Multimodal Large Language Models for MUlti-Subject In-Context Image Generation

Yucheng Zhou, Dubing Chen, Huan Zheng, Jianbing Shen✉
SKL-IOTSC, CIS, University of Macau
yucheng.zhou@connect.um.edu.mo, jianbingshen@um.edu.mo

Abstract

Recent advances in text-to-image (T2I) generation have enabled visually coherent image synthesis from descriptions, but generating images containing multiple given subjects remains challenging. As the number of reference identities increases, existing methods often suffer from subject missing and semantic drift. To address this problem, we propose MUSIC, the first MLLM specifically designed for **MU**lti-**S**ubject **I**n-**C**ontext image generation. To overcome the data scarcity, we introduce an automatic and scalable data generation pipeline that eliminates the need for manual annotation. Furthermore, we enhance the model's understanding of multi-subject semantic relationships through a vision chain-of-thought (CoT) mechanism, guiding step-by-step reasoning from subject images to semantics and generation. To mitigate identity entanglement and manage visual complexity, we develop a novel semantics-driven spatial layout planning method and demonstrate its test-time scalability. By incorporating complex subject images during training, we improve the model's capacity for chained reasoning. In addition, we curate MSIC, a new benchmark tailored for multi-subject in-context generation. Experimental results demonstrate that MUSIC significantly surpasses other methods in both multi- and single-subject scenarios.

1 Introduction

Recent years have witnessed significant advancements in image generation technologies, particularly in text-to-image (T2I) generation (Zhang et al., 2023a; Labs, 2024a; Esser et al., 2024; Podell et al., 2023; Xiao et al., 2024; Yu et al., 2022; Zhou et al., 2026b), where models have shown enhanced capabilities in semantic understanding and image synthesis. With the increasing demand for personalized applications, a growing body of research has focused on personalized image generation (Zhang et al., 2023a; Mou et al., 2024; Gal et al., 2022; Ruiz et al., 2023; Hu et al., 2022). These methods incorporate reference images or subject identities into the generation process to ensure alignment with the input text while preserving the visual identity of the reference subjects. This task holds practical relevance for applications such as multi-person scene synthesis and complex product visualization.

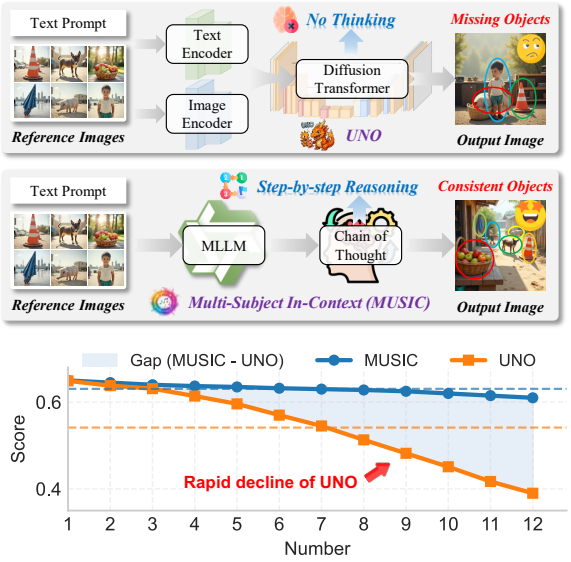

Figure 1: **Top:** Comparison of our MUSIC (bottom) *v.s.*the subject-to-image method UNO (top). **Bottom:** UNO struggles as the number of subject images grows. Our MLLM-based method uses a thinking mechanism and more effectively generates the scene with multi-subjects.

Existing approaches face substantial performance degradation as the number of reference subjects increases, as shown in Figure 1 (Right). Challenges such as subject missing and semantic drift frequently emerge, revealing the scalability limitations of current methods in complex, multi-subject generation scenarios. In contrast to diffusion mod-

✉Corresponding Author. This work was supported by the National Natural Science Foundation of China (No. 624B2002), the Science and Technology Development Fund of Macau SAR (FDCT) under grants 0102/2023/RIA2 and 0154/2022/A3, and the Jiangyin Hi-tech Industrial Development Zone under the Taihu Innovation Scheme (EF2025-00003-SKL-IOTSC).

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 32880–32898
July 2-7, 2026

els (*e.g.*, DiT (Peebles and Xie, 2023)), multimodal large language models (MLLMs) (Ge et al., 2023, 2024; Zhou et al., 2024b) exhibit superior generalization and contextual reasoning capabilities in both vision understanding and language reasoning. It has the potential to handle more complex instructions and intricate semantic relationships across multiple subjects. This raises a critical question: *"how can we effectively harness the reasoning power of MLLMs to enhance the quality and consistency of multi-subject in-context image generation?"*

To address this challenge, we propose a novel MUSIC model, the first MLLM tailored for **MU**lti-**S**ubject **I**n-**C**ontext image generation. We design an automatic data generation pipeline that eliminates the need for manual annotation. As shown in Figure 1 (Left), we introduce a vision chain-of-thought (CoT) that guides the MLLM to perform step-by-step reasoning from subject images to semantic modeling and final image generation. This approach substantially improves the model's ability to understand multi-subject semantic relationships. To handle the visual complexity and mitigate identity entanglement in multi-subject images, we propose a novel semantics-driven spatial layout planning method that assigns semantic ownership to visual content before generation, thereby reducing semantic conflicts. We also explore the test-time scalability potential of this planning method. Furthermore, by incorporating complex subject images in our model training, we enhance the model's capacity for chained reasoning across subject images, semantics, and generated images.

To better evaluate multi-subject image generation, we manually curated a new benchmark dataset, MSIC, from synthetic data, specifically designed for this task. Experimental results on the MSIC dataset and DreamBench (Ruiz et al., 2023) demonstrate that our method significantly outperforms existing approaches in semantic consistency and identity fidelity in both multi- and single-subject image generation. Our main contributions are summarized as below:

- We propose **MUSIC**, the first MLLM designed for multi-subject in-context image generation, integrating vision reasoning capabilities for multi-subject understanding and generation.
- We design an automatic and scalable data generation pipeline that requires no raw data, enabling efficient large-scale training data construction.
- We enhance the model's multi-subject understanding through a vision chain-of-thought mechanism, semantics-driven spatial layout planning, and training on complex subject images. We verify the scalability of layout planning at test time.
- We introduce the MSIC benchmark for multi-subject in-context image generation, and our method achieves state-of-the-art performance on MSIC and DreamBench.

2 Related Work

2.1 Generative Models

Diffusion Models. Diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020; Dhariwal and Nichol, 2021; Song et al., 2020; Peebles and Xie, 2023) generate high-fidelity images via a forward noising and reverse denoising process. DDPM (Ho et al., 2020) first achieved competitive image quality, and subsequent work surpassed GANs (Goodfellow et al., 2020; Brock et al., 2018; Karras et al., 2020) in sample quality (Dhariwal and Nichol, 2021). LDMs (Rombach et al., 2022) reduce computational costs by operating in latent space, while DDIM (Song et al., 2020) and consistency models (Song et al., 2023) accelerate sampling. Scalable transformer architectures such as DiT (Peebles and Xie, 2023) and U-ViT (Bao et al., 2023) further boost performance. Recent advances include architectural improvements via long-skip-connections with spectral constraints (Chen et al., 2025), hierarchical compositional generation (Yang et al., 2025c), and efficient sampling strategies (Wang et al., 2026). For controllable generation, DC-ControlNet (Yang et al., 2025a) decouples inter- and intra-element conditions, while self-rewarding LVLMs (Yang et al., 2025b) optimize T2I prompts via model-driven feedback.

Autoregressive Generative Models. Autoregressive models generate images by modeling pixel or token sequences, inspired by language modeling (Brown et al., 2020; Achiam et al., 2023; Touvron et al., 2023; Yang et al., 2024). Early models like PixelRNN (Van Den Oord et al., 2016) and PixelCNN (Van den Oord et al., 2016) predicted pixels sequentially with low efficiency. Modern approaches use vector-quantized representations such as VQ-VAE (Razavi et al., 2019) and VQ-GAN (Esser et al., 2021) to compress images into discrete tokens (Van Den Oord et al., 2017). Masked methods including MaskGIT (Chang

et al., 2022), Muse (Chang et al., 2023), and MaskBit (Weber et al., 2024) enable parallelized generation via masked token prediction (Gao et al., 2023; Fan et al., 2024). Recent models like LLaMAGen (Sun et al., 2024), Show-o (Xie et al., 2024), Infinity (Han et al., 2024), and Emu3 (Wang et al., 2024b) scale up decoder-only autoregressive frameworks (Achiam et al., 2023; Touvron et al., 2023; Yang et al., 2024). VAR (Tian et al., 2024) redefines generation as next-scale prediction, Fluid (Fan et al., 2024) integrates continuous tokenization with diffusion loss, and RandAR (Pang et al., 2024) enables parallel decoding via position prediction with KV-Cache (Pope et al., 2023). More recently, vision representation compression has improved autoregressive video generation efficiency (Zhou et al., 2026b), diffusion loss has been incorporated to refine condition errors (Zhou et al., 2026a), and entropy-guided optimization balances exploration and stability (Song et al., 2026).

2.2 In-Context Learning.

LLM In-Context Learning In-context learning (ICL) enables LLMs to adapt to new tasks via few-shot prompts without parameter updates (Brown et al., 2020). Subsequent studies explored ICL mechanisms, framing it as implicit Bayesian inference (Xie et al., 2021) and identifying attention-driven task vectors (Olsson et al., 2022). Chain-of-Thought prompting (Wei et al., 2022) significantly boosts reasoning, while retrieval-based example selection (Liu et al., 2021), context calibration (Zhao et al., 2021), and example ordering (Lu et al., 2021) further optimize ICL performance. Recent work shows that weak models can elicit strong performance through multi-capability supervision (Zhou et al., 2025).

Vision In-Context Learning. Vision in-context learning extends ICL to visual tasks using few visual examples (Tsimpoukelli et al., 2021). Flamingo (Alayrac et al., 2022) leverages interleaved image-text prompts for few-shot visual QA, while visual prompting via inpainting (Bar et al., 2022) specifies tasks through image patches. Recent work demonstrates improved compositional understanding through ICL (Nulli et al., 2024) and effective visual task adaptation in LVLMs via in-context examples (Zhou et al., 2024a). Challenges remain in efficiently representing visual prompts and generalizing with limited context (Zhang et al., 2023b; Zhou et al., 2024b).

2.3 Subject-Driven Image Generation

Subject-driven image generation synthesizes specific subjects in novel contexts. Early approaches (Zhang et al., 2023a; Mou et al., 2024; Gal et al., 2022; Ruiz et al., 2023; Hu et al., 2022) require per-subject optimization, limiting generalization. IP-Adapter (Ye et al., 2023), BLIP Diffusion (Li et al., 2023), and ELITE (Wei et al., 2023) achieve zero-shot generation via additional image encoders. Recent methods tackle multi-subject scenarios (Ma et al., 2024; Huang et al., 2025; Wang et al., 2024a; Wu et al., 2025): MS-Diffusion (Wang et al., 2024a) uses layout guidance, UNO (Wu et al., 2025) employs vision in-context learning for flexible multi-subject synthesis, and unified multimodal agents (Chen et al., 2026) combine diverse capabilities for world-grounded image synthesis.

Despite these advances, most existing approaches struggle to generalize to large-scale multi-subject combinations, and constructing diverse, high-quality training datasets for multi-subject generation remains challenging. This work addresses these gaps through a fully automated data construction pipeline and enhanced identity consistency via vision in-context learning, enabling robust multi-subject generation.

3 Methodology

To address the data scarcity for multi-subject in-context generation, we introduce an automated framework for generating diverse multi-subject datasets. We then detail the training of our MUSIC model for multi-subject in-context image generation and discuss strategies for test-time scaling using our semantics-driven spatial layout planning.

3.1 Automated Multi-Subject Dataset Generation Framework

The core of our data generation strategy is a novel, fully automated framework for synthesizing multi-subject training data. Critically, this framework operates without relying on any pre-existing image or text metadata, enabling scalable dataset creation from foundational models. Our pipeline integrates multiple foundation models, *i.e.*, large language model (LLM) (Yang et al., 2024), text-to-image (T2I) model (Labs, 2024b), vision language model (VLM) (Yang et al., 2024), image-to-image (I2I) model (Wu et al., 2025), open-vocabulary detection (OVD) model (Liu et al., 2024), and foundational

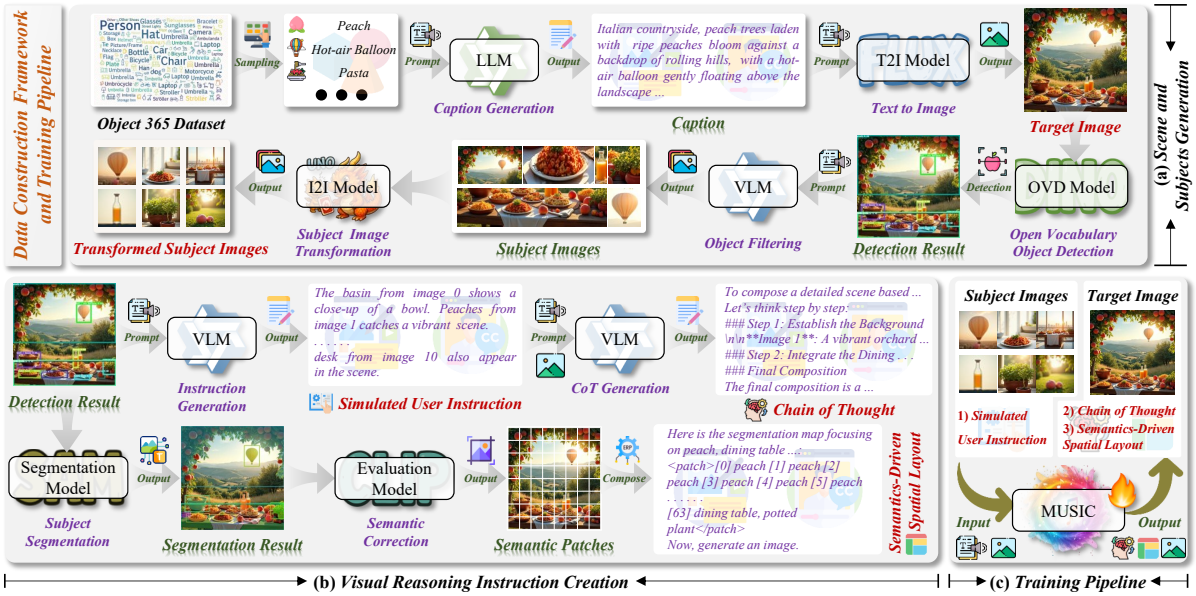

Figure 2: **Overview of our data construction pipeline: (a)** An LLM and T2I model generate the target image, followed by an OVD model detecting subjects. A VLM filters out unsuitable objects, and an I2I model creates transformed subject images. **(b)** A VLM produces simulated user instructions and CoT instructions. A segmentation model generates segmentation masks, yielding a semantics-driven spatial layout textual description. **(c)** Application of the constructed data pair in the training pipeline.

segmentation model (Ravi et al., 2024), to produce subject-consistent image pairs, instructions, and reasoning sequences for training multi-subject in-context generation models. As shown in Figure 2, the pipeline includes several stages: category sampling, caption generation, scene generation, subject detection and filtering, perspective transformation, vision chain-of-thought (CoT) reasoning, and semantics-driven spatial layout planning.

3.1.1 Scene and Subjects Generation

Subject Sampling and Scene Generation. As shown in Figure 2 (a), we start by random select between $2n_{\text{min}}$ and $2n_{\text{max}}$ categories from the Object365 dataset (Shao et al., 2019). Here, n_{min} and n_{max} define the range of subject counts. Randomly sampled categories often lack semantic coherence. For example, "lamp" and "pineapple" are unlikely to appear naturally together. To address this, we use an LLM to reason over the candidate pool and automatically select a semantically related subset. The LLM then generates a caption describing a plausible scene that combines these categories. This caption is fed into a T2I model to generate a target image I_{tgt}. This image serves as the foundation for the next stages and represents a complex multi-subject environment.

Subject Detection and Regeneration. To extract subjects in image I_{tgt}, we apply an OVD model to identify subjects in the target image and obtain their bounding boxes. To remove noisy detections, boxes with areas smaller than a threshold $\delta \cdot \text{Area}(I_{\text{tgt}})$ are filtered out. We then select up to n_{max} subjects per image, prioritizing category diversity and favoring larger bounding boxes.

Since the OVD may produce incorrect detections, we further verify bounding boxes using a VLM. Specifically, we overlay the detected boxes, categories, and confidence scores on the original image and prompt the VLM to assess their correctness. Incorrect boxes are filtered out. Each valid subject is cropped from the target image to create a subject-specific subimage. To avoid the subject-to-image generation model learning trivial copy-paste strategies, we apply an I2I model to produce view-transformed subject images. These transformed images $I_{\text{subj}_1}, I_{\text{subj}_2}, \ldots, I_{\text{subj}_S}$ and the target image I_{tgt} form the paired training data for our model, where S is the number of detected subjects.

3.1.2 Visual Reasoning Instruction Creation

Instruction and Vision CoT Generation. In real scenarios, users typically provide a simple instruction along with a set of subject images to guide scene composition. However, as the number of subjects increases, maintaining subject identity and correct spatial arrangements becomes very challenging. To solve this, we generate CoT data to improve the model's reasoning abilities.

As shown in Figure 2 (b), we use a VLM to generate a simulated user instruction based on the target image annotated with bounding boxes and

category labels. Each detected subject is assigned an ID to clarify object references. The instruction, along with the annotated target image, is then used to prompt another VLM to produce a detailed CoT reasoning sequence. The simulated user instruction may include subject IDs or omit them; we generate both types to cover a wide range of scenarios. Subjects mentioned in the CoT are expected to be linked to their corresponding IDs. This helps the model understand subject interactions within the scene, especially when multiple instances of the same category appear. We also perform string matching to ensure that IDs in the CoT and instruction align with detected subjects, removing any incorrect IDs introduced by hallucination. The CoT outlines a clear, step-by-step reasoning process for scene construction, including spatial arrangements and contextual relationships between subjects.

Semantics-Driven Spatial Layout Planning. To further enhance subject placement, we propose a semantics-driven spatial layout planning approach. Following vision CoT reasoning (Shao et al., 2024), this method aligns visual content more precisely within the 2D spatial domain of the image. Specifically, we incorporate category-level semantic information into designated spatial regions. Inspired by Text4Seg (Lan et al., 2024), we partition the target image into an $M \times M$ grid ($M = 8$ in our implementation) and determine the dominant subject categories for each patch.

To assign categories accurately, we use the foundational segmentation model SAM2 (Ravi et al., 2024) to generate object masks from detected bounding boxes. However, as SAM2 relies solely on spatial inputs and lacks language understanding, its masks may miss the correct regions. To improve robustness, we dynamically choose between using the mask or the bounding box for each subject. Specifically, we employ CLIP (Radford et al., 2021) as the evaluation model to compute cosine similarities between the class text embedding and the average visual features extracted from both the masked and unmasked regions:

$$\begin{aligned} \text{Sim}_{\text{mask}} &= \cos\left(f_{\text{mask}}, f_{\text{class}}\right), \\ \text{Sim}_{\text{unmask}} &= \cos\left(f_{\text{unmask}}, f_{\text{class}}\right). \end{aligned} \tag{1}$$

If $\text{Sim}_{\text{mask}} > \text{Sim}_{\text{unmask}}$, the mask is used, otherwise, we fall back to the bounding box. This approach balances SAM2's fine-grained segmentation with improved robustness.

We compute the intersection-over-union (IoU) between each patch and the subject's mask or bounding box. To avoid filtering out small objects, we use a dynamic IoU threshold:

$$\tau = \lambda \cdot \frac{1}{K} \sum_{i=1}^{K} \text{IoU}(b_i, p_i), \tag{2}$$

where λ is the pre-defined scaling factor, K is the number of non-zero IoUs, and b_i and p_i represent the subject region and the patch, respectively. Patches with IoU larger than τ are assigned the subject's category; otherwise, they default to "others." The final spatial layout prompt is structured as follows:

```
Here is the segmentation map focusing
on flower, frame, pineapple, plate,
potted plant, lamp, couch: <patch> [0]
others [1] others [2] frame [3] others
... [63] others </patch>
Now, generate an image.
```

This segmentation-aware instruction is integrated into the CoT to guide the subject-to-image generation model in accurately composing multi-subject scenes with correct spatial layouts.

3.1.3 Final Training Data Construction

We further enriched our dataset by simulating more realistic user demands, such as generating scenes centered around a specific subject within a complex environment. To achieve this, we used an LLM to construct a category dictionary, where each key represents a class and the corresponding value contains semantically similar classes. When generating transformed subject images with the I2I model, we randomly select multiple similar classes from this dictionary and incorporate them into the instruction. This encourages the I2I model to produce more diverse and complex scenes. As a result, our dataset contains both simple and complex instances, providing a comprehensive training resource. More details can be found in Appendix C.

We further augment the generated data by systematically reducing the number of subjects. We sort subjects by bounding box size and iteratively remove the smallest ones until only two subjects remain. Correspondingly, we update the instruction by removing or reordering subject IDs to match the new subject set. Through this process, we generate data with varying subject counts at no extra cost, helping the model better handle scenes with different levels of complexity.

As shown in Figure 2 (c), each final training instance includes a set of augmented subject images $\{I_{\text{subj}_1}, \ldots, I_{\text{subj}_S}\}$, the target image I_{tgt}, a concise simulated user instruction T_{instr}, a detailed CoT reasoning sequence C_{CoT}, and a semantics-driven spatial layout planning description L_{spatial}. This rich and structured supervision significantly improves the model's ability to reason, compose, and maintain both subject identity and spatial consistency in complex multi-subject scenarios.

3.2 Training MUSIC for Multi-Subject In-Context Image Generation

We train our proposed model, MUSIC (Multi-Subject In-Context Image Generation), using the diverse and structured datasets automatically generated by the framework described in Section 3.1. The architecture of MUSIC is built upon the principles of existing MLLM frameworks, *e.g.*, SEED-X (Ge et al., 2024), which are adept at processing and generating both textual and visual data. Our training strategy decomposes the complex task of multi-subject in-context generation into two distinct capabilities, mirroring the structure of our generated data and the desired inference process:

Capability 1: Visual Reasoning and Spatial Planning. This phase focuses on teaching the model to interpret user intent and plan the scene composition. The input consists of the initial user instruction (text) and the set of augmented subject images $\{I_{\text{subj}_1}, \ldots, I_{\text{subj}_S}\}$. The training objective is to predict the corresponding vision CoT reasoning sequence and the semantics-driven spatial layout plan generated by our pipeline.

Formally, given the input instruction T_{instr} and subject images $\{I_{\text{subj}_i}\}_{i=1}^{S}$, this capability learns a mapping f_1:

$$f_1(T_{\text{instr}}, \{I_{\text{subj}_i}\}_{i=1}^{S}) \rightarrow (\hat{C}_{\text{CoT}}, \hat{L}_{\text{spatial}}) \quad (3)$$

where $\hat{C}_{\text{CoT}}$ is the predicted CoT text sequence and $\hat{L}_{\text{spatial}}$ is the predicted spatial layout grid. This capability is trained using supervision from the ground-truth C_{CoT} and L_{spatial}, with a cross-entropy loss. The loss for this stage is denoted as L_1.

Capability 2: Image Generation from Plan. This phase focuses on synthesizing the final image conditioned on the outputs of the planning stage. The input to this capability consists of the vision CoT and the semantics-driven spatial layout plan. The training objective is to generate the target image I_{tgt} that reflects the intended composition and semantics. This capability learns a mapping f_2:

$$f_2(C_{\text{CoT}}, L_{\text{spatial}}) \rightarrow \hat{I}_{\text{tgt}} \quad (4)$$

where $\hat{I}_{\text{tgt}}$ denotes the synthesized image. During training, we leverage the ground-truth C_{CoT} and L_{spatial} as conditioning signals. Following the SEED-X framework, this stage is optimized by minimizing the distance between the visual representations of the synthesized image $\hat{I}_{\text{tgt}}$ and the ground-truth image I_{tgt}. The corresponding training loss is denoted as L_2.

Overall Training Objective. The MUSIC model is trained by optimizing a combination of the losses from both capabilities. The total loss L can be expressed as:

$$\begin{aligned} L = {} & w_1 L_1(T_{\text{instr}}, \{I_{\text{subj}_i}\}_{i=1}^{S}; C_{\text{CoT}}, L_{\text{spatial}}) \\ & + w_2 L_2(C_{\text{CoT}}, L_{\text{spatial}}; I_{\text{tgt}}) \end{aligned} \quad (5)$$

where w_1 and w_2 are weighting factors. Training is performed end-to-end, allowing the gradients from L_2 to flow back by the predicted CoT and layout.

3.3 Test Time Scaling via Semantics-Driven Spatial Layout Planning

At inference time, we enable scalable and diverse generation by producing multiple candidate spatial layout plans from a single instruction and subject set. MUSIC first generates N candidate planning branches, each comprising a unique CoT and spatial layout:

$$\{(\hat{C}_{\text{CoT}}^{(j)}, \hat{L}_{\text{spatial}}^{(j)})\}_{j=1}^{N} \quad (6)$$

Each plan leads to a synthesized image via the generation module:

$$\hat{I}_{\text{tgt}}^{(j)} = f_2(\hat{C}_{\text{CoT}}^{(j)}, \hat{L}_{\text{spatial}}^{(j)}) \quad (7)$$

To select the best output, we use a CLIP-based verifier to compute similarity scores between each generated image and the original instruction T_{instr}:

$$\begin{aligned} I_{\text{final}} &= \hat{I}_{\text{tgt}}^{(j^*)}, \\ j^* &= \arg\max_j \cos\left(f_{\text{CLIP}}^{\text{image}}(\hat{I}_{\text{tgt}}^{(j)}), f_{\text{CLIP}}^{\text{text}}(T_{\text{instr}})\right) \end{aligned} \quad (8)$$

This strategy introduces minimal overhead while significantly enhancing output diversity and fidelity. It is adaptable to different generation frameworks and supports flexible test-time control over layout complexity.

4 Experiments

4.1 Experiments Setting

Implementation Details. We adopt Qwen-3 (Yang et al., 2024) as our LLM, Qwen-2.5 VL (Yang et al., 2024) as our VLM, FLUX-1.0-DEV (Labs, 2024a) as the T2I model, UNO-FLUX-1.0-DEV (Wu et al., 2025) for I2I generation, GroundingDINO as our OVD model, SAM2 for segmentation, and CLIP-vit-large-patch14 (Radford et al., 2021) for mask evaluation. We set the maximum number of subjects n_{max} to 12, the minimum number n_{min} to 1, the threshould scaling factor δ to 0.01, the patch count to 8×8, and the scaling factor λ to 0.05. We initialize our model using SEED-X and apply Low-Rank Adaptation (LoRA) for efficient fine-tuning. The training is conducted on a synthetic dataset comprising 10,000 samples generated through our customized data generation pipeline. All experiments are performed using a server equipped with 8 × A100 GPUs. LoRA is configured with a rank of 64 and a scaling factor of $\alpha = 64$. We train the model for 10 epochs using a learning rate of 1×10^{-4}. The LoRA weight scaling parameters are fixed at $w_1 = 0.5$ and $w_2 = 0.5$ throughout the training process.

Evaluation Benchmarks and Metrics. We evaluate our proposed MUSIC framework on two benchmark datasets targeting different aspects of in-context image generation. For multi-subject evaluation, we introduce the MSIC dataset, designed to assess model scalability across varying subject counts (1 to 12). For single-subject generation, we adopt DreamBench (Ruiz et al., 2023). Performance is quantified using three automatic metrics: DINO (Oquab et al., 2023), CLIP-I, and CLIP-T (Radford et al., 2021). DINO measures image-level fidelity using self-supervised features. CLIP-I evaluates subject fidelity by comparing the generated image to the reference. CLIP-T measures alignment with the input text prompt. Higher scores indicate better quality for all metrics.

Comparative Methods. To comprehensively assess the effectiveness of MUSIC, we compare it against a range of state-of-the-art baselines across both multi-subject and single-subject in-context image generation tasks. For the multi-subject setting, we include Subject Diffusion (Ma et al., 2024), MIP-Adapter (Huang et al., 2025), MS-Diffusion (Wang et al., 2024a), OmniGen (Xiao et al., 2024), and UNO (Wu et al., 2025) as comparison baselines. For single-subject setting, we include Textual Inversion (Gal et al., 2022), DreamBooth (Ruiz et al., 2023), BLIP-Diffusion (Li et al., 2023), ELITE (Wei et al., 2023), Re-Imagen (Chen et al., 2022), BootPIG (Purushwalkam et al., 2024), SSR-Encoder (Zhang et al., 2024), RealCustom++ (Mao et al., 2024), OmniGen (Xiao et al., 2024), OminiControl (Tan et al., 2024), FLUX.1 IP-Adapter, and UNO (Wu et al., 2025).

Method	DINO ↑	CLIP-I ↑	CLIP-T ↑
Oracle (caption)	-	-	0.339
Subject Diffusion (Ma et al., 2024)	0.513	0.702	0.287
MIP-Adapter (Huang et al., 2025)	0.497	0.715	0.288
MS-Diffusion (Wang et al., 2024a)	0.532	0.714	0.290
OmniGen (Xiao et al., 2024)	0.525	0.714	0.300
UNO (Wu et al., 2025)	0.541	0.721	0.296
MUSIC (Ours)	0.622	0.812	0.322
MUSIC* (Ours)	**0.631**	**0.822**	**0.330**

Table 1: Performance comparison on the MSIC (multi-subject in-context generation). MUSIC* refers to our method with test-time scaling. Oracle uses ground-truth captions.

Method	DINO ↑	CLIP-I ↑	CLIP-T ↑
Oracle (reference images)	0.774	0.885	-
Textual Inversion (Gal et al., 2022)	0.569	0.780	0.255
DreamBooth (Ruiz et al., 2023)	0.668	0.803	0.305
BLIP-Diffusion (Li et al., 2023)	0.670	0.805	0.302
ELITE (Wei et al., 2023)	0.647	0.772	0.296
Re-Imagen (Chen et al., 2022)	0.600	0.740	0.270
BootPIG (Purushwalkam et al., 2024)	0.674	0.797	0.311
SSR-Encoder (Zhang et al., 2024)	0.612	0.821	0.308
RealCustom++ (Mao et al., 2024)	0.702	0.794	0.318
OmniGen (Xiao et al., 2024)	0.693	0.801	0.315
OminiControl (Tan et al., 2024)	0.684	0.799	0.312
FLUX.1 IP-Adapter (team, 2025)	0.582	0.820	0.288
UNO (Wu et al., 2025)	0.760	0.835	0.304
MUSIC (Ours)	0.761	0.837	0.317
MUSIC* (Ours)	**0.768**	**0.840**	**0.321**

Table 2: Performance comparison of different methods on Dreambench (single-subject in-context generation).

4.2 Quantitative Evaluations

Automatic Scores. Table 1 presents results on the MSIC benchmark for multi-subject in-context generation. **MUSIC** outperforms all state-of-the-art baselines across three metrics (DINO 0.622, CLIP-I 0.812, CLIP-T 0.322), demonstrating superior identity preservation and semantic alignment in complex multi-subject prompts. The **MUSIC*** variant, with semantics-driven spatial planning for test-time scaling, achieves the best overall performance. Table 2 shows results on DreamBench (Ruiz et al., 2023) for single-subject generation. Although designed for multi-subject synthesis, **MUSIC** attains competitive results (DINO 0.761, CLIP-I 0.837) comparable to UNO (Wu et al., 2025) (0.760, 0.835), while **MUSIC*** slightly improves results (0.768, 0.840), confirming the benefit of test-time scaling.

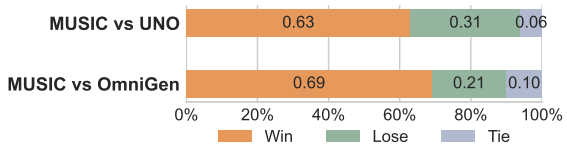

Figure 3: Human evaluation results comparing the MUSIC against OmniGen and UNO.

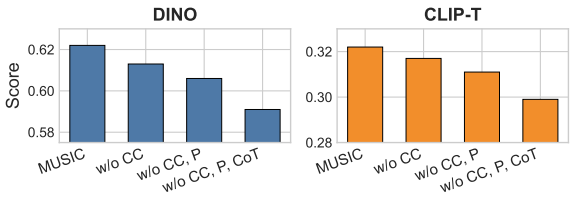

Figure 4: Ablation study. "w/o CC" (Complex Case data augmentation); "w/o P" (Complex Case and spatial layout planning); "w/o CoT" (Complex Case, spatial planning, and Chain-of-Thought reasoning).

Human Evaluation. To complement our automatic quantitative results, we conducted a human evaluation study comparing MUSIC against two strong baselines: OmniGen (Xiao et al., 2024) and UNO (Wu et al., 2025). For a subset of MSIC prompts, human raters compared paired images (MUSIC vs. baseline), evaluating overall quality, subject identity, and adherence to prompt instructions. Figure 3 presents the results. Against OmniGen, MUSIC was preferred in 69% of cases vs. 21% for OmniGen (10% Tie). Against UNO, MUSIC was preferred in 63% vs. 31% for UNO (6% Tie). These results demonstrate a strong human preference for MUSIC's outputs, aligning with our automatic metrics.

4.3 Ablation Study

As shown in Figure 4, an ablation study investigated the contribution of Complex Case augmentation (CC), spatial layout planning (P), and Vision Chain-of-Thought (CoT). Removing CC ("w/o CC") caused a performance drop, suggesting its benefit for varying subject counts. A more significant decline, especially in CLIP-I and CLIP-T, occurred when P was also removed ("w/o CC, P"), highlighting its importance for accurate composition. The largest degradation was observed upon removing all three ("w/o CC, P, CoT"), underscoring CoT's critical role in guiding complex interactions and ensuring fidelity. The study confirms all components are beneficial, with P and CoT particularly impactful for reasoning in multi-subject generation.

Effect of Subject Number. To assess MUSIC's performance under varying subject counts, we evaluated it on the MSIC with 1 to 12 subjects. As expected due to the inherent complexity of multi-subject scenes (maintaining identity, spatial relations, preventing entanglement), Figure 5 shows that DINO and CLIP-T scores decrease with the number of subjects. However, MUSIC maintains relatively strong performance even at high subject counts, demonstrating the robustness of our approach to multi-subject complexity.

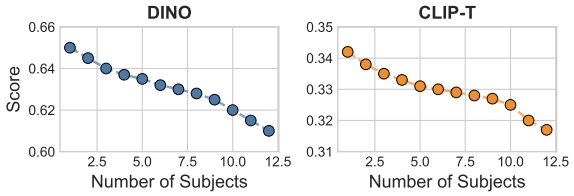

Figure 5: Performance of MUSIC as a function of the number of subjects on the MSIC.

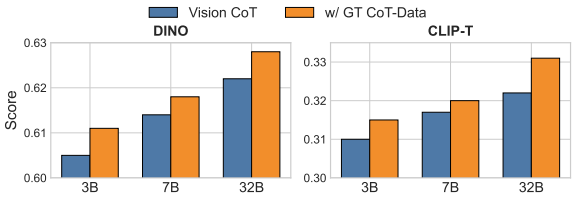

Figure 6: Performance of MUSIC trained on Vision CoT data from Qwen2.5-VL (3B, 7B and 32B). "w/ GT CoT-Data" denotes inference with ground-truth CoT from Qwen2.5-VL, which further boosts generation quality.

Effect of Vision CoT. Vision Chain-of-Thought (CoT) reasoning is crucial for complex multi-subject scene understanding. We evaluated its impact using CoT data from varying Qwen2.5-VL model sizes (3B, 7B, 32B), reflecting VLM reasoning quality. Figure 6 shows that training with CoT from larger VLMs (32B) improves performance (DINO, CLIP-T scores) (left plots). Providing "ground-truth" CoT (from the 32B VLM) at inference significantly boosts performance across all models, regardless of training CoT quality (right plots). This highlights Vision CoT's vitality, the impact of training data quality, and the significant benefit of high-quality CoT during inference for complex multi-subject generation.

4.4 Semantics-Driven Spatial Layout Planning for Test-Time Scaling

Our semantics-driven spatial layout planning mechanism enables effective test-time scaling to enhance performance (Section 3.3). This involves generating N diverse candidate spatial plans and Vision CoT sequences, selecting the best image via CLIP similarity (Pass@N). Figure 8 demonstrates this effect: DINO and CLIP-T scores consistently in-

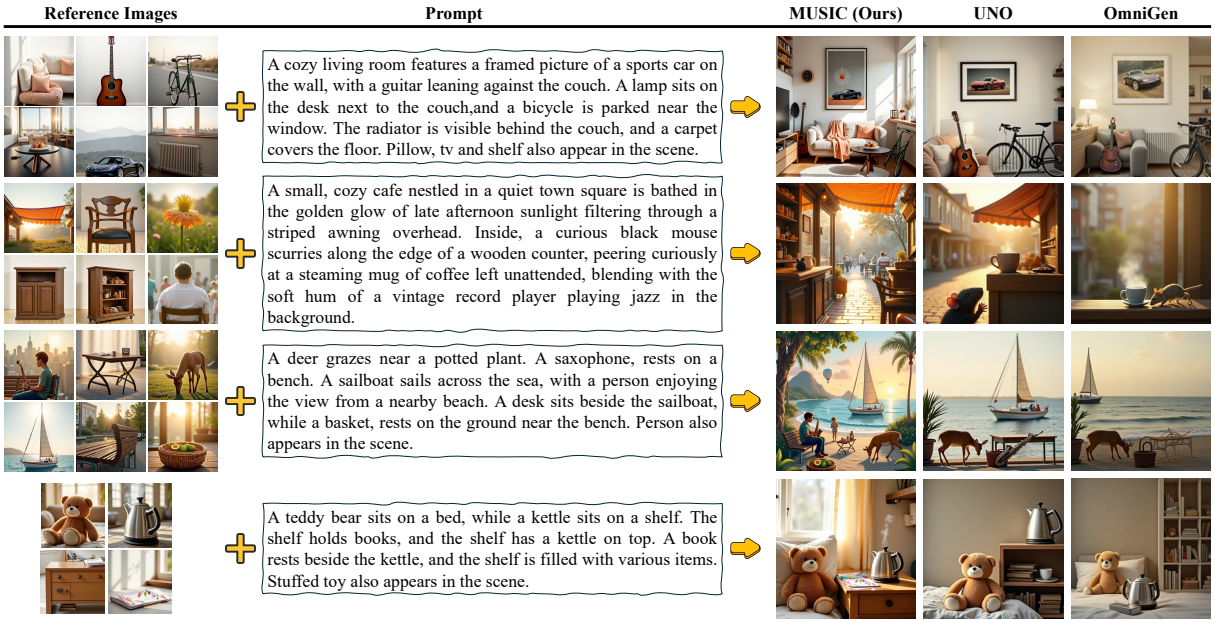

Figure 7: Qualitative comparison of multi-subject image generation results from **MUSIC (Ours)** against **UNO** and **OmniGen**. Each row shows the reference images, input prompt, and generated images.

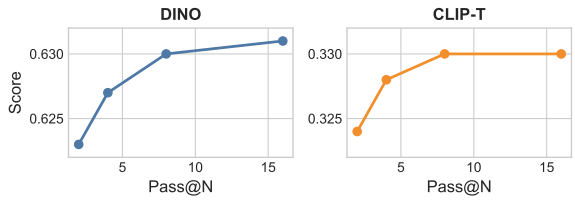

Figure 8: Effectiveness of test-time scaling using Semantics-Driven Spatial Layout Planning.

crease as Pass@N rises from 2 to 16 (*e.g.*, DINO: 0.623 → 0.631, CLIP-T: 0.324 → 0.330). This confirms that generating more planning candidates and selecting the best improves image fidelity and text-alignment, providing a valuable mechanism for trading computation for quality at test time.

4.5 Qualitative Analyses

Figure 7 shows qualitative comparisons between **MUSIC**, UNO (Wu et al., 2025), and OmniGen (Xiao et al., 2024) on multi-subject in-context generation. Each example includes reference subjects, a text prompt, and the corresponding generated images. **MUSIC** synthesizes complex scenes more coherently, accurately preserving subject identities and spatial relationships (e.g., objects in the living room or multiple interacting subjects on the beach). While baselines produce reasonable results, **MUSIC** achieves higher fidelity, more natural layouts, and better semantic consistency, consistent with quantitative and human evaluations.

5 Conclusion

Generating images with multiple specific subjects remains challenging, often leading to missing subjects and semantic drift. We present MUSIC, the first MLLM for **MU**lti-**S**ubject **I**n-**C**ontext image generation. To address data scarcity, we design an automatic scalable data pipeline and enhance multi-subject reasoning through a vision chain-of-thought and semantics-driven spatial layout planning, enabling effective test-time scaling. Training on complex subject compositions further strengthens reasoning ability. We also introduce MSIC, a new benchmark for multi-subject in-context generation. Experiments show that MUSIC substantially surpasses other methods in identity fidelity, semantic consistency, and overall image quality.

Limitations

Although MUSIC significantly outperforms baselines in handling multiple subjects and shows better scaling properties, the performance, as demonstrated in our experiments, still exhibits a degradation trend as the number of subjects increases. Moreover, our test-time scaling via Pass@N improves quality at the cost of linearly increasing inference time, limiting applicability in low-latency scenarios. Future work could explore more efficient plan generation or alternative selection mechanisms.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, and 1 others. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Fan Bao, Shen Nie, Kaiwen Xue, Yue Cao, Chongxuan Li, Hang Su, and Jun Zhu. 2023. All are worth words: A vit backbone for diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22669–22679.

Amir Bar, Yossi Gandelsman, Trevor Darrell, Amir Globerson, and Alexei Efros. 2022. Visual prompting via image inpainting. *Advances in Neural Information Processing Systems*, 35:25005–25017.

Andrew Brock, Jeff Donahue, and Karen Simonyan. 2018. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, and 1 others. 2023. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. 2022. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11315–11325.

Guanjie Chen, Xinyu Zhao, Yucheng Zhou, Xiaoye Qu, Tianlong Chen, and Yu Cheng. 2025. Towards stabilized and efficient diffusion transformers through long-skip-connections with spectral constraints. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17708–17718.

Shuang Chen, Quanxin Shou, Hangting Chen, Yucheng Zhou, Kaituo Feng, Wenbo Hu, Yi-Fan Zhang, Yunlong Lin, Wenxuan Huang, Mingyang Song, and 1 others. 2026. Unify-agent: A unified multimodal agent for world-grounded image synthesis. *arXiv preprint arXiv:2603.29620*.

Wenhu Chen, Hexiang Hu, Chitwan Saharia, and William W Cohen. 2022. Re-imagen: Retrieval-augmented text-to-image generator. *arXiv preprint arXiv:2209.14491*.

Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, and 1 others. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*.

Patrick Esser, Robin Rombach, and Bjorn Ommer. 2021. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883.

Lijie Fan, Tianhong Li, Siyang Qin, Yuanzhen Li, Chen Sun, Michael Rubinstein, Deqing Sun, Kaiming He, and Yonglong Tian. 2024. Fluid: Scaling autoregressive text-to-image generative models with continuous tokens. *arXiv preprint arXiv:2410.13863*.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. 2022. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*.

Shanghua Gao, Pan Zhou, Ming-Ming Cheng, and Shuicheng Yan. 2023. Masked diffusion transformer is a strong image synthesizer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 23164–23173.

Yuying Ge, Sijie Zhao, Ziyun Zeng, Yixiao Ge, Chen Li, Xintao Wang, and Ying Shan. 2023. Making llama see and draw with seed tokenizer. *arXiv preprint arXiv:2310.01218*.

Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying Shan. 2024. Seed-x: Multimodal models with unified multi-granularity comprehension and generation. *arXiv preprint arXiv:2404.14396*.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144.

Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. 2024. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. *arXiv preprint arXiv:2412.04431*.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *NIPS*, 33:6840–6851.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

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

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119.

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

Black Forest Labs. 2024b. Flux: Official inference repository for flux.1 models. Accessed: 2025-02-07.

Mengcheng Lan, Chaofeng Chen, Yue Zhou, Jiaxing Xu, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. 2024. Text4seg: Reimagining image segmentation as text generation. *arXiv preprint arXiv:2410.09855*.

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

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021. What makes good in-context examples for gpt-3? *arXiv preprint arXiv:2101.06804*.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, and 1 others. 2024. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pages 38–55. Springer.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2021. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. *arXiv preprint arXiv:2104.08786*.

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

Zhendong Mao, Mengqi Huang, Fei Ding, Mingcong Liu, Qian He, and Yongdong Zhang. 2024. Realcustom++: Representing images as real-word for real-time customization. *arXiv preprint arXiv:2408.09744*.

Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. 2024. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 4296–4304.

Matteo Nulli, Anesa Ibrahimi, Avik Pal, Hoshe Lee, and Ivona Najdenkoska. 2024. In-context learning improves compositional understanding of vision-language models. *arXiv preprint arXiv:2407.15487*.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, and 1 others. 2022. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*.

Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, and 7 others. 2023. Dinov2: Learning robust visual features without supervision.

Ziqi Pang, Tianyuan Zhang, Fujun Luan, Yunze Man, Hao Tan, Kai Zhang, William T Freeman, and Yu-Xiong Wang. 2024. Randar: Decoder-only autoregressive visual generation in random orders. *arXiv preprint arXiv:2412.01827*.

William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. In *ICCV*, pages 4195–4205.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

Reiner Pope, Sholto Douglas, Aakanksha Chowdhery, Jacob Devlin, James Bradbury, Jonathan Heek, Kefan Xiao, Shivani Agrawal, and Jeff Dean. 2023. Efficiently scaling transformer inference. *Proceedings of Machine Learning and Systems*, 5:606–624.

Senthil Purushwalkam, Akash Gokul, Shafiq Joty, and Nikhil Naik. 2024. Bootpig: Bootstrapping zero-shot personalized image generation capabilities in pretrained diffusion models. *arXiv preprint arXiv:2401.13974*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PmLR.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, and 1 others. 2024. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*.

Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. 2019. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, pages 22500–22510.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. 2024. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *Advances in Neural Information Processing Systems*, 37:8612–8642.

Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. 2019. Objects365: A large-scale, high-quality dataset for object detection. In *CVPR*, pages 8430–8439.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, pages 2256–2265. pmlr.

Han Song, Yucheng Zhou, Jianbing Shen, and Yu Cheng. 2026. From broad exploration to stable synthesis: Entropy-guided optimization for autoregressive image generation. In *The Fourteenth International Conference on Learning Representations*.

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

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. 2023. Consistency models.

Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. 2024. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*.

Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. 2024. Ominicontrol: Minimal and universal control for diffusion transformer. *arXiv preprint arXiv:2411.15098*, 3.

XLabs AI team. 2025. x-flux. Accessed: 2025-02-07.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. 2024. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Advances in neural information processing systems*, 37:84839–84865.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212.

Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, and 1 others. 2016. Conditional image generation with pixelcnn decoders. *Advances in neural information processing systems*, 29.

Aäron Van Den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. 2016. Pixel recurrent neural networks. In *International conference on machine learning*, pages 1747–1756. PMLR.

Aaron Van Den Oord, Oriol Vinyals, and 1 others. 2017. Neural discrete representation learning. *Advances in neural information processing systems*, 30.

Chenglin Wang, Yucheng Zhou, Shawn Chen, Tao Wang, and Kai Zhang. 2026. Ladr: Locality-aware dynamic rescue for efficient text-to-image generation with diffusion large language models. *arXiv preprint arXiv:2603.13450*.

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

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, and 1 others. 2024b. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*.

Mark Weber, Lijun Yu, Qihang Yu, Xueqing Deng, Xiaohui Shen, Daniel Cremers, and Liang-Chieh Chen. 2024. Maskbit: Embedding-free image generation via bit tokens. *arXiv preprint arXiv:2409.16211*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. 2023. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In *CVPR*, pages 15943–15953.

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

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

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. 2024. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2021. An explanation of in-context learning as implicit bayesian inference. *arXiv preprint arXiv:2111.02080*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Hongji Yang, Wencheng Han, Yucheng Zhou, and Jianbing Shen. 2025a. Dc-controlnet: Decoupling inter-and intra-element conditions in image generation with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19065–19074.

Hongji Yang, Yucheng Zhou, Wencheng Han, and Jianbing Shen. 2025b. Self-rewarding large vision-language models for optimizing prompts in text-to-image generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7332–7349.

Hongji Yang, Yucheng Zhou, Wencheng Han, Runzhou Tao, Zhongying Qiu, Jianfei Yang, and Jianbing Shen. 2025c. Hicogen: Hierarchical compositional text-to-image generation in diffusion models via reinforcement learning. *arXiv preprint arXiv:2511.19965*.

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

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, and 1 others. 2022. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023a. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847.

Yuanhan Zhang, Kaiyang Zhou, and Ziwei Liu. 2023b. What makes good examples for visual in-context learning? *Advances in Neural Information Processing Systems*, 36:17773–17794.

Yuxuan Zhang, Yiren Song, Jiaming Liu, Rui Wang, Jinpeng Yu, Hao Tang, Huaxia Li, Xu Tang, Yao Hu, Han Pan, and 1 others. 2024. Ssr-encoder: Encoding selective subject representation for subject-driven generation. In *CVPR*, pages 8069–8078.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International conference on machine learning*, pages 12697–12706. PMLR.

Yucheng Zhou, Hao Li, and Jianbing Shen. 2026a. Condition errors refinement in autoregressive image generation with diffusion loss. In *The Fourteenth International Conference on Learning Representations*.

Yucheng Zhou, Xiang Li, Qianning Wang, and Jianbing Shen. 2024a. Visual in-context learning for large vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15890–15902.

Yucheng Zhou, Zhi Rao, Jun Wan, and Jianbing Shen. 2024b. Rethinking visual dependency in long-context reasoning for large vision-language models. *arXiv preprint arXiv:2410.19732*.

Yucheng Zhou, Jianbing Shen, and Yu Cheng. 2025. Weak to strong generalization for large language models with multi-capabilities. In *The Thirteenth International Conference on Learning Representations*.

Yucheng Zhou, Jihai Zhang, Guanjie Chen, Jianbing Shen, and Yu Cheng. 2026b. Less is more: Vision representation compression for efficient video generation with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 13826–13834.

A Prompt

Caption Generation

Create a scene for an image generation model by semantically choosing at least half of these objects: {classes_str} to fit a coherent setting. Use any style. Describe the scene in one concise English paragraph. no think.

Object Filtering

Image: {Image_with_box_label}
Text:
I have a scene description to evaluate. Here's the info:
- Description: {caption}
Think these questions in mind to check the description:
1. Does the description avoid excessive adjectives or storytelling? (No more than 2 adjectives, no narrative.)
2. Are all classes ({', '.join(classes_str)}) mentioned in the description?
Output:
- If any answer is 'No', list violations like ['Missing class: xx'] and end with 'Please revise.'
- If all answers are 'Yes', say 'Meets all criteria.'

Subject Image Transformation (Simple)

Image: {Image_cropped}
Text:
A {class_name} viewed from a different perspective, maintaining its core features and details.

Subject Image Transformation (Complex)

Image: {Image_cropped}
Text:
i) (*Random Optional*) *With one randomly sampled object:*
A cozy scene shows a {class_name} sitting close to a {random_class}, with their positions suggesting a natural interaction. The environment is filled with random objects, and neither item dominates the scene.
ii) (*Random Optional*) *With two randomly sampled objects:*
The {class_name}, partially out of focus, lies near the camera; a {random_class_1} stands in full view, and a {random_class_2} is seen beyond a pile of scattered objects.
iii) (*Random Optional*) *With three randomly sampled objects:*
In a sun-drenched garden, the {class_name} leans gently against a wooden crate. A {random_class_1} is placed on the grass nearby, with a {random_class_2} resting atop the crate. A {random_class_3} hangs lazily from a tree branch, swaying slightly with the breeze.

Simulated User Instruction Generation

i) (*Optional for Training*) *With subject ID:*
Image: {Image_with_box_label}
Text:
Create a scene description for image annotation using {classes_str}
Use short phrases, and describe how these objects might relate or be arranged in a scene. Clearly refer to each object with its source image (*e.g.*, cat from image 1). Allow slight ambiguity, and mention each given object at least once relative to another or a landmark. Keep it concise, 5-30 words.
Tempelate:
The dog from image 0 rolls around in the grass while the cat from image 1 watches from a rock. A tall tree from image 2 casts shadows across the scene.
ii) (*Optional for Training*) *Without subject ID:*
Image: {Image}
Text:
Create a detailed prompt for a generative model to create an image, depicting a harmonious scene composed of the following objects: {classes_str}. Describe how these objects interact or are arranged in the scene to form a cohesive and visually appealing composition.

CoT Generation

Image: {Image_with_box_label}
Text:
You are an expert in multi-image scene composition and visual grounding. Given initial prompt with subjects from different sub-image, create a detailed composition, step-by-step reasoning (CoT) to describe the entire scene in given target image.
Input - **Initial Prompt**: "{initial_prompt}"
Note 1. Clearly describe spatial relationships and interactions between objects (use terms like beside, behind, above, near, aligned with, etc.).
2. Maintain a clear and logical flow, progressively describing the background, foreground, object positions, and their visual relationships.
3. Include image IDs when needed to distinguish between multiple objects of the same category.
4. Language: English. Provide at least 300 words.

B Data Example

Data Example 1:

Sampled Classes

"toaster", "lamp", "plate", "potted plant", "chair couch", "radiator", "book", "sneakers", "sneakers leather shoes", "orange"

Detected Classes

"microwave", "desk", "potted plant", "desk cabinet", "trolley", "vase", "carpet", "carpet"

Caption

In the cozy living room, the lamp from image 1 casts a warm glow over the table where the toaster from image 0 sits, next to the plate from image 2. The potted plant from image 3 adds a touch of greenery, while the chair couch from image 4 provides a comfortable spot for reading. The radiator from image 5 ensures warmth, and the book from image 6 rests nearby. Sneakers from image 7 and sneakers leather shoes from image 8 hint at someone's presence, and an orange from image 9 adds a splash of color to the scene.

Simulated User Instruction

The microwave from image 0 sits on the desk from image 1, which is covered by a green carpet from image 6. Nearby, a potted plant from image 2 adds greenery, while a desk cabinet from image 3 provides storage. A trolley from image 4 stands nearby, and a vase from image 5 adds color to the scene. Carpet from image 7 also appears in the scene.

CoT

Detailed Composition and Spatial Relationships\n\n#### Background:\nThe scene opens with a cozy backyard setting, featuring a house with a red and white striped awning over the entrance. The house has a wooden deck leading up to the front door, which is partially visible through the awning. The surrounding area is lush with greenery, including bushes and trees, creating a serene and inviting atmosphere. This backdrop sets the stage for the rest of the elements in the scene.\n\n#### Foreground:\nIn the foreground, there is a well-maintained garden area with various plants and flowers. A large, vibrant purple cabbage is prominently displayed in a terracotta pot, adding a pop of color and texture to the scene. Nearby, a small gray cat is sitting on the grass, seemingly relaxed and enjoying the peaceful environment. A few red flowers are scattered around the garden, enhancing the natural beauty of the space.\n\n#### Desk and Microwave:\nMoving forward, we see a wooden desk placed on the patio. On this desk, there is a microwave from image 0, which is positioned centrally. The desk itself is covered by a green carpet from image 7, providing a soft and comfortable surface for the microwave and other items. The desk cabinet from image 3 is situated behind the desk, offering additional storage space. The desk cabinet is aligned with the desk, ensuring that it complements the overall setup without obstructing any view or functionality.\n\n#### Trolley and Vase:\nTo the right of the desk, a blue trolley from image 4 is parked. The trolley is empty but appears ready for use, suggesting it might be used for transporting items within the garden or around the house. Adjacent to the trolley, a vase from image 5 is placed on the ground, adding a touch of elegance and color to the scene. The vase is filled with red flowers, which match the flowers in the garden, creating a cohesive and harmonious look.\n\n#### Potted Plant:\nA potted plant from image 2 is positioned near the trolley, adding more greenery to the scene. The plant is placed on the ground, slightly away from the desk and trolley, allowing it to blend naturally into the environment. Its presence enhances the overall aesthetic of the backyard, making it feel more alive and vibrant.\n\n#### Overall Scene:\nThe entire scene is a harmonious blend of indoor and outdoor elements, creating a functional yet aesthetically pleasing environment. The microwave on the desk, the green carpet, the desk cabinet, the trolley, the vase, and the potted plant all contribute to a well-organized and inviting space. The background of the house and the lush greenery provide a calming and natural backdrop, making the scene feel both practical and serene. The spatial relationships between these objects are carefully arranged to ensure that each element has its own place while still contributing to the overall harmony of the scene."

Spatial Layout Prompt

Here is the segmentation map focusing on piano, person, flower, carpet, chair, stool, frame, desk cabinet, cabinet, blackboard:\n<patch>[0] others [1] others [2] others [3] others [4] others [5] others [6] others [7] others [8] blackboard [9] others [10] others [11] others [12] others [13] others [14] others [15] others [16] blackboard [17] others [18] others [19] others [20] others [21] others [22] cabinet [23] cabinet [24] blackboard [25] frame [26] others [27] others [28] person [29] others [30] cabinet, flower [31] cabinet, flower [32] others [33] frame [34] others [35] person [36] person [37] chair [38] flower, piano [39] flower, piano [40] desk cabinet [41] desk cabinet [42] desk cabinet [43] person [44] chair, person [45] chair [46] piano [47] piano [48] desk cabinet [49] desk cabinet [50] desk cabinet [51] chair, person [52] carpet, chair [53] carpet, chair [54] carpet, piano [55] piano [56] desk cabinet, stool [57] desk cabinet [58] desk cabinet [59] carpet [60] carpet [61] carpet [62] carpet [63] others</patch>\n Now, generate an image.

Data Example 2:

Sampled Classes

"ship", "sports car"

Detected Classes

"ship", "sports car"

Caption

A sleek red sports car speeds along a coastal highway, its shiny body reflecting the golden hues of a setting sun, while in the distance, a large ship cuts through the waves of the ocean, its massive silhouette contrasting with the car's nimble form, capturing the essence of speed and adventure against the vastness of the sea.

Simulated User Instruction

Imagine the ship from image 0 sailing in the distance, with the sports car from image 1 parked on the road near the ocean, both under the warm glow of the setting sun.

CoT

Detailed Composition and Reasoning\n\nThe scene is a harmonious blend of two distinct elements: a ship from image 0 and a sports car from image 1, set against the backdrop of a breathtaking sunset over the ocean. The composition is carefully crafted to ensure that each element contributes to a cohesive and visually appealing narrative.#### Background:\n- **Sunset**: The sky is painted in warm hues of orange and yellow, transitioning into a soft gradient as it meets the horizon. The sun is positioned low on the horizon, casting a golden glow across the entire scene. This creates a serene and tranquil atmosphere, perfect for the contemplative mood suggested by the ship and the dynamic energy of the sports car.\n- **Ocean**: The calm waters of the ocean reflect the vibrant colors of the sunset, creating a shimmering effect that enhances the overall beauty of the scene. The waves gently lap against the rocky cliffs, adding a sense of movement and life to the otherwise still landscape.\n- **Cliffs**: On the right side of the image, rugged cliffs rise steeply from the water's edge. Their dark, textured surfaces contrast sharply with the bright colors of the sunset, providing a dramatic and natural frame for the scene.#### Foreground:\n- **Sports Car (Image 1)**: A sleek red sports car is positioned prominently in the foreground, driving along a coastal road. The car is angled slightly towards the viewer, suggesting motion and speed. Its polished exterior reflects the warm light of the sunset, adding depth and vibrancy to the image. The car's design is modern and aerodynamic, with sharp lines and curves that emphasize its high-performance capabilities.\n- **Road**: The road on which the car is traveling is a two-lane highway with a metal guardrail running parallel to the cliffside. The road appears smooth and well-maintained, curving gently along the coastline. The guardrail adds a sense of safety and structure to the scene, while also emphasizing the proximity of the car to the edge of the cliff.#### Spatial Relationships:\n- **Ship (Image 0)**: The ship is situated further out in the ocean, positioned behind the cliffs and slightly to the left of the sports car. It is a large cargo ship, identifiable by its size and the cranes visible on its deck. The ship is sailing away from the viewer, moving towards the horizon where it blends seamlessly with the sunset. Its presence adds a sense of scale and grandeur to the scene, contrasting with the smaller, more personal scale of the sports car.\n- **Alignment and Interaction**: The sports car is positioned closer to the viewer, creating a sense of depth and perspective. The ship, being farther away, provides a sense of distance and vastness. The cliffs act as a natural barrier between the two elements, subtly separating them yet keeping them within the same frame. The warm light of the sunset unifies the scene, tying together the ship, the car, and the ocean into a single, cohesive composition.#### Visual Flow:\nThe composition flows naturally from the foreground to the background. The sports car, with its vibrant red color and sleek design, draws the viewer's attention immediately. As the eye moves upward, the warm glow of the sunset and the reflective ocean create a sense of serenity and beauty. Finally, the ship in the distance adds a layer of intrigue and scale, completing the scene with a sense of adventure and exploration.In summary, the scene is a masterful blend of motion and tranquility, with the sports car symbolizing human ambition and the ship representing the vastness of the world. The sunset serves as a unifying element, creating a harmonious balance between the different components of the image.

Spatial Layout Prompt

Here is the segmentation map focusing on ship, sports car:\n<patch>[0] others [1] others [2] others [3] others [4] others [5] others [6] others [7] others [8] others [9] others [10] others [11] others [12] others [13] others [14] others [15] others [16] others [17] ship [18] ship [19] others [20] others [21] others [22] others [23] others [24] others [25] ship [26] ship [27] ship [28] others [29] others [30] others [31] others [32] others [33] sports car [34] sports car [35] sports car [36] sports car [37] sports car [38] others [39] others [40] sports car [41] sports car [42] sports car [43] sports car [44] sports car [45] sports car [46] sports car [47] others [48] others [49] others [50] others [51] others [52] sports car [53] sports car [54] sports car [55] others [56] others [57] others [58] others [59] others [60] others [61] others [62] others [63] others</patch>\n Now, generate an image.

Data Example 3:

Sampled Classes

"trumpet", "sushi", "flute", "tissue", "brush", "bracelet", "piano", "trophy", "coffee machine", "candy", "globe"

Detected Classes

"piano", "person", "flower", "carpet", "chair", "stool", "frame", "desk cabinet", "cabinet", "blackboard"

Caption

In a cozy, sunlit studio with floor-to-ceiling windows, a painter sits at an easel, a brush in hand, surrounded by scattered art supplies and a small globe on a nearby shelf, symbolizing inspiration from around the world. To their left, a coffee machine steams quietly, while a tray of sushi and a candy dish sit on the desk, offering a delightful break. Nearby, a trumpet and a flute rest on a piano, suggesting a dual passion for music and art. A trophy sits on the windowsill, reflecting the artist's past achievements, while a tissue and a bracelet lie nearby, hinting at a moment of pause or reflection.

Simulated User Instruction

The person from image 1 sits on a chair from image 4, painting near a piano from image 0. A flower from image 2 adorns the room, and a carpet from image 3 covers the floor. The blackboard from image 9 is visible in the background, and a desk cabinet from image 7 is nearby. Stool from image 5, frame from image 6, cabinet from image 8 also appear in the scene.

CoT

Detailed Composition and Spatial Relationships\n\n#### Background:\nThe scene is set in a cozy, well-lit studio with large windows that offer a serene view of lush greenery outside. The natural light streaming through the windows illuminates the space, creating a warm and inviting atmosphere. In the background, a blackboard from Image 9 is visible, suggesting this might be a creative or educational workspace. The blackboard is positioned against the wall, slightly to the left of the center of the room, providing a focal point for any notes or drawings.\n\n#### Foreground:\nIn the foreground, a person is seated on a wooden chair from Image 4, engrossed in painting on an easel. The individual appears focused, holding a paintbrush in one hand while the other rests on the canvas. The easel is placed on a desk cabinet from Image 7, which is situated to the left of the person. The desk cabinet has several drawers and compartments, offering storage for art supplies and other materials. Adjacent to the desk cabinet, a stool from Image 5 can be seen, likely used by the artist for additional seating or to rest their feet during long sessions.\n\n#### Objects and Their Interactions:\n- **Piano from Image 0**: Positioned to the right of the person, a grand piano adds a touch of elegance and suggests a multifunctional use of the space. The piano's polished surface reflects some of the ambient light, enhancing its prominence in the room.\n- **Flower from Image 2**: A vase of flowers is placed on a small table near the piano, adding a splash of color and life to the scene. The flowers are positioned slightly behind the piano, creating a harmonious balance between the musical instrument and the floral arrangement.\n- **Carpet from Image 3**: A soft pink carpet lies on the wooden floor, extending from the center of the room towards the piano. The carpet not only adds warmth to the space but also provides a comfortable area for the artist to sit or move around.\n- **Desk Cabinet from Image 7**: As mentioned earlier, the desk cabinet serves as a functional piece of furniture, housing various items such as art supplies and personal belongings. It is aligned with the desk where the easel is placed, facilitating easy access to tools and materials.\n- **Frame from Image 6**: A framed picture or artwork hangs on the wall above the desk cabinet, adding an element of personalization and aesthetic appeal to the room. The frame is positioned slightly above and to the left of the desk, drawing the viewer's eye upwards.\n- **Stool from Image 5**: The stool is located to the left of the desk cabinet, providing an alternative seating option for the artist. Its placement ensures that the artist can easily switch between sitting on the chair and using the stool without disrupting their workflow.\n\n#### Overall Flow:\nThe scene is meticulously arranged to create a harmonious and productive environment. The artist is comfortably seated, surrounded by essential tools and personal touches that enhance both functionality and aesthetics. The natural light from the windows complements the warm tones of the wooden furniture and the soft pink carpet, creating a tranquil and inspiring workspace. The presence of the piano and the vase of flowers adds layers of depth and character, making the studio feel like a sanctuary for creativity and relaxation.

Spatial Layout Prompt

Here is the segmentation map focusing on piano, person, flower, carpet, chair, stool, frame, desk cabinet, cabinet, blackboard:\n<patch>[0] others [1] others [2] others [3] others [4] others [5] others [6] others [7] others [8] blackboard [9] others [10] others [11] others [12] others [13] others [14] others [15] others [16] blackboard [17] others [18] others [19] others [20] others [21] others [22] cabinet [23] cabinet [24] blackboard [25] frame [26] others [27] others [28] person [29] others [30] cabinet, flower [31] cabinet, flower [32] others [33] frame [34] others [35] person [36] person [37] chair [38] flower, piano [39] flower, piano [40] desk cabinet [41] desk cabinet [42] desk cabinet [43] person [44] chair, person [45] chair [46] piano [47] piano [48] desk cabinet [49] desk cabinet [50] desk cabinet [51] chair, person [52] carpet, chair [53] carpet, chair [54] carpet, piano [55] piano [56] desk cabinet, stool [57] desk cabinet [58] desk cabinet [59] carpet [60] carpet [61] carpet [62] carpet [63] others</patch>\n Now, generate an image.

C Complex Subject Images Data Generation

Current subject-to-image generation tasks typically focus on generating scenes based on a few subject images corresponding to target objects. However, in real-world applications, users may want to select specific objects from complex scene images and generate new scenes based on those selections. Our method, leveraging *in-context learning*, effectively addresses this more practical and challenging scenario.

To support this, we generate a dedicated dataset where subject objects are embedded into complex scenes and transformed accordingly. Since conventional I2I models (*e.g.*, UNO-FLUX) lack reasoning capabilities, we first need to construct rich and contextually meaningful prompts.

Specifically, we traverse our previously generated dataset to compile a comprehensive list of object categories. This list is then provided to GPT-4o (Hurst et al., 2024), to automatically generate a dictionary mapping each category to a set of semantically related categories.

Method	DINO ↑	CLIP-I ↑	CLIP-T ↑
SDXL + IP-Adapter	0.571	0.738	0.294
PixArt-Σ + IP-Adapter	0.583	0.746	0.298
DiT-XL + Subject Adapter	0.589	0.751	0.301
UNO (FLUX-1.0-DEV)	0.541	0.721	0.296
MUSIC (Ours)	**0.622**	**0.812**	**0.322**

Table 3: Comparison with strong diffusion backbones adapted for subject-driven generation on MSIC.

Pipeline Stage	Failure Rate (%)	Mitigation Strategy
T2I semantic mismatch	9.8	Prompt filtering and regeneration
OVD miss / false detection	14.6	Area threshold + VLM verification
VLM hallucinated validation	6.3	ID-string consistency check
SAM2 incorrect mask	11.2	CLIP-based mask vs. box selection
Final retained samples		**68.1%**

Table 4: Per-stage error rates and mitigation strategies in the data construction pipeline.

For example:

- **Key**: *Chair* → **Value**: *Stool, Bench, Couch*
- **Key**: *Bottle* → **Value**: *Cup, Glass, Jar*

Using this dictionary, we instruct GPT-4o to produce diverse prompt templates that capture natural object interactions and varied scene compositions. Examples include:

- *A cozy scene shows a {class_name} sitting close to a {random_class}, with their positions suggesting a natural interaction. The environment is filled with random objects, and neither item dominates the scene.*
- *The {class_name}, partially out of focus, lies near the camera; a {random_class_1} stands in full view, and a {random_class_2} is seen beyond a pile of scattered objects.*
- *In a sun-drenched garden, the {class_name} leans gently against a wooden crate. A {random_class_1} is placed on the grass nearby, with a {random_class_2} resting atop the crate. A {random_class_3} hangs lazily from a tree branch, swaying slightly with the breeze.*

When embedding subject images into complex scenes, we randomly select both a prompt template and object categories to construct a detailed prompt. This prompt is then provided to the T2I model to generate complex subject cases.

D Additional Experimental Analysis and Validation

D.1 Comparison with Stronger Diffusion Backbones

To better position MUSIC within the current state-of-the-art landscape, we provide additional comparisons with strong diffusion-based backbones adapted for subject-driven generation. While many recent diffusion transformers do not natively support multi-subject in-context conditioning with reference images, we equip them with subject adapters for a controlled and fair comparison. All models are evaluated on the MSIC benchmark using identical prompts and reference subjects. As shown in Table 3, despite stronger diffusion backbones improving base image fidelity, they still exhibit subject omission and identity entanglement as the number of subjects increases. MUSIC consistently outperforms these methods due to its explicit reasoning and planning formulation.

D.2 Error Propagation Analysis of the Automated Data Pipeline

Our automated data construction pipeline integrates multiple pretrained foundation models, each of which may introduce noise. To quantify cumulative error propagation, we measure per-stage failure rates on a random subset of generated samples and summarize the mitigation strategies used at each stage (Table 4). Although individual stages introduce moderate noise, cascading filters significantly reduce error accumulation. Empirically, models trained on this automatically curated dataset achieve consistent gains, indicating robustness to

Method	Patch IoU ↑	Category Coverage ↑
Random layout	0.21	0.48
UNO (implicit layout)	0.34	0.61
MUSIC (predicted layout)	**0.47**	**0.76**

Table 5: Agreement between predicted spatial layouts and segmentation masks.

Method	Relation Accuracy ↑	Identity Fidelity ↑
UNO	0.63	0.72
MUSIC	**0.74**	**0.81**

Table 6: Performance on relative spatial relation prompts on MSIC.

realistic supervision noise.

D.3 Validation of the Learned Spatial Layout Prior

The semantics-driven spatial layout in MUSIC serves as a high-level prior rather than pixel-accurate supervision. To verify that MUSIC learns meaningful spatial structure rather than memorizing noisy layouts, we measure the agreement between predicted layouts and SAM2 segmentation masks on held-out images. As reported in Table 5, the learned spatial layout prior aligns well with object spatial distributions and is not dominated by hallucinated patterns.

D.4 Effectiveness of Vision Chain-of-Thought for Spatial Relations

Vision Chain-of-Thought (CoT) in MUSIC focuses on relative spatial relations (e.g., left, right, behind), which are more robust and natural for complex multi-subject composition. We evaluate relational accuracy on prompts requiring relative spatial reasoning (Table 6).