MoLaGen: Multimodal Conditional Layout Control for Image Generation

Anonymous authors
Paper under double-blind review

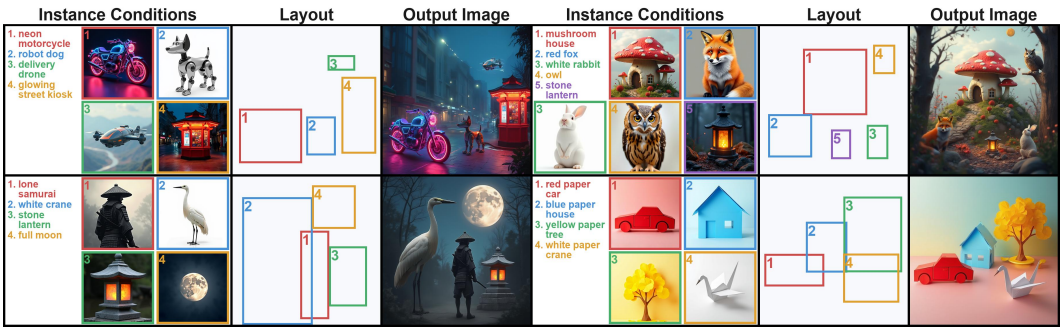


Figure 1: **Multimodal conditional layout generation results of MoLaGen**. For each instance, the local prompt, reference image, and bounding box are marked with the same color and index. MoLaGen preserves the key features conveyed by the instance conditions while adapting each instance to the specified layout and overall scene context.

Abstract

In layout-to-image generation, instances are typically specified by local text prompts and spatial coordinates. However, local text descriptions provide limited precision in specifying instance appearance, whereas reference images offer richer and more fine-grained visual guidance. Achieving such multimodal conditional layout control requires effectively encoding and fusing visual, textual, and spatial representations for instance-level conditioning. Existing approaches typically rely on separate encoders and fusion modules, which may introduce representation mismatch and additional complexity. We therefore propose the **M**ultim**o**dal Conditional **La**yout Controllable **Gen**erative Network (MoLaGen), a unified framework that uses the MM-DiT backbone itself to encode local prompts and reference images and concatenates the resulting tokens as multimodal conditions. For spatial control, we introduce a spatially aware CondLoRA network with dedicated spatial mapping weights and token-dependent routing between visual and textual branches. We further develop a three-stage data construction pipeline for image synthesis, instance-level annotation, and quality filtering. Qualitative and quantitative experiments demonstrate the effectiveness of MoLaGen.

1 Introduction

Diffusion models Ho et al. (2020); Song et al. (2020); Rombach et al. (2022) have achieved remarkable success in image generation through iterative denoising. Image-generation backbones have evolved from U-Net-based architectures, as used in Stable Diffusion Rombach et al. (2022); Podell et al. (2023), toward MM-DiT architectures Peebles & Xie (2023); Esser et al. (2024); Black-Forest-Labs (2024); Labs et al. (2025). Despite these advances, generating scenes with multiple instances remains challenging because text conditions provide limited control over instance locations. Layout-to-image generation addresses this limitation by introducing explicit spatial constraints into the generation process.

Recent layout-to-image methods Li et al. (2023c); Zhou et al. (2024); Zhang et al. (2025c) achieve explicit instance-level control by pairing local text prompts with corresponding spatial coordinates.

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.

However, text descriptions may not fully capture the distinctive appearance of each instance, making fine-grained identity control difficult. Reference images provide direct visual cues for instance appearance and complement the semantic information conveyed by text. Combining these heterogeneous conditions poses a key challenge: *how to jointly encode and fuse visual, textual, and spatial representations within a unified generative framework*.

Existing methods typically rely on modality-specific encoders. Spatial coordinates are commonly transformed via Fourier embeddings Mildenhall et al. (2021) to capture high-frequency positional variations. For the textual modality, several works Li et al. (2023c); Zhou et al. (2024); Wang et al. (2024) rely on CLIP Radford et al. (2021) to encode local prompts. However, the resulting representations may not be fully aligned with the internal latent space of the generative model and may further suffer from information loss when fused with spatial tokens. For visual conditions, earlier approaches Ye et al. (2023); Li et al. (2023a); Zhang et al. (2024); Patel et al. (2024) extract reference features using pretrained image encoders. In contrast, recent MM-DiT–based methods Tan et al. (2025); Wu et al. (2025) encode reference images using the generative backbone itself along with its associated VAE, thereby better preserving pretrained diffusion knowledge. Since MM-DiT jointly processes visual and textual tokens, it provides a natural architecture for unified multimodal conditioning. To address this challenge, we propose MoLaGen, which jointly incorporates instance-level visual, textual, and spatial conditions, as illustrated in Figure 1.

Layout conditioning is commonly integrated at the cost of increased computational complexity. Classical approaches introduce additional cross-attention layers to fuse noisy image tokens with layout conditioning tokens. Laytrol Huang et al. (2026) further improves generation quality by initializing layout control modules through parameter copying from existing MM-DiT blocks. To reduce this computational overhead, we propose a spatially aware CondLoRA fine-tuning mechanism. Extending the standard LoRA formulation, we introduce an additional mapping matrix $\mathbf{C}$ to project spatial tokens into a low-rank subspace and integrate them with tokens derived from reference images and local prompts. Compared with adapter-based fusion approaches, this low-rank, modality-aware fusion strategy substantially reduces the number of additional trainable parameters. Furthermore, CondLoRA adopts a dual-branch architecture that dynamically routes conditioning tokens to visual or textual branches according to their modalities. This modality-specific routing accommodates the heterogeneous representations of visual and textual conditions.

Constructing a high-quality training dataset for this task requires accurate annotations and diverse training samples. Segmentation-based methods Peng et al. (2025); Xu et al. (2026); Wang et al. (2025) employ SAM Kirillov et al. (2023); Ravi et al. (2024); Carion et al. (2026) to extract reference instances directly from source images, but the resulting references provide limited variations in object appearance and background, restricting subject generalization. Diptych-based approaches Tan et al. (2025); Wu et al. (2025) generate original-reference image pairs through a cascaded process, but may produce imprecise image boundaries or inconsistent subjects. To address these limitations, we propose a three-stage data construction pipeline comprising image synthesis, instance annotation, and quality filtering. We generate reference images using Kontext Labs et al. (2025) with prompts designed to preserve entity identity while varying the background, pose, and viewpoint. Additionally, we apply multi-dimensional quality filtering to improve annotation accuracy. These diverse and accurately annotated samples help MoLaGen generate coherent scenes that preserve condition-specific features while appropriately adapting each instance to the specified layout and global context, as shown in Figure 1.

Our contributions are summarized as follows:

- We propose **CondLoRA**, a low-rank adaptation method that introduces an additional mapping matrix to project spatial tokens into a low-rank subspace for efficient multimodal condition fusion.
- We introduce the **M**ultim**o**dal Conditional **La**yout Controllable **Gen**erative Network (**MoLaGen**), which combines dual-branch CondLoRA with modality-aware routing to integrate visual, textual, and spatial conditions within a unified generative framework.
- We develop a three-stage data construction pipeline comprising image synthesis, instance annotation, and quality filtering to improve reference diversity and annotation accuracy.

2 Related Work

2.1 Text-to-Image Generation

Early studies employed Variational Autoencoders (VAEs) Kingma & Welling (2013); Larsen et al. (2016) or Generative Adversarial Networks (GANs) Goodfellow et al. (2014); Mirza & Osindero (2014); Reed et al. (2016); Zhang et al. (2017) for text-to-image (T2I) generation. Diffusion models (Ho et al., 2020; Song et al., 2020) learn the mapping between noise (Li, 2024; Zhang et al., 2025b) and data, while the Latent Diffusion Model (LDM) (Rombach et al., 2022) improves training efficiency by learning this mapping in a latent space encoded by a VAE (Kingma & Welling, 2013) rather than directly in pixel space. Using T5 Raffel et al. (2020) instead of CLIP Radford et al. (2021) improves text understanding in models such as Imagen Saharia et al. (2022), DALL·E 3 Betker et al. (2023), and PixArt-α Chen et al. (2024). To improve the scalability of U-Net-based architectures Ronneberger et al. (2015), Diffusion Transformers (DiT) Peebles & Xie (2023) replace the U-Net backbone with a Transformer Vaswani et al. (2017). With large-scale training, recent models such as Stable Diffusion 3 Esser et al. (2024) and FLUX Black-Forest-Labs (2024); Black Forest Labs (2025) achieve strong generation performance.

2.2 Subject-Driven Image Generation

Subject-driven image generation aims to synthesize images containing specified objects or styles. Textual Inversion Gal et al. (2023) and DreamBooth Ruiz et al. (2023) achieve subject-driven generation through fine-tuning strategies that use reference images as training data. However, fine-tuning a separate model for each subject is computationally expensive and time-consuming. Several methods Ye et al. (2023); Li et al. (2023a); Zhang et al. (2024); Patel et al. (2024) design and train dedicated visual encoders and fusion modules to extract subject-specific representations from reference images. In contrast, IC-LoRA Huang et al. (2024) and OminiControl Tan et al. (2025) treat reference images as visual context and employ LoRA Hu et al. (2022) to fine-tune the visual branch of MM-DiT for visual-context understanding. Similarly, Kontext Labs et al. (2025) formulates subject-driven image generation as a context-understanding problem and is trained from scratch on the FLUX Black-Forest-Labs (2024) architecture.

2.3 Layout-Driven Image Generation

In classical layout-driven image generation, each instance is specified by a bounding box and an associated local prompt. Training-free methods Bar-Tal et al. (2023); Xie et al. (2023); Yang et al. (2023; 2024) use attention masking to couple latent features within each bounding box with the corresponding local-prompt embeddings. These approaches typically incur increased inference time and computational overhead while providing relatively low layout-control accuracy, posing a deployment challenge (An et al., 2026; Shao & Li, 2026). Training-based methods such as GLIGEN Li et al. (2023c) and MIGC Zhou et al. (2024) are built on U-Net-based Stable Diffusion. They encode bounding-box coordinates with Fourier embeddings Mildenhall et al. (2021) and local prompts with CLIP. The resulting feature streams are fused into layout-aware representations, which interact with image latents through newly trained adapter layers. More recently, SiamLayout Zhang et al. (2025c) and Laytrol Huang et al. (2026) extend layout control to MM-DiT architectures.

Beyond text-guided layout control, subject-guided layout control specifies each instance using a bounding box and a reference image. LAMIC Chen et al. (2026) uses attention-based interactions for training-free layout control. Several methods Wang et al. (2025); Ma et al. (2024); Peng et al. (2025); Li et al. (2025) introduce dedicated visual encoding and fusion modules to process reference images. However, improving image quality and subject similarity remains challenging.

3 Preliminaries

3.1 Multimodal Diffusion Transformer

The Multimodal Diffusion Transformer (MM-DiT) architecture is widely adopted in image generation models, such as Stable Diffusion 3 Esser et al. (2024), FLUX Black-Forest-Labs (2024), and

Kontext Labs et al. (2025). MM-DiT jointly processes image tokens $\mathbf{I} \in \mathbb{R}^{n \times d}$ and text tokens $\mathbf{T} \in \mathbb{R}^{m \times d}$ within a unified framework. MM-DiT consists of double-stream and single-stream blocks. Double-stream blocks employ separate projection layers for image and text tokens, whereas single-stream blocks use shared projection layers to jointly process both modalities. These projection layers map image tokens to $\mathbf{Q_I}, \mathbf{K_I}, \mathbf{V_I}$ and text tokens to $\mathbf{Q_T}, \mathbf{K_T}, \mathbf{V_T}$. In addition, 3D Rotary Position Embedding (RoPE) Su et al. (2024) is applied to model the relative positions among tokens. Specifically, each image token is assigned a positional triplet (t, i, j), where t identifies the image sequence and (i, j) denote the spatial coordinates of the corresponding image patch. In Kontext Labs et al. (2025), target image tokens use $t = 0$, while reference image tokens use $t = 1$. For text tokens, the positional triplet is set to $(t, i, j) = (0, 0, 0)$.

The positional triplet (t, i, j) parameterizes the 3D RoPE rotation matrix $\mathbf{R}$, which is applied to the query $\mathbf{Q}$ and key $\mathbf{K}$ before attention. Subsequently, the concatenated tokens are processed via scaled dot-product attention:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}} + \mathbf{M}\right)\mathbf{V}, \tag{1}$$

where $\mathbf{M}$ is an additive attention mask whose entries are 0 for valid interactions and $-\infty$ otherwise.

3.2 Low-Rank Adaptation (LoRA)

For a pretrained weight matrix $\mathbf{W}_0 \in \mathbb{R}^{k \times d}$, LoRA Hu et al. (2022) parameterizes its update using two learnable matrices $\mathbf{A} \in \mathbb{R}^{r \times d}$ and $\mathbf{B} \in \mathbb{R}^{k \times r}$, where $r \ll \min(k, d)$. During fine-tuning, $\mathbf{W}_0$ remains frozen and only $\mathbf{A}$ and $\mathbf{B}$ are optimized, reducing the number of trainable parameters from kd to $r(d + k)$. Given an input $\mathbf{x} \in \mathbb{R}^d$, the LoRA-augmented forward pass is

$$\mathbf{x}' = \mathbf{W}_0\mathbf{x} + \mathbf{B}\mathbf{A}\mathbf{x}. \tag{2}$$

Matrix $\mathbf{A}$ is randomly initialized, while $\mathbf{B}$ is initialized to zeros, such that $\mathbf{B}\mathbf{A} = 0$ and the layer initially preserves the pretrained mapping.

3.3 Task Definition

In multimodal conditional layout generation, the generation process is guided by a global prompt p_g and local conditions for N entities $\mathbf{e} = \{\mathbf{e}_i\}_{i=1}^N$. For each entity $\mathbf{e}_i$, a set of local conditions is provided, including a local prompt p_i, a reference image I_i, and a bounding box $\mathbf{b}_i$. The synthesized image X should follow the global semantics specified by p_g. Within each bounding box $\mathbf{b}_i$, the corresponding instance should match the local description p_i while preserving the identity of the subject in the reference image I_i.

4 MoLaGen

4.1 Image Generation under Multimodal Conditions

As described in Section 3.3, the task takes a global prompt p_g and N instance-level tuples $\{(I_i, p_i, \mathbf{b}_i)\}_{i=1}^N$ as conditions to synthesize an image X. These conditions span three modalities: visual (I_i), textual (p_g and p_i), and spatial ($\mathbf{b}_i$). Accordingly, we leverage MM-DiT to jointly process the visual and textual conditions, while incorporating the spatial conditions through CondLoRA.

Existing methods Li et al. (2023c); Ye et al. (2023); Zhang et al. (2024) typically employ separate encoders and fusion modules for different conditioning modalities. The representations obtained from the encoders used in these approaches (e.g., CLIP Radford et al. (2021)) may exhibit representational discrepancies with respect to the internal representations of MM-DiT. In contrast, recent methods Tan et al. (2025); Wu et al. (2025); Labs et al. (2025) leverage MM-DiT and its native VAE to encode and incorporate reference images I_i. These methods acquire visual in-context generation capabilities through LoRA fine-tuning Hu et al. (2022) or full-model training. LoRA-based variants introduce few trainable parameters while largely preserving pretrained visual knowledge. Since MM-DiT jointly processes visual and textual tokens, its native textual pathway can likewise incorporate local prompts p_i for localized control over specific regions.

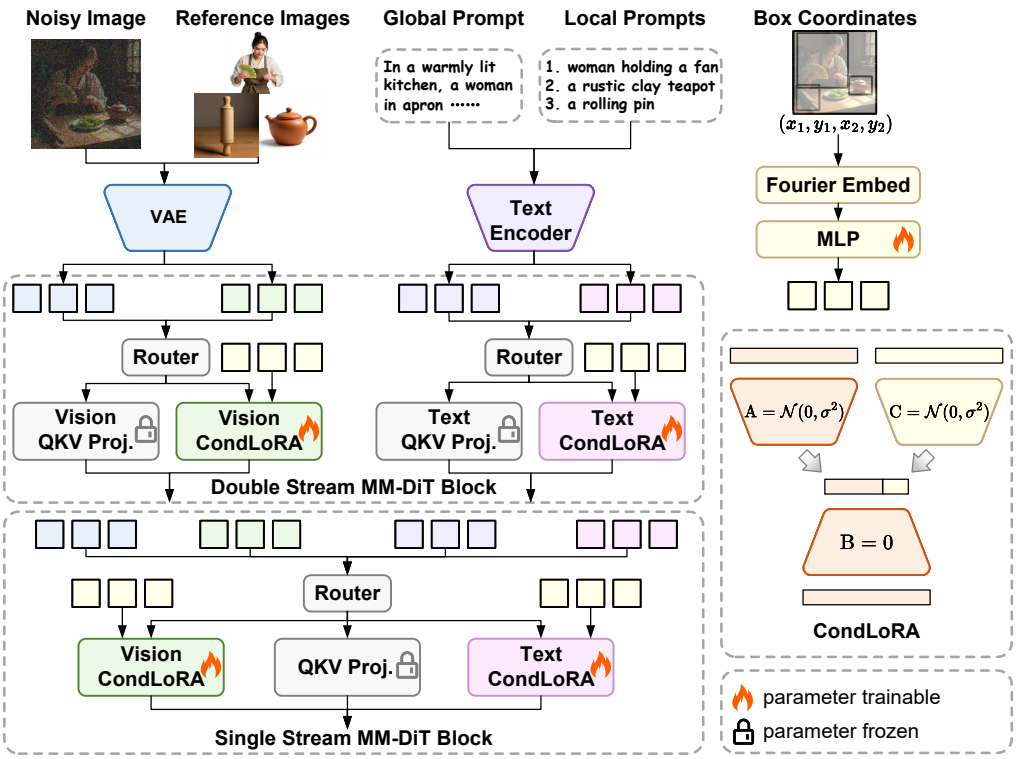


Figure 2: **Overview of the MoLaGen framework**. CondLoRA projects spatial tokens into a low-rank subspace and concatenates them with the low-rank features of the corresponding input tokens. MoLaGen employs Vision CondLoRA and Text CondLoRA, which are selectively activated for reference-image tokens and local-prompt tokens, respectively.

However, MM-DiT lacks an explicit mechanism for encoding and incorporating spatial conditions. We therefore introduce spatially aware CondLoRA, as described in Section 4.2. Specifically, CondLoRA introduces a mapping matrix $\mathbf{C}$ that projects spatial tokens into a low-rank space and fuses them with the low-rank features of the corresponding reference-image and local-prompt tokens. To handle visual and textual tokens separately, we design a routing strategy that selectively activates the corresponding CondLoRA branch, as detailed in Section 4.3. Finally, we introduce a dedicated attention mask to prevent information leakage across different instance-level conditions, as described in Section 4.4.

4.2 Spatially Aware CondLoRA

Each entity $\mathbf{e}_i$ is associated with a bounding box $\mathbf{b}_i = (x_1, y_1, x_2, y_2)$ that specifies its spatial extent. We encode $\mathbf{b}_i$ into a spatial token $\mathbf{s}_i \in \mathbb{R}^{d_s}$ using Fourier features Mildenhall et al. (2021) followed by a multi-layer perceptron (MLP):

$$\mathbf{s}_i = \mathrm{MLP}(\mathrm{Fourier}(\mathbf{b}_i)). \tag{3}$$

To incorporate spatial tokens into LoRA, we introduce a learnable projection matrix $\mathbf{C} \in \mathbb{R}^{r_s \times d_s}$ based on the formulation in Section 3.2. Accordingly, $\mathbf{B}$ is expanded from $\mathbb{R}^{k \times r}$ to $\mathbb{R}^{k \times (r + r_s)}$. Given an input token $\mathbf{x}$ and its corresponding spatial token $\mathbf{s}_i$, the CondLoRA forward pass is

$$\mathbf{x}' = \mathbf{W}_0\mathbf{x} + \mathbf{B}\big(\mathbf{A}\mathbf{x} \oplus \mathbf{C}\mathbf{s}_i\big), \tag{4}$$

where $\oplus$ denotes the concatenation operation. Following LoRA, $\mathbf{A}$ and $\mathbf{C}$ are randomly initialized, while $\mathbf{B}$ is initialized to zeros, ensuring that the residual branch is initially zero and the pretrained mapping remains unchanged. Compared with cross-attention-based layout controllers, CondLoRA introduces only low-rank projection parameters, providing a parameter-efficient mechanism for spatial conditioning.

4.3 CondLoRA Routing Strategy

MoLaGen builds on the pretrained Kontext Labs et al. (2025) backbone. During data preprocessing, noisy images, reference images, global prompts, and local prompts are encoded by a VAE encoder and a T5 text encoder Raffel et al. (2020) into their corresponding token representations, denoted as $\mathbf{I}$, $\mathbf{I_C}$, $\mathbf{T}$, and $\mathbf{T_C}$, respectively. The reference-image and local-prompt token sequences are concatenated across the N instances:

$$\mathbf{I_C} = \bigoplus_{i=1}^{N} \mathbf{I_C^i}, \qquad \mathbf{T_C} = \bigoplus_{i=1}^{N} \mathbf{T_C^i}. \tag{5}$$

The MM-DiT input sequence $\mathbf{X}$ is formed by concatenating the four token sequences:

$$\mathbf{X} = \mathbf{I} \oplus \mathbf{T} \oplus \mathbf{I_C} \oplus \mathbf{T_C}. \tag{6}$$

The routing strategy applies CondLoRA only to the condition tokens. Noisy-image tokens $\mathbf{I}$ and global-prompt tokens $\mathbf{T}$ follow the pretrained pathway without activating CondLoRA. To accommodate the different representations of $\mathbf{I_C}$ and $\mathbf{T_C}$, we introduce two specialized branches: Vision CondLoRA and Text CondLoRA. In double-stream blocks, Vision CondLoRA and Text CondLoRA are attached to the projection layers of the visual and textual streams, respectively. In single-stream blocks, both branches are attached to the shared projection layers, where the routing strategy activates Vision CondLoRA for reference-image tokens and Text CondLoRA for local-prompt tokens. For each instance i, both $\mathbf{I_C^i}$ and $\mathbf{T_C^i}$ are paired with the corresponding spatial token $\mathbf{s}_i$ when routed through their respective CondLoRA branches.

4.4 Instance-Isolated Attention Mask

To prevent cross-instance information leakage, we design an instance-isolated attention mask that restricts interactions among condition tokens from different instances. Specifically, noisy-image tokens retain global attention and can attend to all token groups. Global-prompt tokens attend only to noisy-image tokens and do not directly interact with instance-level condition tokens. In addition to noisy-image tokens, local-prompt and reference-image tokens can attend only to condition tokens belonging to the same instance, while cross-instance attention is blocked. The resulting attention pattern is illustrated in Figure 3.

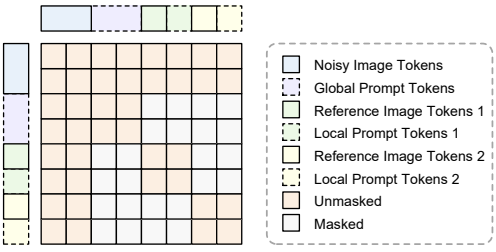


Figure 3: **Instance-isolated attention mask**. Attention between condition tokens from different instances is blocked to prevent cross-instance information leakage.

4.5 Construction of Training Data

In multimodal conditional layout generation, each training sample must provide accurate bounding boxes, local prompts, and reference images while maintaining sufficient diversity between target and reference instances. To meet these requirements, we construct the training data through three stages: image synthesis, instance annotation, and quality filtering, as illustrated in Figure 4. The first stage generates diverse target scenes, the second constructs instance-level spatial, textual, and visual conditions, and the final stage removes low-quality or inconsistent samples. The resulting dataset contains approximately 400,000 images. Detailed motivation and implementation are provided in Appendix A.

5 Experiments

5.1 Experimental Details

Training Settings Our model is built on FLUX.1 Kontext [dev] Labs et al. (2025), a pretrained backbone that supports in-context image generation. We augment the backbone with the proposed

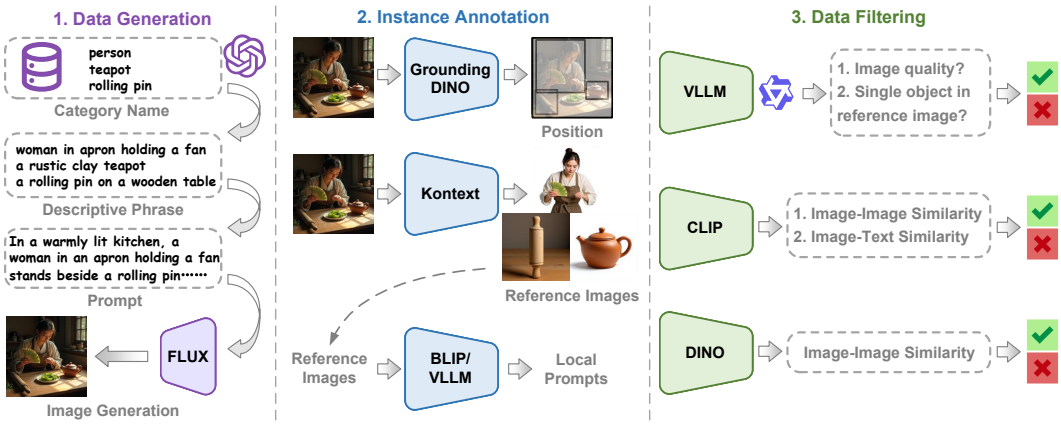


Figure 4: **Overview of the training data construction pipeline**. Image synthesis improves sample diversity, instance annotation constructs reference conditions with varied backgrounds, poses, and viewpoints, and multi-stage data filtering ensures annotation quality.

Table 1: Quantitative comparison on COCO-MIG Zhou et al. (2024). Image-guided methods (*) use pre-generated reference images. **Bold** and underline indicate the best and second-best results.

Method	SR ↑	I-SR ↑	mIoU ↑	G-C ↑	L-C ↑
GLIGEN Li et al. (2023c)	58.25	84.88	71.90	24.73	20.78
MIGC Zhou et al. (2024)	58.00	84.50	71.52	25.28	21.25
CreatiLayout Zhang et al. (2025c)	55.25	81.69	71.36	24.91	20.83
3DIS Zhou et al. (2025)	47.75	78.99	69.19	24.07	20.72
EliGen Zhang et al. (2025a)	53.13	80.09	69.99	24.95	20.70
LAMIC Chen et al. (2026)*	2.88	20.94	26.44	21.50	18.42
ContextGen Xu et al. (2026)*	53.25	81.94	**74.11**	**25.79**	21.63
MS-Diffusion Wang et al. (2025)*	10.00	41.75	43.92	25.53	20.80
MoLaGen (Ours)*	**58.63**	**85.31**	72.94	25.40	**21.81**

MoLaGen modules and fine-tune it on the dataset described in Section 4.5. We optimize the model for 200,000 iterations using AdamW with an initial learning rate of 2×10^{-5}, a weight decay of 0.01, and a cosine learning-rate schedule. The first 200 iterations use linear warm-up. Training is conducted with a global batch size of 32 across 8 NVIDIA H200 GPUs using DeepSpeed ZeRO Stage 2 and bf16 mixed precision Wang & Kanwar (2019). Target images are processed at a resolution of 1024×1024. Each training sample contains up to eight instances, and its reference images are resized to 256×256. During training, we drop global prompts with an initial probability of 0.5 to encourage reliance on instance-level conditions and gradually anneal this probability to zero. We set the main and spatial ranks of CondLoRA to $r = 512$ and $r_s = 128$, respectively. We insert CondLoRA into both the attention and MLP layers and encode Fourier-transformed spatial coordinates as 512-dimensional features.

Evaluation Settings We conduct evaluations on COCO-MIG Zhou et al. (2024) and MS-Bench Wang et al. (2025), both of which focus on multi-instance generation with spatial conditions. On COCO-MIG, we report Success Rate (SR), Instance Success Rate (I-SR), mean Intersection-over-Union (mIoU), Global CLIP Score (G-C), and Local CLIP Score (L-C). Since the original fixed HSV-based color check may reject valid colors under variations in illumination and appearance, we remove this check and compute SR, I-SR, and mIoU. MS-Bench comprises 4,488 samples formed by combining different instances. We measure image–image similarity using CLIP Radford et al. (2021) and DINO Caron et al. (2021), denoted as CLIP-I and DINO, and image–text similarity using CLIP-T. Each similarity score is computed per instance and then averaged within each image. For spatial evaluation, Grounding DINO Liu et al. (2024) detects object regions, and each prediction is

Table 2: Quantitative comparison of spatial layout accuracy, image–image similarity, and image–text similarity on MS-Bench Wang et al. (2025). Layout-controllable methods are marked with *. **Bold** and underline indicate the best and second-best results, respectively.

Method	mIoU ↑	AP ↑	AP50 ↑	AP75 ↑	DINO ↑	CLIP-T ↑	CLIP-I ↑
λ-ECLIPSE Patel et al. (2024)	–	–	–	–	0.419	0.316	0.724
SSR-Encoder Zhang et al. (2024)	–	–	–	–	0.425	0.303	**0.725**
MS-Diffusion Wang et al. (2025)*	60.21	32.99	71.80	25.36	0.425	**0.341**	0.698
ContextGen Xu et al. (2026)*	48.57	17.57	48.98	7.47	**0.481**	0.319	0.709
LAMIC Chen et al. (2026)*	38.06	10.42	33.98	3.00	0.320	0.316	0.642
MoLaGen (Ours)*	**66.09**	**46.33**	**76.18**	**46.70**	0.455	0.330	0.705

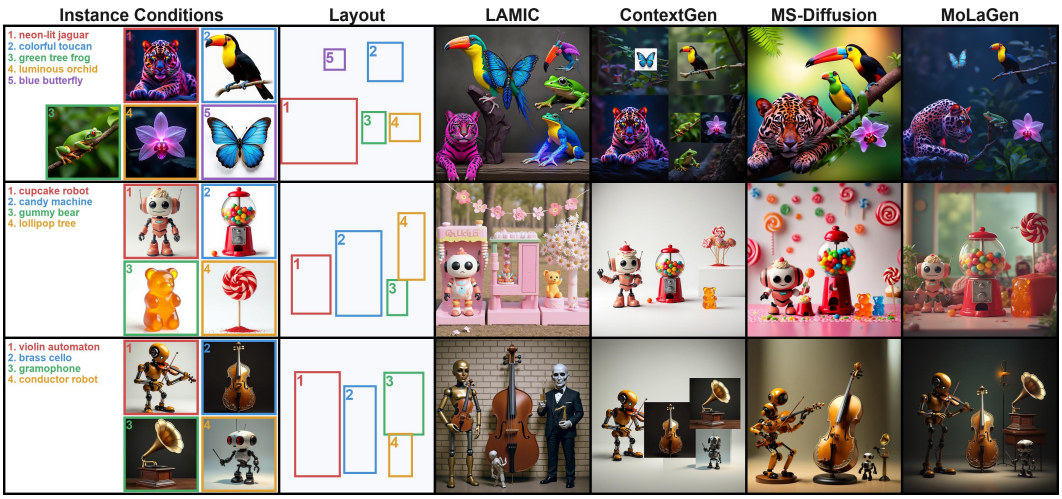


Figure 5: **Qualitative comparisons across diverse multi-instance scenes**. MoLaGen preserves condition-specific appearance and semantics while adapting each instance to the specified layout and scene context.

matched to the ground-truth box with the maximum Intersection-over-Union (IoU). We report mIoU, AP, AP50, and AP75 as the layout metrics.

Baselines For COCO-MIG, we compare with the layout-controllable baselines GLIGEN Li et al. (2023c), MIGC Zhou et al. (2024), CreatiLayout Zhang et al. (2025c), 3DIS Zhou et al. (2025), and EliGen Zhang et al. (2025a), as well as the image-guided multi-instance baselines LAMIC Chen et al. (2026), ContextGen Xu et al. (2026), and MS-Diffusion Wang et al. (2025). For MS-Bench, the layout comparison includes MS-Diffusion, ContextGen, and LAMIC, while the similarity comparison additionally includes λ-ECLIPSE Patel et al. (2024) and SSR-Encoder Zhang et al. (2024). Since the latter two methods lack explicit spatial control, their layout metrics are not reported.

5.2 Qualitative Results

Diverse Multi-Instance Scenes Figure 5 compares different methods on dense multi-instance scenes containing diverse concept combinations. MoLaGen preserves condition-specific appearance and semantics while adapting instance pose and scale to the specified regions and scene context. In contrast, competing methods more frequently omit instances, lose identifying characteristics, or produce less coherent compositions. Rather than transferring reference foregrounds directly, MoLaGen re-synthesizes each instance according to the target scene. This allows the viewpoint, pose, scale, and illumination to change while retaining identity-relevant characteristics, resulting in more consistent object boundaries and spatial relationships with the surrounding content.

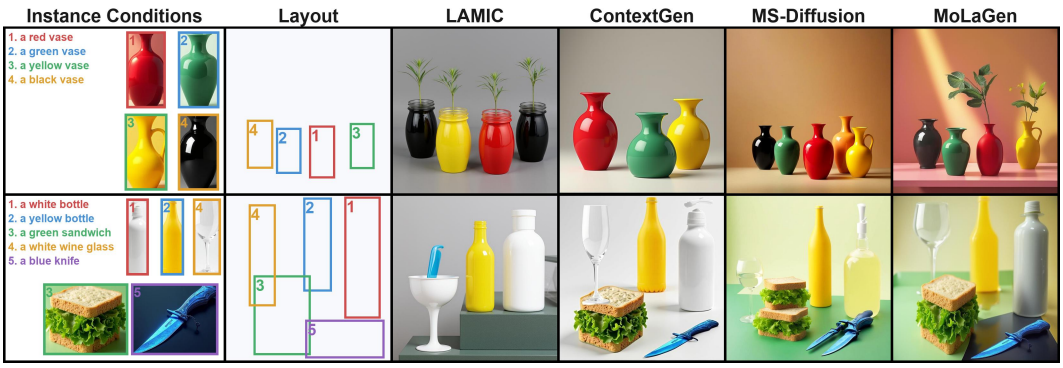


Figure 6: **Qualitative comparisons on COCO-MIG**. MoLaGen follows the specified layouts while preserving the appearance and semantics of multiple instance conditions.

COCO-MIG As shown in Figure 6, MoLaGen consistently places the specified instances within their target regions while retaining category- and appearance-specific attributes. For visually similar objects, such as the differently colored vases and bottles, competing methods more frequently omit instances, confuse their attributes, or deviate from the requested layout. MoLaGen renders these conditioned objects under shared scene lighting and perspective instead of reproducing isolated reference crops. The resulting instances exhibit more plausible relative scales and stronger visual integration, making the output appear as a unified scene rather than a composition of pasted elements. Additional qualitative comparisons on MS-Bench are provided in Appendix B.

5.3 Quantitative Results

COCO-MIG Table 1 reports the quantitative comparison on COCO-MIG. MoLaGen ranks first in SR, I-SR, and L-C and second in mIoU. The strong SR and I-SR results show that MoLaGen reliably generates the requested instances, while its mIoU confirms accurate placement within the target regions. The leading L-C score demonstrates effective preservation of instance-level semantics, and the competitive G-C score indicates a favorable balance between local control and global semantic coherence.

MS-Bench Table 2 compares spatial layout accuracy and visual and semantic similarity on MS-Bench. MoLaGen achieves the best results on all four layout metrics and ranks second in DINO and CLIP-T, demonstrating accurate instance localization together with strong visual and semantic consistency. Reference-based similarity metrics reward close replication and may penalize valid changes in pose, viewpoint, scale, or illumination. Consequently, high DINO or CLIP-I scores can accompany copy-and-paste artifacts, as observed for ContextGen in Figure 7. MoLaGen instead re-synthesizes instances to fit the target scene, providing a better balance among condition fidelity, spatial control, and scene realism.

6 Conclusion

In this work, we propose MoLaGen, a multimodal conditional layout-to-image generation framework that enables accurate instance-level control. By leveraging MM-DiT to encode both reference images and local prompts within a unified conditional space, MoLaGen achieves more efficient and semantically aligned multimodal encoding. For spatial control, we introduce a spatially-aware CondLoRA network with token-dependent routing, which adaptively activates modality-specific branches while maintaining low computational overhead. In addition, we develop a new dataset construction pipeline for synthetic data generation, instance-level annotation, and data filtering. Experiments demonstrate that MoLaGen enables efficient and precise layout control.

AI USE STATEMENT

Generative AI tools were used to polish the manuscript and assist with code development. All AI-assisted text and code were reviewed and verified by the authors. The authors take full responsibility for the final content of this work.

REFERENCES

Hongjun An, Wenhan Hu, Sida Huang, Siqi Huang, Ruanjun Li, Yuanzhi Liang, Jiawei Shao, Yiliang Song, Zihan Wang, Cheng Yuan, Chi Zhang, Hongyuan Zhang, Wenhao Zhuang, and Xuelong Li. AI Flow: Perspectives, scenarios, and approaches. *Vicinagearth*, 3(1), 2026. doi: 10.1007/s44336-025-00031-y.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025. URL https://arxiv.org/abs/2511.21631.

Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. In *International Conference on Machine Learning*, pp. 1737–1752. PMLR, 2023.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. https://cdn.openai.com/papers/dall-e-3.pdf, 2023.

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

Black Forest Labs. FLUX.2: Frontier visual intelligence. https://bfl.ai/blog/flux-2, 2025.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. SAM 3: Segment anything with concepts. In *The Fourteenth International Conference on Learning Representations*, 2026.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9650–9660, October 2021.

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

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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(4):3174–3182, 2026. doi: 10.1609/aaai.v40i4.37311.

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.

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*, 2023.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

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

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Lianghua Huang, Wei Wang, Zhi-Fan Wu, Yupeng Shi, Huanzhang Dou, Chen Liang, Yutong Feng, Yu Liu, and Jingren Zhou. In-context lora for diffusion transformers. *arXiv preprint arXiv:2410.23775*, 2024.

Sida Huang, Siqi Huang, Ping Luo, and Hongyuan Zhang. Laytrol: Preserving pretrained knowledge in layout control for multimodal diffusion transformers. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(7):5113–5121, 2026. doi: 10.1609/aaai.v40i7.37425.

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

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4015–4026, 2023.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.

Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric. In *International conference on machine learning*, pp. 1558–1566. PMLR, 2016.

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

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 12888–12900. PMLR, 17–23 Jul 2022.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 23–29 Jul 2023b.

Wei Li, Hebei Li, Yansong Peng, Siying Wu, Yueyi Zhang, and Xiaoyan Sun. Create anything anywhere: Layout-controllable personalized diffusion model for multiple subjects. *arXiv preprint arXiv:2505.20909*, 2025.

Xuelong Li. Positive-incentive noise. *IEEE Transactions on Neural Networks and Learning Systems*, 35(6):8708–8714, 2024. doi: 10.1109/TNNLS.2022.3224577.

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22511–22521, 2023c.

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. In *European Conference on Computer Vision*, pp. 38–55. Springer, 2024.

Jian Ma, Junhao Liang, Chen Chen, and Haonan Lu. Subject-diffusion: Open domain personalized text-to-image generation without test-time fine-tuning. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–12, 2024.

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*, 65(1):99–106, 2021.

Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. URL https://arxiv.org/abs/2304.07193.

Maitreya Patel, Sangmin Jung, Chitta Baral, and Yezhou Yang. λ-eclipse: Multi-concept personalized text-to-image diffusion models by leveraging clip latent space. *Transactions on Machine Learning Research*, 2024, 2024.

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

Fei Peng, Junqiang Wu, Yan Li, Tingting Gao, Di Zhang, and Huiyuan Fu. Muse: Multi-subject unified synthesis via explicit layout semantic expansion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15885–15895, 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. *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.

Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International conference on machine learning*, pp. 1060–1069. PMLR, 2016.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pp. 234–241. Springer, 2015.

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.

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. *Advances in neural information processing systems*, 35:36479–36494, 2022.

Jiawei Shao and Xuelong Li. AI Flow at the network edge. *IEEE Network*, 40(1):330–336, 2026.

Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8430–8439, 2019.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Ominicontrol: Minimal and universal control for diffusion transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14940–14950, 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Shibo Wang and Pankaj Kanwar. Bfloat16: The secret to high performance on cloud tpus. *Google Cloud Blog*, 4(1), 2019.

Xierui Wang, Siming Fu, Qihan Huang, Wanggui He, and Hao Jiang. Ms-diffusion: Multi-subject zero-shot image personalization with layout guidance. In *The Thirteenth International Conference on Learning Representations*, 2025.

Xudong Wang, Trevor Darrell, Sai Saketh Rambhatla, Rohit Girdhar, and Ishan Misra. Instancediffusion: Instance-level control for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6232–6242, 2024.

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.

Jinheng Xie, Yuexiang Li, Yawen Huang, Haozhe Liu, Wentian Zhang, Yefeng Zheng, and Mike Zheng Shou. Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7452–7461, 2023.

Ruihang Xu, Dewei Zhou, Fan Ma, and Yi Yang. Contextgen: Contextual layout anchoring for identity-consistent multi-instance generation. In *The Fourteenth International Conference on Learning Representations*, 2026.

Ling Yang, Zhaochen Yu, Chenlin Meng, Minkai Xu, Stefano Ermon, and Bin Cui. Mastering text-to-image diffusion: Recaptioning, planning, and generating with multimodal llms. In *Forty-first International Conference on Machine Learning*, 2024.

Zhengyuan Yang, Jianfeng Wang, Zhe Gan, Linjie Li, Kevin Lin, Chenfei Wu, Nan Duan, Zicheng Liu, Ce Liu, Michael Zeng, et al. Reco: Region-controlled text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14246–14255, 2023.

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.

Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 5907–5915, 2017.

Hong Zhang, Zhongjie Duan, Xingjun Wang, Yingda Chen, and Yu Zhang. Eligen: Entity-level controlled image generation with regional attention. *arXiv preprint arXiv:2501.01097*, 2025a.

Hongyuan Zhang, Sida Huang, Yubin Guo, and Xuelong Li. Variational positive-incentive noise: How noise benefits models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47 (9):8313–8320, 2025b. doi: 10.1109/TPAMI.2025.3575295.

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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18487–18497, 2025c.

Yuxuan Zhang, Yiren Song, Jiaming Liu, Rui Wang, Jinpeng Yu, Hao Tang, Huaxia Li, Xu Tang, Yao Hu, Han Pan, et al. Ssr-encoder: Encoding selective subject representation for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8069–8078, 2024.

Dewei Zhou, You Li, Fan Ma, Xiaoting Zhang, and Yi Yang. Migc: Multi-instance generation controller for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6818–6828, 2024.

Dewei Zhou, Ji Xie, Zongxin Yang, and Yi Yang. 3DIS: Depth-driven decoupled image synthesis for universal multi-instance generation. In *The Thirteenth International Conference on Learning Representations*, 2025.

A Dataset Construction Details

A.1 Motivation and Existing Construction Strategies

In multimodal conditional layout generation, each training sample provides a bounding box, local prompt, and reference image for every instance. Dataset construction must therefore ensure annotation accuracy while providing sufficient diversity across target and reference images. Existing reference-image construction methods mainly follow segmentation-based and diptych-based paradigms. Segmentation-based methods such as MUSE and ContextGen Peng et al. (2025); Xu et al. (2026) extract reference instances directly from target images using SAM variants Kirillov et al. (2023); Ravi et al. (2024); Carion et al. (2026). Consequently, the reference and target instances share nearly identical appearances, while the segmented references contain little background context, limiting the appearance and contextual variation available for learning subject generalization. MS-Diffusion Wang et al. (2025) improves reference diversity by sampling different video frames, but video collection and frame-level processing introduce additional costs. Diptych-based methods such as OminiControl and UNO Tan et al. (2025); Wu et al. (2025) prompt a generative model to produce target and reference images side by side. However, the generated diptychs may contain imprecise boundaries or inconsistent instances across the two views.

A.2 Image Generation with Diverse Subjects

The constructed dataset contains 645 fine-grained categories organized into 11 super-categories, covering human-related accessories, household items, transportation, food, electronics, animals, and other daily-life objects. Detailed statistics are provided in Table 3. Some categories are inherited from Objects365 Shao et al. (2019), while additional categories are introduced to improve diversity and coverage. Subsequently, for each object category, multiple descriptive phrases are generated using GPT. These phrases are randomly combined to form diverse text prompts, which are then fed into the FLUX Black-Forest-Labs (2024) model to synthesize corresponding images.

A.3 Entity-Level Information Annotation

Grounding DINO Liu et al. (2024) localizes each target instance and provides its bounding box. Kontext Labs et al. (2025) then generates the corresponding reference image using carefully designed prompts. As summarized in Table 4, these prompts are designed to amplify differences in background, pose, and viewpoint between the reference and original images, while preserving entity identity. Notably, the background variations are not restricted to pure white settings but also include context-appropriate environments. For human instances, pose is controlled via dedicated prompt templates, covering standing, seated, and walking configurations. Finally, BLIP Li et al. (2022; 2023b) and Qwen3-VL Bai et al. (2025) generate candidate local prompts for each reference image, and the highest-scoring candidate is retained.

A.4 Data Filtering Based on Quality and Consistency

To ensure a high-quality dataset, we progressively evaluate images and prompts across multiple criteria. First, Qwen3-VL is used to score the overall quality of images and to determine whether each reference image contains only a single entity. Next, CLIP Radford et al. (2021) evaluates image–image similarity and image–text alignment to remove semantically inconsistent pairs. Finally, DINO Caron et al. (2021); Oquab et al. (2024) measures visual consistency between each reference image and its corresponding target region. This multi-stage filtering process improves annotation reliability and overall data quality.

Table 3: Category statistics of the constructed dataset.

Super-category	No. of Categories
Human and related accessories	42
Clothes	35
Living room	86
Transportation	61
Kitchen	52
Office supplies	45
Food and vegetables	110
Instrument	68
Electronics	57
Animal	59
Bathroom	30
Total	**645**

Table 4: Candidate prompt phrases for diversifying reference images.

Type	Prompt phrase
Background	a professional studio with plain white background
	a professional studio with plain light gray background
	a context-appropriate setting background
Pose (human only)	standing frontal pose
	standing side profile pose
	seated upright pose
	walking motion pose
	sitting pose with arms crossed
	leaning pose
Viewpoint	front view, eye level
	side profile view
	overhead angle view (high angle)
	low angle view (looking up)

B Additional Qualitative Results

B.1 Qualitative Results on MS-Bench

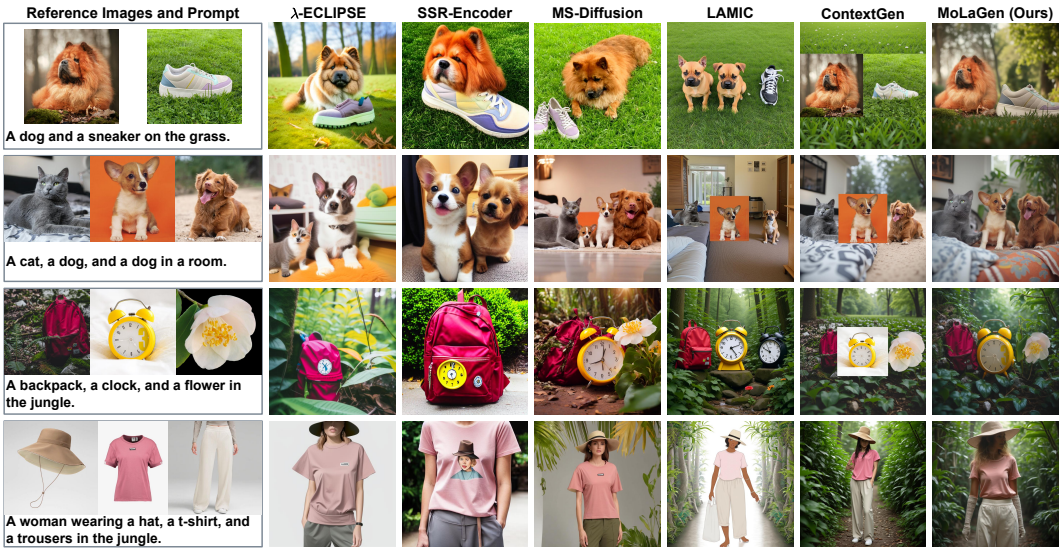


Figure 7: **Qualitative comparisons on MS-Bench**. MoLaGen preserves instance-specific appearance while integrating multiple instances naturally into coherent scenes.

Figure 7 compares MoLaGen with other baselines on MS-Bench. MoLaGen preserves instance-specific appearance while integrating each instance naturally into the generated scene and suppressing irrelevant reference backgrounds. In contrast, ContextGen exhibits visible copy-and-paste artifacts, particularly for the dogs in the first two rows and the clock in the third row, where the reference foregrounds are inserted with limited adaptation. MoLaGen produces more coherent layouts and better aligns instance-level characteristics with the global scene, while allowing the background, pose, and viewpoint to differ from the reference composition.

B.2 Qualitative Results on Diverse Multi-Instance Scenes

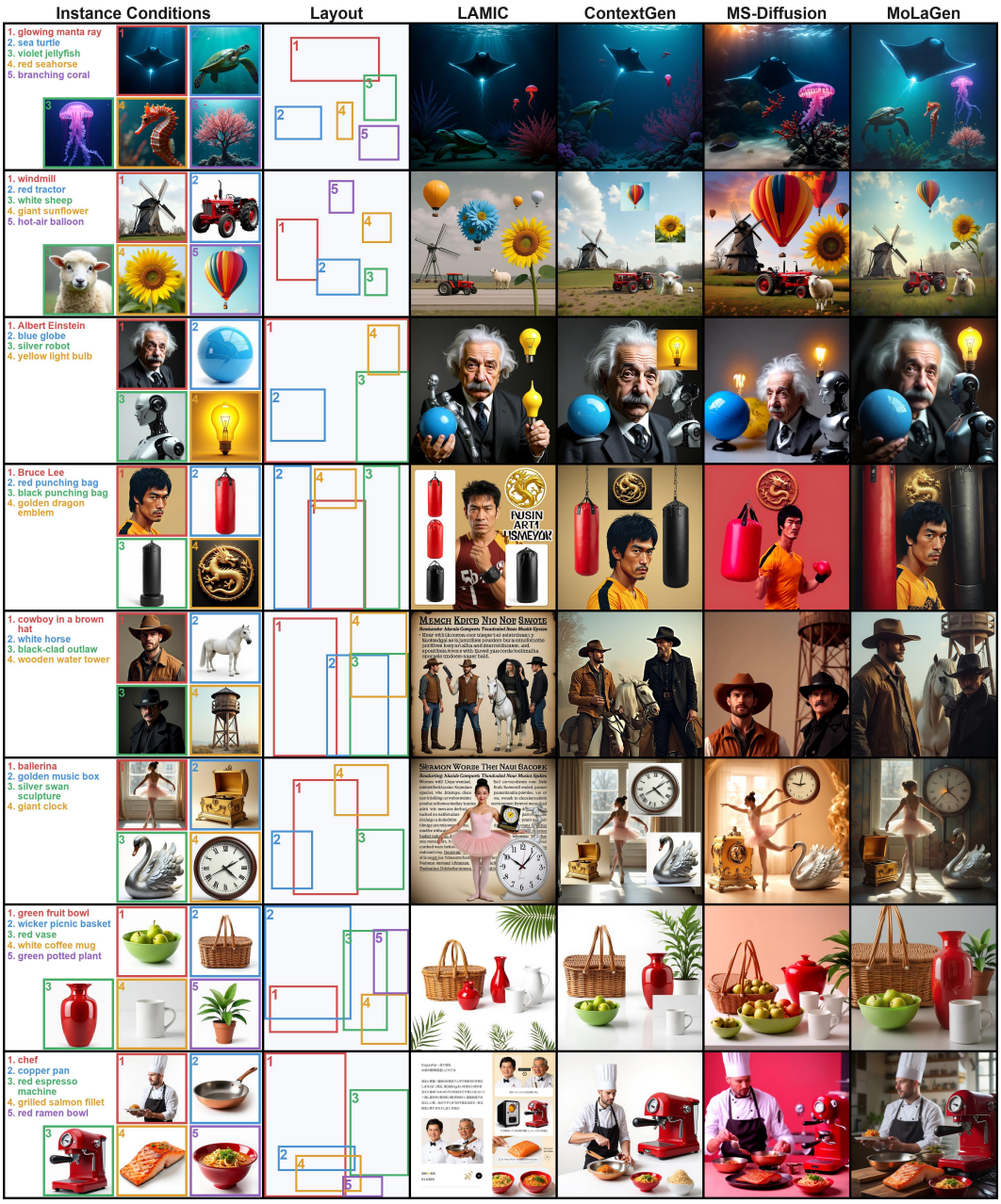


Figure 8: **Qualitative comparisons across diverse multi-instance scenes**. MoLaGen preserves condition-specific appearance and semantics while adapting instances to the specified layouts and scene contexts.

Figure 8 presents additional comparisons on dense multi-instance scenes with diverse object combinations. MoLaGen more consistently preserves condition-specific attributes while adapting instance scale and pose to the target layout and global composition.