self-Attention MMDiT (Deco-DiT)**, a training-free framework for multi-concept image generation. Deco-DiT decouples multi-reference conditioning into **Multi-Reference Cross-Attention (MR-CA)** for localized identity routing and **Spatial-Consistent Self-Attention (SC-SA)** for reference-target structural modeling. MR-CA routes encoder-level visual features to LLM-parsed spatial regions, while SC-SA enables target tokens to access structural and appearance cues from reference latents. Together with a stage-wise denoising schedule, Deco-DiT improves identity preservation, text alignment, and compositional consistency across challenging multi-reference settings without additional training.

## 1 INTRODUCTION

Text-to-image models Ho et al. (2020); Podell et al. (2023); Esser et al. (2024) have greatly advanced text-to-image generation, enabling the synthesis of high-quality and semantically aligned images from textual descriptions. With the growing demand for personalized image generationGuo et al. (2024); Wang et al. (2024b;a); Xing et al. (2024), multi-concept generation has become an important task, requiring models to preserve the identity and structure of multiple reference concepts while composing them into a coherent image.

Existing multi-subject generation methods can be categorized into three groups. The first group relies on subject-specific optimization or fine-tuning Ruiz et al. (2023); Kumari et al. (2023), achieving high fidelity but requiring additional training for each concept. The second group introduces reference encoders, fusion modules, or routing mechanisms learned from multi-concept data Wang et al. (2024c); Huang et al. (2025); Mou et al. (2025); Chen et al. (2025), improving scalability at the cost of dedicated training. The third group explores training-free inference strategies Feng et al. (2026); Dalva et al. (2026); Chen et al. (2026), avoiding optimization but still struggling with concept interference. Despite these advances, preserving identity and structural consistency across multiple references remains challenging. This issue is amplified in Diffusion Transformer architectures Esser et al. (2024), where different reference signals interact through shared attention operations, leading to identity leakage, attribute transfer, and degraded structural fidelity.

To address this challenge, we propose **Deco-DiT**, a training-free framework for multi-concept image generation based on Multimodal Diffusion Transformers (MMDiT). Deco-DiT decouples multi-reference conditioning into two complementary attention pathways. For identity preservation, Deco-DiT routes encoder-level reference features to their corresponding foreground regions, which enables each concept to be injected into its assigned spatial area while suppressing cross-concept mixing. We implement this process as **Multi-Reference Cross-Attention**. For structural consistency, Deco-DiT introduces reference latents into self-attention, allowing target regions to access fine-grained shape and appearance cues from the references. We implement this process as **Spatial-Consistent Self-Attention**. In addition, a stage-wise denoising schedule separates early foreground-background formation from later global semantic refinement, improving overall compo-



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

sitional coherence. With these designs, Deco-DiT faithfully preserves multiple reference concepts without fine-tuning or additional training.

The main contributions of this work are summarized as follows:

- We propose **Deco-DiT**, a training-free framework for high-fidelity multi-concept image generation. Built on **MMDiT**, it composes multiple reference concepts into a coherent scene without fine-tuning or test-time optimization.
- We introduce a heterogeneous attention decoupling mechanism that separates multi-reference conditioning into MR-CA for localized identity routing and SC-SA for reference-target structural modeling.
- Extensive experiments demonstrate that Deco-DiT achieves strong identity preservation, text alignment, and compositional coherence across different reference scales, especially in challenging multi-reference scenarios.

## 2 Related Work

### 2.1 Single-concept Personalized Generation

Single-concept personalized generation aims to synthesize a specific subject or object from a few reference images using pretrained diffusion models. Early methods rely on subject-specific optimization. Textual Inversion Gal et al. learns a dedicated text embedding, DreamBooth Ruiz et al. (2023) fine-tunes the diffusion model, and Custom Diffusion Choi et al. (2023) reduces the tuning cost by updating only a small subset of parameters. Although these methods achieve strong subject fidelity, they still require optimization for each new concept.

Later methods avoid subject-specific test-time optimization by introducing reference-based conditioning. IP-Adapter Ye et al. (2023) injects image features through decoupled cross-attention, while PhotoVerse Chen et al. (2023) and BLIP-Diffusion Li et al. (2023) improve subject-driven generation with image-text conditioning and vision-language pretraining. These methods provide important foundations for reference-guided generation, but they mainly focus on single-concept personalization and do not directly address preserving multiple reference concepts within one coherent image.

### 2.2 Multi-concept Image Generation

Multi-concept image generation aims to synthesize a coherent image from multiple reference inputs. Compared with single-concept personalization, this task is more challenging because different concepts may interfere with each other, leading to identity blending, attribute leakage, and reduced subject fidelity.

Recent methods address this problem from different perspectives. MS-Diffusion Wang et al. (2024c) uses layout guidance and subject-specific tokens to organize multiple concepts. OmniGen Xiao et al. (2025) and OmniGen2 Wu et al. (2025) unify text- and image-conditioned generation within a single framework. DreamO Mou et al. (2025) studies unified customization under the DiT architecture, while XVerse Chen et al. (2025) improves multi-subject identity control through modulation in large-scale DiTs.

Despite their progress, these methods often rely on large-scale training data, or additional model components. Moreover, their generation quality can degrade in complex multi-reference scenarios, where identity blending, attribute leakage, and weak compositional coherence remain common issues.

## 3 Method

### 3.1 Overview

Given multiple reference images $R = I_1, I_2, \ldots, I_N$ and a text prompt $P$, our goal is to generate an image that faithfully preserves each reference concept while maintaining coherent multi-concept composition.

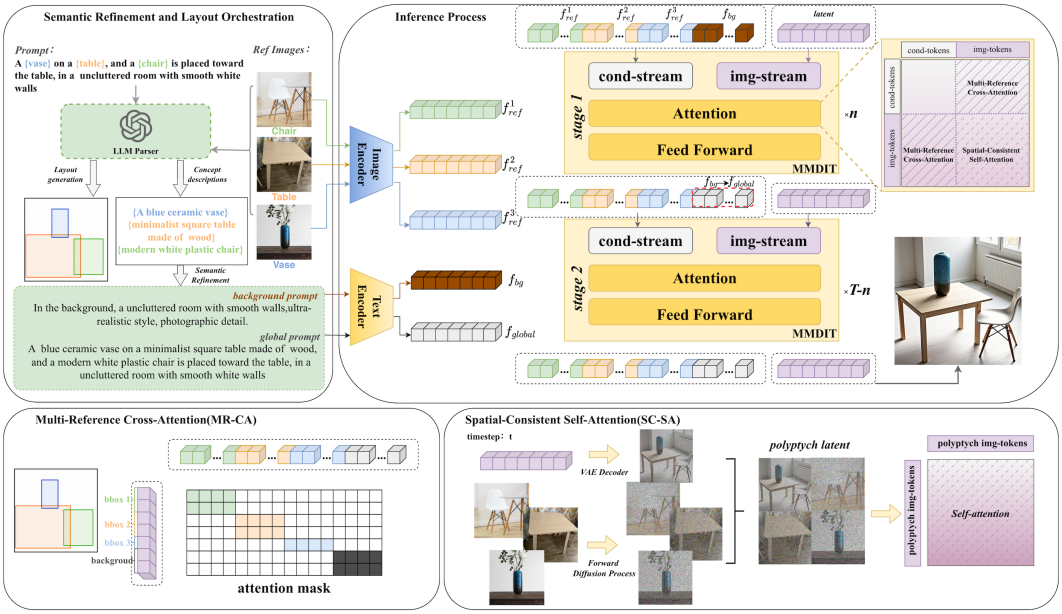



Figure 1: Overview of the Deco-DiT framework. Deco-DiT decouples identity preservation and scene composition via three components: LLM-driven Semantic ParsingChang et al. (2024) for layout prior construction, Multi-Reference Cross-Attention (MR-CA) for localized identity injection, and Spatial-Consistent Self-Attention (SC-SA) for structural harmonization. A progressive denoising strategy further enables localized concept formation and global scene integration.

To address the coupling between identity preservation and structural consistency in multi-reference generation, we propose Deco-DiT, a training-free framework that decouples multi-reference conditioning into localized identity routing and reference-target structural modeling. An LLM-based parser first extracts entity-level descriptions and coarse spatial priors from the prompt. Guided by these priors, **Multi-Reference Cross-Attention** routes encoder-level reference features to the corresponding target regions, suppressing cross-concept mixing. In parallel, **Spatial-Consistent Self-Attention** uses reference latents as structural anchors, enabling target regions to preserve fine-grained appearance and structural details.

Deco-DiT further adopts a stage-wise denoising strategy, which separates early local concept formation from later global semantic refinement to improve overall compositional coherence.

### 3.2 Semantic Layout Prior Construction

Although MR-CA and SC-SA provide decoupled pathways for reference-specific identity injection and structure-aware reference-target interaction, they still require explicit guidance on where each concept should be injected. In multi-concept prompts, different entities, attributes, and spatial relations are often described in a free-form manner. This makes it difficult for the model to establish reliable concept-to-region correspondence during denoising. To address this problem, we construct a semantic layout prior with an LLM-based parser. The parser converts the input prompt into structured entity-level representations, including concept descriptions and spatial constraints, which are then used to guide subsequent attention routing.

Given a text prompt $P$, the parser generates a semantic layout prior:

$$\mathcal{S} = \left( \{(T_i, B_i)\}_{i=1}^{N}, T_{\text{bg}} \right), \tag{1}$$

where $T_i$ denotes the concept-level textual description of the $i$-th reference concept, $B_i$ represents its corresponding spatial bounding box, and $T_{bg}$ describes the background context.

The concept descriptions provide attribute-aware semantic guidance for identity preservation, while the spatial layouts serve as explicit constraints for reference feature routing in MR-CA. The background description is further used for background-aware conditioning in the early denoising stage, which helps reduce semantic interference between foreground concepts and the surrounding environment.

### 3.3 Region-Guided Identity Preservation

The semantic layout prior assigns each reference concept to a specific spatial region. Based on this prior, we aim to preserve the visual identity of each concept within its corresponding target area. In multi-concept generation, visual features from different reference images may interfere with each other when they are injected through shared attention. Such interference can lead to identity mixing or attribute leakage. To address this problem, we guide the identity features of each reference concept toward its assigned region, while suppressing irrelevant information from other references.

**Multi-Reference Cross-Attention**: Existing MMDiT-based models process text tokens and image tokens within a unified attention space. This design supports flexible multimodal interactions. However, it does not explicitly specify which reference concept should guide each target region. As a result, target image tokens may attend to multiple unrelated references at the same time, which can cause identity leakage, attribute confusion, and concept blending in multi-reference generation.

To address this problem, we introduce Multi-Reference Cross-Attention (MR-CA), which decouples reference feature injection from the original unified attention pathway. MR-CA introduces a spatially constrained attention route for reference-specific visual features. Instead of allowing all reference concepts to interact with the whole target canvas, it routes each reference concept only to its assigned generation region according to the semantic layout prior. This design establishes an explicit correspondence between reference concepts and target regions, which helps preserve concept identity and reduce cross-concept interference.

Given $N$ reference images $\{I_i\}_{i=1}^{N}$,visual features for each reference concept with image encoder, denoted as $f_{\text{ref}}^{i} = \text{ImageEncoder}(I_i)$, where $f_{\text{ref}}^{i}$ represents the visual feature of the $i$-th reference concept. The extracted features are concatenated into a unified multi-reference condition representation, denoted as $F_{\text{ref}}$, which aggregates visual information from all reference concepts.

At each selected MMDiT block, MR-CA computes cross-attention between the target image tokens and reference features. Specifically, the query, key, and value representations are obtained as:

$$Q = X_{\text{img}} W_Q, \quad K = F_{\text{ref}} W_K, \quad V = F_{\text{ref}} W_V, \tag{2}$$

where $X_{\text{img}}$ denotes the target image tokens and $W_Q$, $W_K$, and $W_V$ are the projection matrices inherited from the pretrained MMDiT block. The attention operation is formulated as:

$$\text{MRCA}(Q, K, V, M) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}} + M\right) V, \tag{3}$$

where $d$ denotes the feature dimension and $M$ is an additive routing mask derived from the semantic layout prior.

Specifically, for an image token at spatial position $p_i$ and a reference feature token belonging to the $k$-th concept, the mask is defined as:

$$M_{i,j} = \begin{cases} 0, & p_i \in B_k,\ j \in f_{\text{ref}}^{k}, \\ -\infty, & \text{otherwise}, \end{cases} \tag{4}$$

where $B_k$ represents the bounding box of the $k$-th concept and $j$ denotes the index of reference feature tokens. This formulation ensures that the visual features of each reference concept mainly affects its assigned spatial region while preventing undesired interactions with other concepts.

By replacing unrestricted reference feature aggregation with spatially constrained concept routing, MR-CA reduces identity leakage and attribute interference. This helps preserve multiple reference concepts faithfully during generation.

### 3.4 Structural Consistency Modeling

MR-CA preserves concept identity by injecting reference features into spatially assigned target regions. However, it mainly focuses on region-level identity guidance and does not fully capture the structural correspondence between reference images and generated objects. As a result, the generated concepts may keep the correct category and attributes, but still suffer from shape distortion, missing details, or inconsistent appearance.

To address this problem, we introduce Spatial-Consistent Self-Attention (SC-SA). SC-SA treats reference latents as structural anchors and concatenates them with the target latent tokens in a shared self-attention space. This allows target regions to directly interact with their corresponding reference regions. Unlike MR-CA, which injects reference features in a directional manner, SC-SA builds bidirectional structural correspondence between reference concepts and generated regions. This helps preserve fine-grained shapes, textures, and appearance details during synthesis.

Given a set of reference images $\{I_i\}_{i=1}^{N}$, we first obtain their latent representations $\{z_{\text{ref}}^i\}_{i=1}^{N}$. Then, we construct timestep-aligned noisy reference latents through the forward noising process:

$$z_{\text{ref}}^i(t) = \mathcal{N}_t\left(z_{\text{ref}}^i, \epsilon^i\right), \quad \epsilon^i \sim \mathcal{N}(0, I), \tag{5}$$

where $\mathcal{N}_t$ denotes the scheduler-defined forward noisingHo et al. (2020) operation at timestep $t$. The obtained latent $z_{\text{ref}}^i(t)$ preserves the structural information of the corresponding reference concept at timestep $t$.

Instead of treating reference images as independent external conditions, we concatenate the reference latents with the target latent to construct a unified **polyptych latent space**. For example, when integrating three reference concepts, the polyptych latent can be spatially arranged as:

$$Z_{\text{poly}} = \begin{bmatrix} z_{\text{tg}} & z_{\text{ref}}^1 \\ z_{\text{ref}}^2 & z_{\text{ref}}^3 \end{bmatrix}. \tag{6}$$

Self-attention is then performed over the unified polyptych latent $Z_{\text{poly}}$, allowing target tokens to directly interact with reference tokens within a shared latent space. By performing self-attention in the concatenated reference-target latent space, target tokens can directly attend to structural cues from the reference latents. This helps the generated regions better preserve fine-grained shapes, textures, and appearance details from the corresponding reference concepts.

Compared with MR-CA, which focuses on localized identity injection, SC-SA provides complementary structural guidance at the latent level. It builds direct correspondence between reference regions and target regions through self-attention. Together, MR-CA and SC-SA separate the modeling of identity preservation and structural consistency. This enables high-fidelity multi-concept synthesis while maintaining a coherent overall composition.

Table 1: Quantitative comparison on multi-concept synthesis across different reference scales.

| **Method** | Single-Reference | | | Two-Reference | | | Three-Reference | | |
|---|---|---|---|---|---|---|---|---|---|
| | CLIP-T ↑ | CLIP-I ↑ | DINO-I ↑ | CLIP-T ↑ | CLIP-I ↑ | DINO-I ↑ | CLIP-T ↑ | CLIP-I ↑ | DINO-I ↑ |
| MS-Diffusion | 0.3564 | 0.8420 | 0.6029 | 0.3514 | 0.7400 | 0.4610 | 0.3381 | 0.7096 | 0.4522 |
| MIP-Adapter | 0.3461 | 0.9119 | 0.8126 | 0.3481 | 0.8407 | 0.5658 | 0.3020 | 0.8033 | 0.5409 |
| OmniGen2 | 0.3567 | 0.9478 | 0.8697 | 0.3597 | 0.8815 | 0.6891 | 0.3643 | 0.8276 | 0.6249 |
| DreamO | 0.3583 | 0.9547 | 0.8710 | 0.3588 | **0.9018** | **0.7366** | 0.3563 | 0.8385 | 0.6442 |
| Xverse | 0.3565 | **0.9656** | **0.8998** | 0.3516 | 0.8817 | 0.7073 | 0.3660 | 0.8448 | 0.6571 |
| **Ours** | **0.3638** | 0.9371 | 0.8524 | **0.3649** | 0.8874 | 0.7156 | **0.3662** | **0.8531** | **0.6705** |

### 3.5 Stage-wise Denoising Schedule

Although MR-CA and SC-SA provide region-guided pathways for identity preservation and structural consistency modeling, textual conditioning may still introduce semantic competition during denoising. If the full prompt is applied to the whole image from the beginning, different concepts may compete for the same textual cues. This can disturb the early formation of individual concepts, especially when coarse object layouts and foreground-background structures are being established.

To reduce this interference, we introduce a **Stage-wise Denoising Schedule**. This strategy keeps the reference-based guidance for foreground concepts unchanged, and only adjusts the textual guidance across different denoising stages. Let $t$ denote the denoising timestep, and let $\tau$ denote the transition timestep.

In the early stage ($t > \tau$), Deco-DiT uses the background description to guide only the background region, while the foreground regions are guided by reference visual features, reducing interference between foreground concepts and the background text. In the later stage ($t \leq \tau$), the background text is replaced with the original full prompt and applied to the whole image, allowing the foreground concepts and background to be adjusted under a unified semantic context and improving overall structural consistency and semantic coherence.

Therefore, the proposed stage-wise schedule separates early foreground-background formation from later global semantic refinement. In the early stage, reference visual features guide the foreground concepts, while the background text guides the background region. In the later stage, the full prompt is applied to the whole image to enhance global consistency. Combined with MR-CA and SC-SA, this strategy helps Deco-DiT preserve multiple reference concepts while producing a coherent multi-concept composition.

## 4 Experiments

### 4.1 Implementation Details

We adopt **FLUX.1-dev** Black Forest Labs (2024) as the base generative model. The reference images are encoded with SigLIP Zhai et al. (2023), and the text prompts are encoded with T5 Raffel et al. (2020). The denoising process is divided into two stages. In the first **20%** of the sampling steps, Deco-DiT applies localized conditioning: foreground regions are guided by their corresponding reference visual features, while the background region is guided by the background text. In the remaining **80%** of the sampling steps, the original full prompt is applied to the whole image for global semantic refinement. This stage improves overall structural consistency and semantic coherence.

We employ the FlowMatch Euler discrete scheduler Lipman et al. (2022); Liu et al. (2022) with 40 inference steps and a fixed guidance scale. Unless otherwise specified, all images are generated at a resolution of $768 \times 768$ on NVIDIA A800 GPUs with 80 GB memory.

### 4.2 Datasets, Baselines, and Evaluation Metrics

**Datasets:** Inspired by DreamO, we construct a multi-concept evaluation benchmark using **DreamBench++** Peng et al. (2024) and the **CustomDiffusion** dataset Choi et al. (2023) as the concept bank. We use an LLM to select semantically plausible concept combinations and generate prompts that cover diverse scenarios. The generated combinations and prompts are further checked to remove ambiguous or unrealistic cases.

For each concept number, we construct **10** concept combinations. Each combination is paired with **10** scenario-specific prompts, and **10** images are generated for each prompt using different random seeds. This results in **1,000** images for each concept number.

**Baselines:** We compare our method with representative multi-reference generation approaches across diverse architectural paradigms, including diffusion-based methods **MS-Diffusion** Wang et al. (2024c) and **MIP-Adapter** Huang et al. (2025), unified generative models **OmniGen** Xiao et al. (2025) and **OmniGen2** Wu et al. (2025), and reference-based generation models **DreamO** Mou et al. (2025) and **XVerse** Chen et al. (2025).




Figure 2: Comparative results of **Deco-DiT** and baseline models across diverse multi-reference composition scenarios.

**Evaluation Metrics:** We evaluate multi-concept generation from two aspects: reference preservation and prompt alignment. Since each generated image contains multiple subjects and background content, image-level similarity may be biased by irrelevant regions. Therefore, we use SAM2 Ravi et al. (2024) to extract foreground subject regions and compute subject-level similarities between each reference concept and its corresponding generated region.

For reference preservation, we report CLIP-I and DINO-I, which measure cosine similarity in the CLIP Radford et al. (2021) and DINO Oquab et al. (2023) feature spaces, respectively. CLIP-I reflects high-level semantic and appearance consistency, while DINO-I is more sensitive to fine-grained structures and visual details. The subject-level scores are averaged over all reference concepts in each generated image. For prompt alignment, we report CLIP-T, which measures the semantic consistency between the complete generated image and the input prompt.

### 4.3 Quantitative Comparison

Table 1 compares Deco-DiT with representative baselines under one-, two-, and three-reference settings. Deco-DiT achieves the highest CLIP-T score across all reference scales, with scores of 0.3638, 0.3649, and 0.3662, respectively. This indicates that our method maintains strong prompt alignment while incorporating multiple reference concepts.

In the one-reference setting, some trained or reference-encoder-based methods obtain slightly higher identity-preservation scores. This is reasonable, as these methods can benefit from learned reference modeling or additional adaptation, whereas Deco-DiT is training-free and operates on frozen MMDiT representations. Nevertheless, Deco-DiT remains competitive on CLIP-I and DINO-I while

achieving the best CLIP-T score, showing a favorable balance between subject fidelity and text alignment.

As the number of references increases, the advantage of Deco-DiT becomes more evident. In the two-reference setting, Deco-DiT achieves the best CLIP-T score and the second-best DINO-I score, indicating stable prompt adherence and structural preservation under stronger cross-concept interference. In the more challenging three-reference setting, Deco-DiT achieves the best performance across all three metrics, with CLIP-T, CLIP-I, and DINO-I scores of 0.3662, 0.8531, and 0.6705, respectively. These results demonstrate that Deco-DiT scales more robustly to complex multi-concept compositions, achieving better overall consistency without fine-tuning or additional training.

### 4.4 Qualitative Comparison

Figure 2 presents qualitative comparisons across diverse two- and three-reference composition scenarios. Deco-DiT preserves the identity of reference concepts while maintaining reasonable object structures and overall composition.

In the two-reference setting, differences among methods are already apparent. XVerse and MIP-Adapter often suffer from attribute leakage and semantic entanglement, while DreamO may lose fine-grained textures and appearance details for complex concepts. In contrast, Deco-DiT better retains the distinctive visual characteristics of each reference concept and produces more plausible subject interactions.

In more challenging three-reference scenarios, cross-concept interference becomes stronger. Although OmniGen2 benefits from a unified architecture, its results may still show weak integration between foreground subjects and the background. Existing methods also tend to struggle with rare or unusual concept combinations, where identity confusion and structural distortion become more visible. By combining localized identity routing, reference-target structural modeling, and stage-wise denoising, Deco-DiT reduces these artifacts and generates more reliable multi-concept compositions.

### 4.5 Ablation Study

We conduct ablation studies on stage-wise text guidance, spatial layout guidance, and spatial-consistent self-attention (SC-SA). **Global Guid.** denotes the later full-prompt guidance, **Background Guid.** denotes the early background guidance, **Layout Control** denotes spatial masks, and **SC-SA** denotes the reference-latent interaction in self-attention. Quantitative results are reported in Table 2, and qualitative comparisons are shown in Figure 3.

**Stage-wise Text Guidance.** We first analyze the proposed stage-wise text guidance strategy. Deco-DiT uses the background prompt in the early denoising stage to guide background formation, and switches to the full prompt in the later stage for global semantic refinement. This design aims to reduce foreground-background conflicts while improving overall compositional consistency.

To validate its effectiveness, we compare Deco-DiT with two variants: using only the background prompt or only the full prompt throughout denoising. As shown in Table 2, removing later global guidance reduces CLIP-T from 0.3662 to 0.3115, while removing early background guidance decreases DINO-I from 0.6705 to 0.6182, demonstrating the importance of both guidance stages. Figure 3 further shows that background-only guidance causes foreground-background semantic inconsistency, whereas full-prompt-only guidance introduces conflicts between global semantics and local visual features, reducing reference fidelity.

**Spatial Layout Guidance.** We also evaluate the effect of spatial layout guidance by comparing results with and without predicted bounding boxes. As shown in Table 2, removing layout guidance leads to a decline across all three metrics. The drop in CLIP-T is the most noticeable, decreasing from 0.3662 to 0.3395.

The qualitative results in Figure 3 show a similar trend. Without layout guidance, the model relies mainly on global prompts to organize multiple concepts. This often leads to spatial overlap, incorrect object placement, and identity blending. In contrast, the predicted bounding boxes provide a structured spatial prior. This prior enables MR-CA to route each reference feature to its assigned

region, helping each concept appear in the correct location while reducing interference from other concepts.

Table 2: Ablation study of different model settings

| Method | CLIP-T ↑ | CLIP-I ↑ | DINO-I ↑ |
|---|---|---|---|
| w/o global guid. | 0.3115 | 0.8513 | 0.6673 |
| w/o background guid. | 0.3589 | 0.8469 | 0.6182 |
| w/o layout control | 0.3395 | 0.8162 | 0.6312 |
| w/o SC-SA | 0.3630 | 0.7966 | 0.5634 |
| **Ours** | **0.3662** | **0.8531** | **0.6705** |

**Spatial-Consistent Self-Attention.** We evaluate the contribution of SC-SA by replacing it with standard self-attention. In this ablation setting, the target latents can no longer directly interact with the reference latents through the proposed spatial-consistent pathway. As shown in Table 2, removing SC-SA leads to a clear performance drop. CLIP-I decreases from 0.8531 to 0.7966, and DINO-I drops from 0.6705 to 0.5634. This indicates that SC-SA is important for preserving reference identity and fine-grained structure.

The fourth column of Figure 3 provides qualitative evidence. Without SC-SA, the generated concepts often lose detailed geometric structures from the reference images. This results in weaker identity preservation and more visible structural distortion. These results show that SC-SA provides complementary structural guidance beyond the region-guided identity injection of MR-CA.

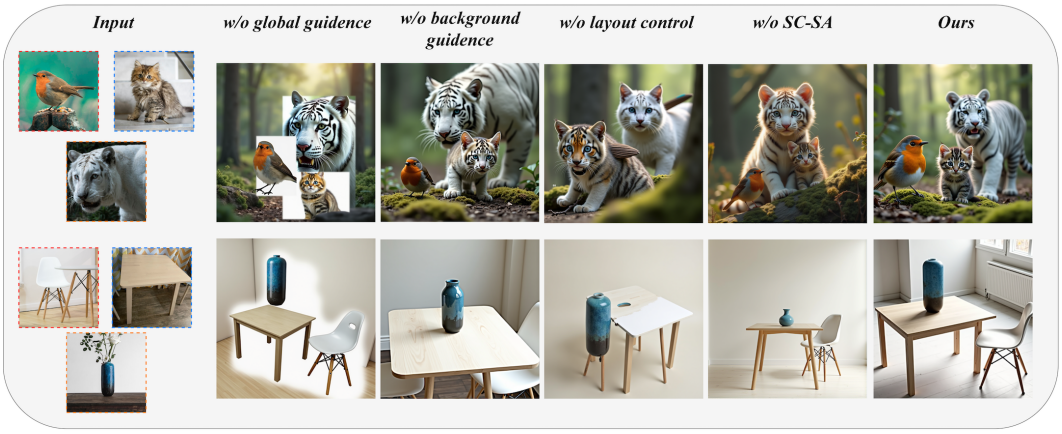



Figure 3: Ablation study of Deco-DiT. We evaluate the contribution of each component, under three-reference settings.

## 5 Conclusion and Limitation

In this paper, we presented **Deco-DiT**, a training-free framework for multi-concept image generation with MMDiT. Deco-DiT uses MR-CA for localized identity preservation and SC-SA for structural and appearance consistency. A stage-wise denoising strategy further improves the coordination between multiple concepts and the overall scene. Extensive experiments demonstrate strong performance in identity preservation, text alignment, and compositional coherence.

Despite these advantages, Deco-DiT still has some limitations. As a training-free method, its performance depends on the generation capability of the base model and the representation quality of the visual encoder. In addition, introducing multiple reference latents increases the attention sequence length, leading to higher memory consumption and inference latency. Future work will explore more efficient reference modeling to improve scalability with a larger number of concepts.

## REFERENCES


Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45, 2024.

Bowen Chen, Mengyi Zhao, Haomiao Sun, Li Chen, Xu Wang, Kang Du, and Xinglong Wu. Xverse: Consistent multi-subject control of identity and semantic attributes via dit modulation. *arXiv preprint arXiv:2506.21416*, 2025.

Li Chen, Mengyi Zhao, Yiheng Liu, Mingxu Ding, Yangyang Song, Shizun Wang, Xu Wang, Hao Yang, Jing Liu, Kang Du, et al. Photoverse: Tuning-free image customization with text-to-image diffusion models. *arXiv preprint arXiv:2309.05793*, 2023.

Yuzhuo Chen, Zehua Ma, Jianhua Wang, Kai Kang, Shunyu Yao, and Weiming Zhang. Lamic: Layout-aware multi-image composition via scalability of multimodal diffusion transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 3174–3182, 2026.

Jooyoung Choi, Yunjey Choi, Yunji Kim, Junho Kim, and Sungroh Yoon. Custom-edit: Text-guided image editing with customized diffusion models. *arXiv preprint arXiv:2305.15779*, 2023.

Yusuf Dalva, Hidir Yesiltepe, and Pinar Yanardag. Lorashop: Training-free multi-concept image generation and editing with rectified flow transformers. *Advances in Neural Information Processing Systems*, 38:167841–167876, 2026.

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. In *Forty-first international conference on machine learning*, 2024.

Haoran Feng, Zehuan Huang, Lin Li, and Lu Sheng. Personalize anything for free with diffusion transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 3921–3929, 2026.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *The Eleventh International Conference on Learning Representations*.

Zinan Guo, Yanze Wu, Zhuowei Chen, Lang Chen, Peng Zhang, and Qian He. Pulid: Pure and lightning id customization via contrastive alignment. *Advances in neural information processing systems*, 37:36777–36804, 2024.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Qihan Huang, Siming Fu, Jinlong Liu, Hao Jiang, Yipeng Yu, and Jie Song. Resolving multi-condition confusion for finetuning-free personalized image generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 3707–3714, 2025.

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1931–1941, June 2023.

Dongxu Li, Junnan Li, and Steven Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *Advances in Neural Information Processing Systems*, 36:30146–30166, 2023.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Chong Mou, Yanze Wu, Wenxu Wu, Zinan Guo, Pengze Zhang, Yufeng Cheng, Yiming Luo, Fei Ding, Shiwen Zhang, Xinghui Li, et al. Dreamo: A unified framework for image customization. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 1–12, 2025.

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.

Yuang Peng, Yuxin Cui, Haomiao Tang, Zekun Qi, Runpei Dong, Jing Bai, Chunrui Han, Zheng Ge, Xiangyu Zhang, and Shu-Tao Xia. Dreambench++: A human-aligned benchmark for personalized image generation. *arXiv preprint arXiv:2406.16855*, 2024.

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. *arXiv preprint arXiv:2307.01952*, 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 *International conference on machine learning*, pp. 8748–8763. PmLR, 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. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22500–22510, 2023.

Haofan Wang, Matteo Spinelli, Qixun Wang, Xu Bai, Zekui Qin, and Anthony Chen. Instantstyle: Free lunch towards style-preserving in text-to-image generation. *arXiv preprint arXiv:2404.02733*, 2024a.

Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, Anthony Chen, Huaxia Li, Xu Tang, and Yao Hu. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*, 2024b.

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*, 2024c.

Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, et al. Omnigen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13294–13304, 2025.

Peng Xing, Haofan Wang, Yanpeng Sun, Qixun Wang, Xu Bai, Hao Ai, Renyuan Huang, and Zechao Li. Csgo: Content-style composition in text-to-image generation. *arXiv preprint arXiv:2408.16766*, 2024.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11975–11986, 2023.