Squeezing Capacity from Multimodal Large Language Models for Subject-driven Generation

Shuhong Zheng[1*] **Aashish Kumar Misraa**[2] **Yu-Teng Li**[2]
Yu-Jhe Li[3*†] **Igor Gilitschenski**[1†]
[1]University of Toronto & Vector Institute [2]Adobe [3]Google

Abstract

Subject-driven image generation aims to synthesize new images that preserve the identity of the given subject while following textual instructions. Existing approaches often encode text and reference images separately. This limits cross-modal reasoning abilities and causes copy-paste artifacts. Recent frameworks that connect multimodal models and diffusion models improve instruction following, but largely overlook identity preservation. To address these limitations, we condition diffusion models on Multimodal Large Language Models (MLLMs) that jointly encode text and reference images, and augment it with VAE-based identity conditioning. A novel Dual Layer Aggregation (DLA) module is designed to aggregate multi-level MLLM features for optimal conditioning, and a multi-stage denoising strategy is applied to progressively balance the semantic information from MLLM and fine-detail identity from VAE during inference. Extensive experiments demonstrate that our approach harmonizes multimodal understanding with identity preservation, mitigates copy-paste issues, and achieves superior performance regarding human preference on subject-driven image generation. Our project website is available at https://zsh2000.github.io/squeeze-mllm-subject-gen/.

1 Introduction

Subject-driven image generation aims to synthesize new content while preserving the visual identity of a specific subject. Early approaches [2, 9, 19, 48, 64, 65, 98, 106], such as DreamBooth [83] and Textual Inversion [23], personalize pretrained diffusion models via per-subject fine-tuning, achieving strong identity fidelity at the cost of scalability. Subsequent works [17, 38, 66, 67, 76, 87, 127] adopt reference-image conditioning to avoid retraining, where models like IP-Adapter [123] extract subject features at inference time. More recent efforts [6, 18, 22, 72, 110, 124] further enhance zero-shot subject generalization through VAE-based (Variational Autoencoder-based [45]) token conditioning. However, these pipelines still process text and reference images separately, limiting multimodal understanding and often producing copy-paste artifacts or identity drift on complex prompts.

In parallel, multimodal large language models (MLLMs) [61, 62] have demonstrated strong abilities in joint text-image reasoning and structured control [90]. Systems [15, 89] such as DreamEngine [11], Qwen-Image [107], and EasyRef [131] integrate MLLMs into diffusion decoders to parse interleaved multimodal instructions, enabling more flexible prompt interpretation. Yet, these designs typically rely only on the MLLM's final-layer features (*e.g.,* Qwen-Image, EasyRef), or combine ViT features which contain fine details, with final-layer outputs via scalar mixing (*e.g.,* DreamEngine). These models often neglect fine-grained visual cues which are crucial for identity, thereby leading to suboptimal identity preservation.

[†]Joint Advising
[*]Work done in Adobe

Preprint.

In this work, we unify these two directions by introducing an MLLM-driven subject conditioning framework that jointly encodes text and reference images within a shared multimodal space, and enhances ID preservation with VAE conditioning. This joint encoding enables the model to perform multimodal reasoning and coherently preserve subject identity, beyond the representational limits of pure VAE-based encoders. However, this unification is non-trivial due to the different feature structures of text and image tokens in MLLMs. The discrepancy between text and image features makes it fundamentally inadequate to directly fuse modalities or rely on a single-layer representation for conditioning. To effectively align MLLM embeddings with diffusion features, we design an innovative Dual Layer Aggregation (DLA) mechanism, that adopts layerwise attention pooling to separately aggregate text and visual embeddings. Instead of conditioning solely on the MLLM's final layer feature, the DLA takes the aggregated features from all transformer layers in the MLLM as input, to fully leverage its multimodal prompt understanding capability. We also justify the mechanism of aggregation by analyzing the roles and effectiveness of different layer groups (*i.e.,* early, middle, and late layers) within MLLM in the experimental study.

In addition, directly combining MLLM embeddings with VAE-based identity enhancement can cause embedding conflicts, as both contain overlapping visual representations. To reconcile these signals, a two-stage training strategy is invented to first enable multimodal conditioning from MLLM, before combining the optimization with the high-frequency identity details from VAE features. To further balance the multimodal conditioning from the MLLM and the identity details provided by the VAE, we propose a multi-stage denoising strategy: the diffusion model first denoises under MLLM guidance to establish global semantics, then jointly refines with both modalities, and finally focuses on VAE-conditioned fine details. As shown in Figure 1, this staged process effectively harmonizes the two embedding sources, alleviating copy-paste artifacts common in VAE-based pipelines, while providing richer reasoning ability and instruction-aware, identity-preserving generation compared to existing frameworks. Our contributions can be summarized as follows: we propose a Dual Layer Aggregation (DLA) module to aggregate text and visual embeddings across MLLM layers for improved conditioning, along with a multi-stage denoising strategy that balances semantic reasoning and fine-grained identity during generation. Also, we provide a detailed analysis of MLLM layer representations and their roles in diffusion conditioning under different fusion strategies. Extensive experiments demonstrate competitive performance in multimodal understanding and identity preservation over prior subject-driven methods.

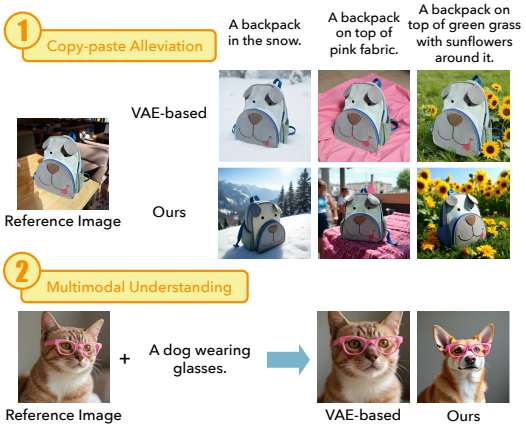

Figure 1: Benefits of leveraging MLLMs for subject-driven generation. MLLMs mitigate the *copy-paste issue* within VAE-based methods, and enables the *multimodal understanding* of the subject-driven generation pipeline by jointly modeling input image and text, while VAE-based methods encode them separately.

2 Related Work

Subject-driven Generation focuses on preserving the identity or visual characteristics of a specific subject within the synthesized images. Early optimization-based approaches [1, 12, 24, 32] such as DreamBooth [83], Textual Inversion [23], and LoRA [35] adapt pretrained diffusion models to new identities by introducing subject-specific parameters, but require costly per-subject fine-tuning. To eliminate this need, recent methods employ explicit reference encoders or adapters that extract identity features directly from input images and condition the diffusion process at inference time (*e.g.*, IP-Adapter [123], BLIP-Diffusion [50]). Transformer-based diffusion decoders (DiT) have further incorporated such reference conditioning [36, 58] through lightweight modules like IC-LoRA [37]. Subsequent research [39, 44, 51, 59, 116] enhances facial fidelity [52, 77, 95, 103, 112, 117], multi-reference composition [31, 43, 84, 86, 88, 96, 99, 100, 105, 118–120, 126, 128], computational efficiency [16, 41, 53, 54, 56, 102, 121, 122], and multimodal controllability [21, 30, 33, 40, 49, 60, 92, 101, 113–115]. Recently, UNO [110], UMO [14], USO [111], and DreamO [69] achieve zero-shot

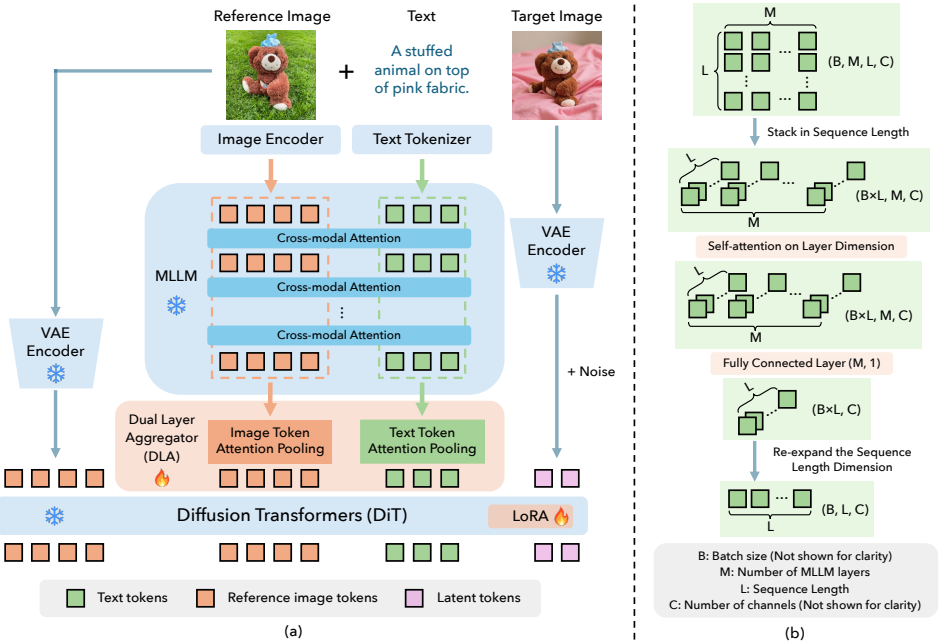

Figure 2: Overview of our framework. (a) The full model architecture, consisting of an MLLM for multimodal understanding, a VAE encoder for mapping images into latent space, a DiT backbone for diffusion denoising, and a Dual Layer Aggregator (DLA) that aligns MLLM embeddings for DiT. (b) Details of the token attention pooling module inside the DLA module, including its layerwise attention and pooling operations, to form MLLM aggregated embeddings from all MLLM layers.

generation conditioned by multiple images leveraging VAE-based token conditioning. However, these identity-preserving and control-oriented pipelines remain largely decoupled from large multimodal language models (MLLMs), lacking the semantic reasoning and contextual understanding necessary for flexible, instruction-aware identity control. Due to the limit of space, more discussions on the related work can be found in Section C in the Appendix.

3 Method

Given a text prompt $\mathcal{T}$ and a set of reference images $\mathcal{I} = \{I_n\}_{n=1}^{N}$, our method produces an image $\hat{I} = \mathcal{G}(\mathcal{T}, \mathcal{I})$ that aligns with the textual description while preserving the identity of the reference images. Our approach, visualized in Figure 2(a), is built on top of a Diffusion Transformer (DiT) backbone (Section 3.1) conditioned on a Multimodal Large Language Model (MLLM) and a VAE encoder. Specifically, we propose to use layerwise attention pooling (Section 3.2) and propose a Dual Layer Aggregator (DLA) module (Section 3.3) that allows to extract aggregated features from MLLM layers for text and image modalities. The architecture unifies MLLM for multimodal understanding and VAE for deriving high-fidelity identity details. To better reconcile capabilities of MLLM and VAE, we propose a multi-stage denoising process (Section 3.4) that allows integrating different conditioning branches and design a two-stage training strategy (Section 3.5).

3.1 Background: Diffusion Transformers

Diffusion models learn a mapping from a simple prior distribution to the data manifold through iterative denoising. Given a data sample $\mathbf{x}_0 \sim q(\mathbf{x}_0)$, the forward process gradually perturbs it with Gaussian noise under a variance schedule $\{\alpha_t\}_{t=1}^{T}$:

$$q(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\alpha_t}\mathbf{x}_0, (1-\alpha_t)\mathbf{I}), \tag{1}$$

where $\mathbf{x}_T$ approaches an isotropic Gaussian. The reverse process is learned by predicting either the added noise or the clean sample $\mathbf{x}_0$ with the denoising network $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, \mathbf{c})$, conditioned on control signals $\mathbf{c}$ such as text or image embeddings. Recently, Rectified Flow [63] reformulates diffusion as a

deterministic *transport process* parameterized by a time-dependent velocity field $\mathbf{v}_\theta(\mathbf{x}_t, t)$:

$$\frac{\mathrm{d}\mathbf{x}_t}{\mathrm{d}t} = \mathbf{v}_\theta(\mathbf{x}_t, t), \quad \mathbf{x}_0 = \mathbf{x}_1 + \int_0^1 \mathbf{v}_\theta(\mathbf{x}_t, t)\,\mathrm{d}t. \tag{2}$$

This rectified formulation stabilizes training and simplifies inference by eliminating stochastic sampling steps. The objective becomes a velocity-matching loss:

$$\mathcal{L}_{\text{RF}} = \mathbb{E}_{t,\mathbf{x}_0,\mathbf{x}_1}\left[\|\mathbf{v}_\theta(\mathbf{x}_t, t) - (\mathbf{x}_0 - \mathbf{x}_1)\|_2^2\right]. \tag{3}$$

Building on this, Diffusion Transformers (DiT) [73] replace the standard UNet backbone with a transformer that operates on *patch tokens*. At each timestep t, the noisy image $\mathbf{x}_t$ is first projected into a latent representation $\mathbf{z}_t \in \mathbb{R}^{H\times W\times D}$. This latent is then flattened into a sequence of patch embeddings, augmented with timestep and conditioning tokens, and processed through self-attention layers to predict either the velocity or noise tokens for denoising.

In our experiments, we adopt *FLUX.1 dev* [5], a recent DiT-based architecture employing rectified flow parameterization as our backbone due to its training stability, synthesis capability, and modular conditioning design. Flux provides a flexible transformer-based diffusion decoder that seamlessly integrates multimodal embeddings, making it a strong foundation for our proposed MLLM-driven subject conditioning framework.

3.2 Basic Module: Layerwise Attention Pooling

Existing methods that connect MLLMs with diffusion models mainly focus on text-to-image generation and typically extract the single final layer feature as conditioning tokens [107, 131], assuming that the last layer contains the most informative semantic representation after multimodal reasoning. However, this strategy is suboptimal for subject-driven generation, where both *text adherence* and *identity preservation* are equally important.

Motivation. Since most MLLMs are optimized for high-level reasoning tasks such as VQA, their image tokens tend to lose fine-grained texture and appearance details in deeper layers. As also observed in [10], the visual representation in MLLMs shifts from low-level appearance to high-level semantics across layers when the layer dives deeper. This creates a *representation mismatch*: no single layer provides both the semantic completeness required for text alignment and the fine-grained fidelity required for identity preservation. To alleviate this issue, we leverage a Layerwise Attention Pooling (LAP) mechanism that integrates features across multiple MLLM layers to retain both higher-level semantic and lower-level structural information.

LAP Module. Given MLLM feature maps from all transformer layers $\mathcal{F} = \{F_i\}_{i=0}^{M}$ (M is the number of MLLM layers), where $F_i \in \mathbb{R}^{B\times L\times C}$ (B is the batch size, L is the sequence length, and C is the channel number), LAP produces a summarized representation $\hat{F} \in \mathbb{R}^{B\times L\times C}$ via attention over the layer axis. Concretely, LAP implements a lightweight multi-head attention mechanism where the layer index is treated as the sequence dimension, followed by a fully connected projection for adaptive layer weighting, as shown in Figure 2(b).

3.3 Dual Layer Aggregator

Observation and Motivation from Single LAP Module. As illustrated in Figure 3(a), preliminary experiments using a single LAP module to jointly summarize text and image tokens revealed a trade-off between identity preservation and text alignment for different checkpoints obtained during the optimization process. When trained together, the model tends to overfit to one modality, degrading the performance of the other. Further analysis in Figure 3(b) on text-to-image (T2I) and image-to-image (I2I) reconstruction tasks breaks down this issue, and shows that the layer-wise attention obtained from text and image tokens differ significantly, reflecting distinct hierarchical information patterns for each modality.

DLA for Multimodal Processing. Motivated by the observed issue, we introduce a Dual Layer Aggregator (DLA) that decouples layerwise aggregation across modalities. DLA consists of two separate LAP modules: one for text tokens and one for image tokens. Each LAP specializes in summarizing layerwise features most relevant to its modality—text LAP emphasizes on semantic fidelity and the prompt, while image LAP focuses on subject appearance and identity consistency.

Importantly, this design does not sacrifice cross-modal interaction, as MLLMs already enable multimodal information to flow within intermediate layers, which means image tokens inside MLLM already absorb cross-modal information from text, and vice versa. Therefore, DLA avoids redundant multimodal fusion and instead focuses on modality-aware layerwise information processing.

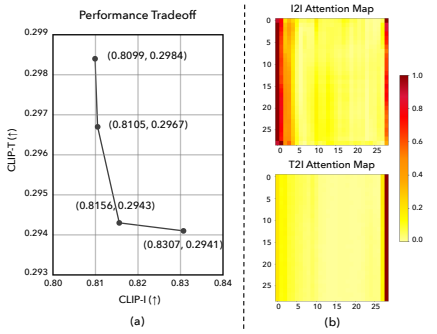

Figure 3: Preliminary analysis on single Layer-wise Attention Pooling (LAP) across text and image modalities. (a) Performance tradeoff with different checkpoints when optimizing with a single LAP. (b) Attention maps from the model trained solely on I2I task and the model trained on T2I task, where both use single LAP model.

With the designed DLA module, each modality-specific LAP can focus on effectively aggregating intra-modal information without redundant fusion learning. Empirically, we observe that early and late layers in the MLLM often exhibit stronger activations corresponding to appearance and semantic cues, respectively. To maintain model-agnostic flexibility, we apply LAP to all MLLM layers, allowing DLA to adaptively learn each layer's contribution to identity or text following. This ensures robustness when adapting to different MLLM architectures with varying attention behaviors.

3.4 Multi-stage Timestep-aware Denoising

The VAE encoder in diffusion models serves as a strong visual tokenizer that effectively captures detailed subject identity from reference images [14, 110]. While VAEs preserve fine-grained appearance, they often suffer from copy-paste artifacts and lack semantic understanding. In contrast, MLLMs jointly encode text and images, offering better reasoning and layout understanding, but relatively weaker identity fidelity. To address the above limitations with single-source features, we leverage both conditioning sources to combine the complementary strengths of VAEs and MLLMs, and propose a *multi-stage denoising process* that activates different conditioning branches along the denoising timesteps. This design aligns with the inherent coarse-to-fine nature of diffusion: earlier steps capture semantics and global layout, and later steps refine local details. Specifically, MLLM conditioning is used in early steps for semantic and compositional reasoning; both MLLM and VAE conditioning are applied in the middle for balanced control; and only VAE conditioning is used in late steps for detailed identity refinement.

Formulation. The denoising network predicts the clean sample at each step as:

$$\hat{\mathbf{x}}_{t-1} = f_\theta(\mathbf{x_t}, \mathbf{c}_{\text{MLLM}} \cdot M_{\text{MLLM}}(t), \mathbf{c}_{\text{VAE}} \cdot M_{\text{VAE}}(t)), \tag{4}$$

where f_θ denotes the denoising transformer, and $\mathbf{c}_{\text{MLLM}}$ and $\mathbf{c}_{\text{VAE}}$ are conditioning embeddings from the two encoders. The timestep-dependent masks $M_{\text{MLLM}}(t), M_{\text{VAE}}(t) \in \{0, 1\}$ control which branches are active. During training, the reference image input for either branch (MLLM or VAE) is randomly dropped to ensure robustness. As a result, the whole system can naturally handle scenarios when only one of the sources has the reference image input.

We define three denoising stages parameterized by τ_1 and τ_2:

$$M_{\text{MLLM}}(t), M_{\text{VAE}}(t) = \begin{cases} (1,0), & t \geq \tau_1 \quad \text{(early)} \\ (1,1), & \tau_2 \leq t < \tau_1 \quad \text{(middle)} \\ (0,1), & t < \tau_2 \quad \text{(late)}. \end{cases} \tag{5}$$

Integration with rectified flow. This stage-aware conditioning naturally integrates with our rectified flow objective. As the rectified flow continuously transports samples from noise to data, the conditioning signal shifts from semantic alignment via the MLLM, to fine-detailed identity refinement via the VAE near the data manifold, achieving coherent and instruction-aware subject generation.

3.5 Two-stage Training Strategy

Training a diffusion system conditioned on both MLLM and VAE embeddings presents a unique challenge. Since our timestep-aware denoising process requires the model to function when only

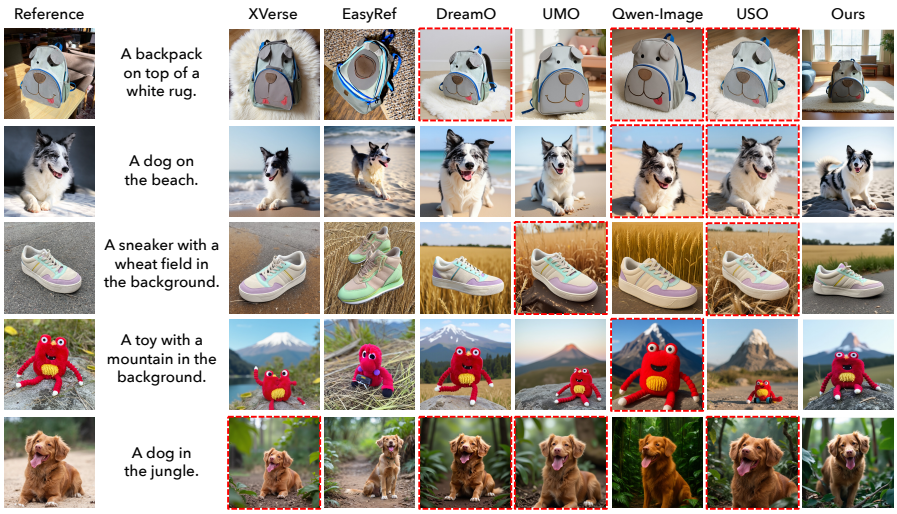

Figure 4: Comparisons of our method with state-of-the-art subject-driven generation approaches. ***Red dashed lines*** indicate instances where other methods suffer from the copy-paste issue. In contrast, our method produces images that maintain strong subject identity, exhibit creative pose variations, and respect underlying physical constraints.

one of the two modalities (MLLM or VAE) is present, both encoders must independently learn to contribute meaningful signals for subject-driven generation. To achieve this, we adopt the following *two-stage training strategy*.

In the first stage, we train the diffusion transformer using only MLLM-derived conditioning. This stage encourages the MLLM to fully exploit its multimodal reasoning ability, and capture identity-related cues from the reference images as well. In the second stage, we jointly train the entire framework—MLLM, VAE, and DiT—enabling the model to balance high-level reasoning from the MLLM with fine-grained identity features from the VAE.

This staged optimization prevents the VAE from dominating identity preservation too early. If trained jointly from scratch, the VAE tends to absorb most of the identity learning, leaving the MLLM under-optimized and ineffective in the early denoising steps—where global structure and appearance are primarily determined. Consequently, once identity information is far off track in the early timesteps, it cannot be recovered later even when VAE conditioning is introduced. Our two-stage strategy therefore ensures both conditioning pathways to contribute effectively throughout the denoising process, leading to harmonized identity fidelity and prompt alignment.

4 Experiments

4.1 Experimental Settings

Dataset. To explore the potential of MLLMs for subject-driven generation, we use only public datasets throughout our experiments. Our model is trained on the publicly available UNO-1M [110], which contains approximately 400K image pairs after filtering with MLLM-based scoring criteria. Each pair features a subject with matched images of the same identity.

Implementation Details. Following the two-stage training strategy described in Section 3.5, we first train the MLLM-DiT framework for 25K steps, and then incorporate both MLLM and VAE as conditioning signals for an additional 10K steps. Training is conducted on 8 NVIDIA H100 GPUs, each with a batch size of 16, using a constant learning rate of 1e-5.

We adopt InternVL3-8B [130] as the MLLM and FLUX.1 dev [5] as the DiT, with a LoRA rank of 512 for finetuning the DiT attention blocks. The MLLM and other weights in DiT are all frozen. During inference, we set timestep-aware denoising thresholds to $\tau_1 = 0.95$ and $\tau_2 = 0.85$, use a cosine denoising schedule, and apply a classifier-free guidance (CFG) value of 2.5 for all stages when evaluating metrics. Section 4.3 further analyzes how these parameters affect performance and provide users with finer control over identity fidelity, pose variation, and overall image quality.

4.2 Comparisons with Existing Methods

In this section, we conduct comprehensive experiments on various aspects to demonstrate the capability of our method. Besides the **(1)** standard benchmark evaluations, **(2)** we propose an evaluation criteria to quantify the copy-paste issue and illustrate that the issue gets largely mitigated. Also, **(3)** better multimodal understanding capability is revealed with both qualitative and quantitative results. Additionally, **(4)** automatic human-aligned evaluation and **(5)** user study demonstrate that our method receives more preference from users compared to existing models.

(1) Standard Benchmark Performance. We compare our model with state-of-the-art subject-driven generation methods including OminiControl [94], OmniGen2 [108], UNO [110], XVerse [8], DreamO [69], USO [111], and UMO [14], as well as recent approaches that connect MLLMs with diffusion models, including DreamEngine [11], Qwen-Image [107], and EasyRef [131]. As many existing systems rely on private high-quality subject-driven datasets, we also re-train UNO which has public training code with the same UNO-1M data to better show the potential of our method. Following previous works, DreamBench [83] is adopted as our main evaluation benchmark for experimental analysis and ablation study. DINO-I [7] and CLIP-I [78] are used for measuring identity similarity, and CLIP-T is used for text-image alignment. As shown in Table 1, our MLLM-only model (only trained for the first stage with the MLLM-DiT framework) already reaches performance on par with UNO trained under the same conditions, demonstrating the strength of our DLA in extracting multimodal features and identity signals from MLLMs. With both MLLM and VAE conditioning, our full model—trained entirely on public data—achieves performance comparable to state-of-the-art methods. Qualitative comparisons in Figure 4 show that our approach produces *more diverse poses* while preserving identity, and yields *more physically coherent* scenes, avoiding artifacts such as subjects floating above backgrounds. Beyond the standard DreamBench, we also include evaluations on additional benchmarks including XVerseBench [8] and a multi-subject LAMICBench [13] on our model with slight multi-subject adaptation in the Appendix.

Table 1: Quantitative comparison on DreamBench. †Training with the same public data as ours (single-subject data from UNO-1M). First block indicates VAE-based methods while the second block indicates MLLM-based approaches.

Method	DINO-I (↑)	CLIP-I (↑)	CLIP-T (↑)
OminiControl [94]	0.5987	0.7840	0.3186
OmniGen2 [108]	0.7323	0.8268	0.3185
UNO [110]	0.7484	0.8354	0.3040
UNO† [110]	0.6860	0.8161	0.3071
XVerse [8]	0.7215	0.8175	0.3098
DreamO [69]	**0.7537**	0.8356	0.3086
USO [111]	0.7478	0.8263	**0.3213**
UMO [14]	0.7481	0.8339	0.3022
DreamEngine [11]	0.5195	0.7428	0.3006
Qwen-Image [107]	0.7317	0.8261	0.3158
EasyRef [131]	0.6961	0.8153	0.3031
Ours (MLLM only)	0.6788	0.8228	0.2988
Ours (MLLM + VAE)	0.7482	**0.8443**	0.3010

(2) Copy-paste Issue Alleviation. A common failure mode in subject-driven generation—particularly in VAE-based methods—is the *copy-paste* effect, where the generated subject closely mimics the reference pose with minimal variation. This issue is largely overlooked in the prior evaluations in existing works that mainly focus on identity preservation and text alignment. As illustrated in Figure 4, many existing approaches suffer from this behavior (highlighted with red dashed boxes), whereas our method produces subjects with noticeably more diverse and creative poses. To quantify this effect, we adopt Orient Anything [104] to estimate the azimuth and polar angles of subjects in both the reference and generated images, and compute their average orientation discrepancy. We further propose a *"Recall"@ k°* metric—the percentage of generated samples whose orientation angles (both azimuth and polar) are below k°, and report the *Average "Recall" Rate* metric, which is averaged over $k^\circ \in \{5^\circ, 10^\circ, 15^\circ, 20^\circ\}$ for *"Recall"@ k°*. As shown in Table 2, our MLLM-based conditioning significantly mitigates the copy-paste issue compared to previous methods. While the diversity of the generated subject can also be reflected in other factors (*e.g.*, posture), orientation is a crucial and easily measurable indicator, especially for rigid objects, so it serves as a practical proxy for evaluating the copy-paste artifacts.

Table 2: Quantitative comparisons of subject variation between the reference and generated images, measuring the copy-paste issue. We evaluate differences in azimuth and polar angles to assess the subject pose diversity produced by the model, showing its ability to generate more varied poses and reduce copy-paste artifacts.

Method	Azimuth (↑)	Polar (↑)	Average "Recall" Rate (↓)
OmniGen2 [108]	22.6	7.0	0.486
DreamO [69]	22.1	9.6	0.372
USO [111]	20.8	9.6	0.401
Qwen-Image [107]	17.6	7.8	0.460
Ours	**25.7**	**10.4**	**0.349**

(3) Reasoning Capability. The text prompts in DreamBench are relatively simple and require limited cross-modal reasoning, making differences in text-following performance across models

Figure 5: Comparisons on reasoning capability show that VAE-based methods often fail on complex user prompts, producing copy-pasted subjects or incorrect concept binding. MLLM-DiT pipelines like Qwen-Image also struggle in understanding these challenging user prompts, demonstrating the prior solution for connecting MLLM and DiT is suboptimal. In contrast, our method, conditioned solely on MLLM signals, accurately interprets the prompts and associates concepts with the appropriate visual elements.

Table 3: Quantitative comparisons on the constructed benchmark with 350 samples for multimodal reasoning capability in subject-driven generation.

Metric	UNO	DreamO	Qwen-Image	Ours
CLIP-T (↑)	0.2851	0.2888	0.3099	**0.3208**

Table 4: Quantitative comparisons on the human-aligned MLLM-based scores on the subject categories in DreamBench++.

Method	MLLM-based Scores (0-4 scale) (↑)							
	(A)	(B)	(C)	(D)	(E)	(F)	(G)	Average
DreamO	2.837	2.892	2.802	3.402	2.737	2.462	2.737	2.838
UNO	2.539	2.753	2.474	3.027	2.303	2.103	2.576	2.539
USO	2.790	2.868	2.798	3.410	2.663	2.400	2.668	2.800
Ours	**3.119**	**2.969**	**3.006**	**3.568**	**2.962**	**2.601**	**2.847**	**3.010**

Table 5: User study from 30 participants with a total of 1,500 votes on samples from DreamBench and XVerseBench.

Method	XVerse	DreamO	USO	UMO	Ours
Score (1-10 scale) (↑)	5.75	6.31	6.74	6.02	**7.26**

small. Methods like USO can achieve high scores despite occasionally *"copy-pasting"* the subject and placing it on top of the prompted background. More challenging scenarios arise when the user prompt refers to concepts that are not the sole focus of the reference image, or when concept binding needs to be figured out between the text and the visual input. Figure 5 illustrates such cases, where correct generation depends on understanding and reasoning over both modalities. For instance, in the first row, the model must associate the "hat" mentioned in the prompt with the correct region of the reference image, while ignoring irrelevant distractors such as the cat. VAE-based pipelines struggle here because they encode text and reference images independently, limiting their ability to jointly interpret user intent. Qwen-Image with the MLLM-DiT structure also shows multiple failure cases, suggesting that conditioning the DiT solely on the final layer features does not fully leverage the multimodal reasoning capacity of MLLMs. In contrast, our model successfully aligns text and image cues, producing coherent outputs. Notably, even though our model is *trained only on single-subject data*, it can take two reference images as input (last row of Figure 5) and still correctly bind textual concepts to the appropriate visual regions. To further quantitatively verify the multimodal reasoning capability, we construct a benchmark consisting of 350 samples similar to the examples in Figure 5, and evaluate on the text following capability that largely reflects the correctness of the generated images from user instructions. Detailed information about the constructed benchmark can be referred in the supplementary material. As shown in Table 3, our method greatly outperforms the existing models on multimodal understanding on these complex scenarios, because the MLLM in our model jointly encode both images and text that enables cross-modal concept binding and reasoning.

(4) Human-aligned Evaluation. To provide an additional perspective regarding human-aligned preference, we perform evaluation on DreamBench++ [74], which adopts an MLLM-based scoring metric that is aligned with human perception. MLLMs are expected to give an overall score from 0-4 on subject consistency following prompts used in [47, 74, 129] for the generated images, considering multiple aspects including shape, color, and texture. More details about the prompts used for MLLMs are described in the supplementary material. We select seven MLLMs with different architectures and sizes to foster the soundness of the evaluation: (A) GPT-4o [71] (original choice from DreamBench++), (B) Gemma 3 27B [28], (C) Gemini 2.5 Flash [27], (D) Gemini 3 Flash [29], (E) Qwen3-VL-30B-A3B-Thinking [3], (F) Qwen3-VL-235B-A22B-Thinking [3], and (G) Mistral-Small-3.2-24B-Instruct [68]. Table 4 demonstrate the superiority on subject consistency of our method evaluated with all types of MLLMs.

(5) User Study. To further calibrate subject-driven generation quality with human perception, we conduct user study on the images generated by XVerse, DreamO, USO, UMO, and our method. We randomly select 10 samples from DreamBench and XVerseBench, and ask the volunteers to score the generated assets in a scale of 1-10 on the overall quality, where the participants are guided to focus on subject fidelity, text following, *etc*. that are considered important factors for subject-driven generation.

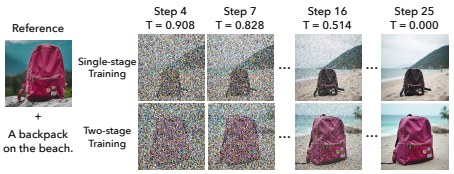

Figure 6: Single-stage training prevents the model from leveraging timestep-aware denoising, limiting both potential performance gains and the flexibility for user control.

Table 7: Comparison regarding single-stage training, with and without timestep-aware denoising (TAD). The results highlight the importance of our two-stage training strategy, which first establishes a well-trained MLLM-DiT system before introducing the VAE.

Method	DINO-I (↑)	CLIP-I (↑)	CLIP-T (↑)
Single-stage Training w/o TAD	0.7184	0.8245	0.2971
Single-stage Training with TAD	0.5763	0.7686	0.2995
Two-stage Training	**0.7482**	**0.8443**	**0.3010**

Details on the instructions and interface of our user study can be referred in the supplementary material. There are 30 volunteers participating in our user study, and a total of 1,500 votes are collected. The results of the user study in Table 5 show that our method is also more subjectively preferred from real user experience.

4.3 Ablation Study

In this section, we provide the ablation study and analysis on the design and training mechanisms of our method. Due to the limit of space, more ablation study and analysis can be found in the Appendix.

(1) Strategies for Leveraging MLLM. We conduct a systematic analysis to identify effective ways to squeeze MLLM capacity to diffusion models for subject-driven generation. We ablate several strategies, including those from prior works: last-layer feature extraction from Qwen-Image [107], scalar blending of MLLM ViT feature with last-layer feature in DreamEngine [11] (blend ViT), and concatenation of MLLM ViT image feature with last-layer text feature (mix ViT). Additionally, we explore different layer selections in our LAP module by partitioning InternVL3-8B's 28 layers into early (0–9), middle (10–19), and late (20–28) layers, and evaluate residual connections to the last layer. Table 6 shows that existing strategies are suboptimal under limited data and resource constraints. Using a single LAP for both text and image preserves identity reasonably but severely compromises text alignment. Residual connections to the last layer do not improve performance and can even degrade it, suggesting that overemphasizing last-layer features may be harmful.

Table 6: Analysis on different strategies of connecting MLLM features to the DiT. The first block reports baselines that rely on last-layer conditioning and their variants. The second block evaluates single LAP configurations, showing that a single LAP for both text and image tokens preserves identity reasonably well, but severely weakens text following.

Method	Selected Layers	Residual Connection	DINO-I (↑)	CLIP-I (↑)	CLIP-T (↑)
Last layer [107]	-	-	0.6566	0.8128	0.2893
Last layer (blend ViT) [11]	-	-	0.7118	0.8286	0.2850
Last layer (mix ViT)	-	-	0.7097	0.8233	0.2946
Single LAP	0-9	×	0.7167	0.8391	0.2969
Single LAP	10-19	×	0.7315	0.8463	0.2957
Single LAP	20-28	×	0.7325	0.8386	0.2981
Single LAP	0-28	×	0.7524	0.8502	0.2878
Single LAP	0-9	✓	0.7282	0.8497	0.2944
Single LAP	10-19	✓	0.7109	0.8377	0.2958
Single LAP	20-28	✓	0.7246	0.8389	0.2963
Single LAP	0-28	✓	0.7242	0.8379	0.2974
DLA (Dual LAP)	0-28	✓	0.7275	0.8401	0.3013
DLA (Dual LAP)	0-28	×	0.7482	0.8443	0.3010

(2) Two-stage Training. Optimizing a framework that integrates both MLLM and VAE for subject-driven generation is non-trivial. In Section 3.5, we propose a two-stage training strategy: first training the MLLM-DiT framework in the initial stage and then adding the VAE in the second stage. Without this staged approach, the capacity from MLLMs for identity preservation cannot be sufficiently unleashed, which prevents the model from fully leveraging the timestep-aware denoising process. As shown in Figure 6, the initial denoising steps conditioned on MLLM largely determine the final appearance of the image; if MLLM has not developed decent identity preservation capability, the VAE in the later stages cannot correct the denoising trajectory. Table 7 further shows that the model trained with single-stage strategy has inferior performance in both identity preservation and text alignment, and its failure to utilize timestep-aware denoising to boost the performance. This occurs because VAE tokens, which are originally optimized for reconstruction, dominate the generation process, therefore reducing the information contribution from MLLM features. As a result, the DiT loses one source of identity information, and the cross-modal reasoning and understanding capabilities from the MLLM are diminished as well.

5 Conclusion

We study towards the optimal strategy to utilize MLLMs for subject-driven generation, with the introduced Dual Layer Aggregation (DLA) module. Our analysis shows that aggregating represen-

tations across all layers and aligning text and visual modalities separately, is critical to achieving strong multimodal understanding and identity preservation. Combined with the VAE's strength in capturing fine-grained visual details, our multi-stage denoising framework and two-stage training strategy further harmonize the conditioning signals from both MLLM and VAE, and provide users with more flexibility during generation.

References

[1] Moab Arar, Rinon Gal, Yuval Atzmon, Gal Chechik, Daniel Cohen-Or, Ariel Shamir, and Amit H. Bermano. Domain-agnostic tuning-encoder for fast personalization of text-to-image models. In *SIGGRAPH Asia*, 2023. 2

[2] Omri Avrahami, Kfir Aberman, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. Break-A-Scene: Extracting multiple concepts from a single image. In *SIGGRAPH Asia*, 2023. 1

[3] 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. *arXiv preprint arXiv:2511.21631*, 2025. 8

[4] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025. 20, 27

[5] Black Forest Labs. FLUX: Official inference repository for FLUX.1 models. `https://github.com/black-forest-labs/flux`, 2024. Accessed: 2025-02-07. 4, 6, 23, 24

[6] Shengqu Cai, Eric Ryan Chan, Yunzhi Zhang, Leonidas Guibas, Jiajun Wu, and Gordon. Wetzstein. Diffusion self-distillation for zero-shot customized image generation. In *CVPR*, 2025. 1

[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 7

[8] Bowen Chen, Mengyi Zhao, Haomiao Sun, Li Chen, Xu Wang, Kang Du, and Xinglong Wu. XVerse: Consistent multi-subject control of identity and semantic attributes via DiT modulation. In *NeurIPS*, 2025. 7, 23, 24, 32, 33

[9] Hong Chen, Yipeng Zhang, Simin Wu, Xin Wang, Xuguang Duan, Yuwei Zhou, and Wenwu Zhu. DisenBooth: Identity-preserving disentangled tuning for subject-driven text-to-image generation. In *ICLR*, 2024. 1

[10] Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *ECCV*, 2024. 4

[11] Liang Chen, Shuai Bai, Wenhao Chai, Weichu Xie, Haozhe Zhao, Leon Vinci, Junyang Lin, and Baobao Chang. Multimodal representation alignment for image generation: Text-image interleaved control is easier than you think. In *ICCV Findings*, 2025. 1, 7, 9, 19

[12] Wenhu Chen, Hexiang Hu, Yandong Li, Nataniel Ruiz, Xuhui Jia, Ming-Wei Chang, and William W. Cohen. Subject-driven text-to-image generation via apprenticeship learning. In *NeurIPS*, 2023. 2

[13] Yuzhuo Chen, Zehua Ma, Jianhua Wang, Kai Kang, Shunyu Yao, and Weiming Zhang. LAMIC: Layout-aware multi-image composition via scalability of multimodal diffusion transformer. In *AAAI*, 2026. 7, 23, 24

[14] Yufeng Cheng, Wenxu Wu, Shaojin Wu, Mengqi Huang, Fei Ding, and Qian He. UMO: Scaling multi-identity consistency for image customization via matching reward. In *CVPR*, 2026. 2, 5, 7, 21, 23, 28, 32, 33

[15] Yusuf Dalva, Guocheng Gordon Qian, Maya Goldenberg, Tsai-Shien Chen, Kfir Aberman, Sergey Tulyakov, Pinar Yanardag, and Kuan-Chieh Jackson Wang. Canvas-to-Image: Compositional image generation with multimodal controls. *arXiv preprint arXiv:2511.21691*, 2025. 1

[16] Yusuf Dalva, Hidir Yesiltepe, and Pinar Yanardag. LoRAShop: Training-free multi-concept image generation and editing with rectified flow transformers. In *NeurIPS*, 2025. 2

[17] Jiahua Dong, Wenqi Liang, Hongliu Li, Duzhen Zhang, Meng Cao, Henghui Ding, Salman Khan, and Fahad Khan. How to continually adapt text-to-image diffusion models for flexible customization? In *NeurIPS*, 2024. 1

[18] Ruixiao Dong, Zhendong Wang, Keli Liu, Li Li, Ying Chen, Kai Li, Daowen Li, and Houqiang Li. EchoGen: Generating visual echoes in any scene via feed-forward subject-driven auto-regressive model. In *ICLR*, 2026. 1

[19] Ziyi Dong, Pengxu Wei, and Liang Lin. DreamArtist: Towards controllable one-shot text-to-image generation via contrastive prompt-tuning. *arXiv preprint arXiv:2211.11337*, 2022. 1

[20] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024. 18

[21] Hao Fang, Zechao Zhan, Weixin Feng, Ziwei Huang, Xubin Li, and Tiezheng Ge. TBStar-Edit: From image editing pattern shifting to consistency enhancement. *arXiv preprint arXiv:2510.04483*, 2025. 2

[22] Zhoujie Fu, Xianfang Zeng, Jinghong Lan, Xinyao Liao, Cheng Chen, Junyi Chen, Jiacheng Wei, Wei Cheng, Shiyu Liu, Yunuo Chen, Gang Yu, and Guosheng Lin. iMontage: Unified, versatile, highly dynamic many-to-many image generation. *arXiv preprint arXiv:2511.20635*, 2025. 1

[23] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *ICLR*, 2023. 1, 2

[24] Rinon Gal, Moab Arar, Yuval Atzmon, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. Encoder-based domain tuning for fast personalization of text-to-image models. *ACM TOG*, 2023. 2

[25] Yuying Ge, Yixiao Ge, Ziyun Zeng, Xintao Wang, and Ying Shan. Planting a SEED of vision in large language model. *arXiv preprint arXiv:2307.08041*, 2023. 19

[26] Yuying Ge, Sijie Zhao, Ziyun Zeng, Yixiao Ge, Chen Li, Xintao Wang, and Ying Shan. Making LLaMA see and draw with seed tokenizer. In *ICLR*, 2024. 19

[27] Gemini Team. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 8

[28] Gemma Team. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025. 8

[29] Google. A new era of intelligence with Gemini 3, 2025. 8

[30] Zinan Guo, Yanze Wu, Zhuowei Chen, Lang Chen, Peng Zhang, and Qian He. PuLID: Pure and lightning ID customization via contrastive alignment. In *NeurIPS*, 2024. 2

[31] Zinan Guo, Pengze Zhang, Yanze Wu, Chong Mou, Songtao Zhao, and Qian He. MUSAR: Exploring multi-subject customization from single-subject dataset via attention routing. *arXiv preprint arXiv:2505.02823*, 2025. 2, 20, 28

[32] Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris Metaxas, and Feng Yang. SVDiff: Compact parameter space for diffusion fine-tuning. In *ICCV*, 2023. 2

[33] Runze He, Yiji Cheng, Tiankai Hang, Zhimin Li, Yu Xu, Zijin Yin, Shiyi Zhang, Wenxun Dai, Penghui Du, Ao Ma, Chunyu Wang, Qinglin Lu, Jizhong Han, and Jiao Dai. Re-Align: Structured reasoning-guided alignment for in-context image generation and editing. *arXiv preprint arXiv:2601.05124*, 2026. 2

[34] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 18

[35] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022. 2

[36] Junjie Hu, Tianyang Han, Kai Ma, Jialin Gao, Song Yang, Xianhua He, Junfeng Luo, Xiaoming Wei, and Wenqiang Zhang. PositionIC: Unified position and identity consistency for image customization. In *CVPR*, 2026. 2

[37] 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. 2

[38] Mengqi Huang, Zhendong Mao, Mingcong Liu, Qian He, and Yongdong Zhang. RealCustom: Narrowing real text word for real-time open-domain text-to-image customization. In *CVPR*, 2024. 1

[39] Ziwei Huang, Ying Shu, Hao Fang, Quanyu Long, Wenya Wang, Qiushi Guo, Tiezheng Ge, and Leilei Gan. From competition to synergy: Unlocking reinforcement learning for subject-driven image generation. *arXiv preprint arXiv:2510.18263*, 2025. 2

[40] Sangwon Jang, Jaehyeong Jo, Kimin Lee, and Sung Ju Hwang. Identity decoupling for multi-subject personalization of text-to-image models. In *NeurIPS*, 2024. 2

[41] Xuhui Jia, Yang Zhao, Kelvin C. K. Chan, Yandong Li, Han Zhang, Boqing Gong, Tingbo Hou, Huisheng Wang, and Yu-Chuan Su. Taming encoder for zero fine-tuning image customization with text-to-image diffusion models. *arXiv preprint arXiv:2304.02642*, 2023. 2

[42] Xin Jiang, Jingwen Chen, Yehao Li, Yingwei Pan, Kezhou Chen, Zechao Li, Ting Yao, and Tao Mei. DreamVAR: Taming reinforced visual autoregressive model for high-fidelity subject-driven image generation. In *ICASSP*, 2026. 19

[43] Qiaoqiao Jin, Siming Fu, Dong She, Weinan Jia, Hualiang Wang, Mu Liu, and Jidong Jiang. FocusDPO: Dynamic preference optimization for multi-subject personalized image generation via adaptive focus. In *AAAI*, 2026. 2

[44] Qiaoqiao Jin, Dong She, Hualiang Wang, Siming Fu, Mu Liu, and Jidong Jiang. Consis-GCPO: Consistency-preserving group causal preference optimization for vision customization. In *ICLR*, 2026. 2

[45] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014. 1

[46] Jing Yu Koh, Daniel Fried, and Russ R Salakhutdinov. Generating images with multimodal language models. In *NeurIPS*, 2023. 19

[47] Komal Kumar, Rao Muhammad Anwer, Fahad Shahbaz Khan, Salman Khan, Ivan Laptev, and Hisham Cholakkal. DEFT: Decompositional efficient fine-tuning for text-to-image models. In *NeurIPS*, 2025. 8, 21

[48] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *CVPR*, 2023. 1

[49] Duong H Le, Tuan Pham, Sangho Lee, Christopher Clark, Aniruddha Kembhavi, Stephan Mandt, Ranjay Krishna, and Jiasen Lu. One diffusion to generate them all. In *CVPR*, 2025. 2

[50] Dongxu Li, Junnan Li, and Steven Hoi. BLIP-Diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. In *NeurIPS*, 2023. 2, 19

[51] Guandong Li and Zhaobin Chu. EditID: Training-free editable ID customization for text-to-image generation. In *EMNLP Findings*, 2025. 2

[52] Guandong Li and Zhaobin Chu. EditIDv2: Editable ID customization with data-lubricated ID feature integration for text-to-image generation. *arXiv preprint arXiv:2509.05659*, 2025. 2

[53] Guandong Li and Yijun Ding. DVI: Disentangling semantic and visual identity for training-free personalized generation. *arXiv preprint arXiv:2512.18964*, 2025. 2

[54] Guandong Li and Mengxia Ye. Inject where it matters: Training-free spatially-adaptive identity preservation for text-to-image personalization. *arXiv preprint arXiv:2602.13994*, 2026. 2

[55] Kevin Li, Manuel Brack, Sudeep Katakol, Hareesh Ravi, and Ajinkya Kale. UniFusion: Vision-language model as unified encoder in image generation. *arXiv preprint arXiv:2510.12789*, 2025. 19

[56] Pengzhi Li, Qiang Nie, Ying Chen, Xi Jiang, Kai Wu, Yuhuan Lin, Yong Liu, Jinlong Peng, Chengjie Wang, and Feng Zheng. Tuning-free image customization with image and text guidance. In *ECCV*, 2024. 2

[57] Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. Mini-Gemini: Mining the potential of multi-modality vision language models. *arXiv preprint arXiv:2403.18814*, 2024. 19

[58] Yaowei Li, Xiaoyu Li, Zhaoyang Zhang, Yuxuan Bian, Gan Liu, Xinyuan Li, Jiale Xu, Wenbo Hu, Yating Liu, Lingen Li, Jing Cai, Yuexian Zou, Yancheng He, and Ying Shan. IC-Custom: Diverse image customization via in-context learning. *arXiv preprint arXiv:2507.01926*, 2025. 2

[59] Zhaoyang Li, Dongjun Qian, Kai Su, qishuai diao, Xiangyang Xia, Chang Liu, Wenfei Yang, Tianzhu Zhang, and Zehuan Yuan. BindWeave: Subject-consistent video generation via cross-modal integration. In *ICLR*, 2026. 2

[60] Zhong-Yu Li, Ruoyi Du, Juncheng Yan, Le Zhuo, Zhen Li, Peng Gao, Zhanyu Ma, and Ming-Ming Cheng. VisualCloze: A universal image generation framework via visual in-context learning. In *ICCV*, 2025. 2

[61] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 1

[62] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, 2024. 1

[63] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *ICLR*, 2023. 3

[64] Zhiheng Liu, Ruili Feng, Kai Zhu, Yifei Zhang, Kecheng Zheng, Yu Liu, Deli Zhao, Jingren Zhou, and Yang Cao. Cones: Concept neurons in diffusion models for customized generation. In *ICML*, 2023. 1

[65] Zhiheng Liu, Yifei Zhang, Yujun Shen, Kecheng Zheng, Kai Zhu, Ruili Feng, Yu Liu, Deli Zhao, Jingren Zhou, and Yang Cao. Customizable image synthesis with multiple subjects. In *NeurIPS*, 2023. 1

[66] Jian Ma, Junhao Liang, Chen Chen, and Haonan Lu. Subject-Diffusion: Open domain personalized text-to-image generation without test-time fine-tuning. In *SIGGRAPH*, 2024. 1

[67] Zhendong Mao, Mengqi Huang, Fei Ding, Mingcong Liu, Qian He, and Yongdong Zhang. RealCustom++: Representing images as real textual word for real-time customization. *IEEE TPAMI*, 48(2):2078–2095, 2026. 1

[68] Mistral AI. Mistral small 3.2, 2025. 8

[69] Chong Mou, Yanze Wu, Wenxu Wu, Zinan Guo, Pengze Zhang, Yufeng Cheng, Yiming Luo, Fei Ding, Shiwen Zhang, Xinghui Li, et al. DreamO: A unified framework for image customization. In *SIGGRAPH Asia*, 2025. 2, 7, 21, 23, 28, 32, 33

[70] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. In *ICML*, 2022. 18

[71] OpenAI. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 8

[72] Ziheng Ouyang, Yiren Song, Yaoli Liu, Shihao Zhu, Qibin Hou, Ming-Ming Cheng, and Mike Zheng Shou. The consistency critic: Correcting inconsistencies in generated images via reference-guided attentive alignment. In *CVPR*, 2026. 1

[73] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023. 4, 19

[74] Yuang Peng, Yuxin Cui, Haomiao Tang, Zekun Qi, Runpei Dong, Jing Bai, Chunrui Han, Zheng Ge, Xiangyu Zhang, and Shu-Tao Xia. DreamBench++: A human-aligned benchmark for personalized image generation. In *ICLR*, 2025. 8, 21, 24

[75] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *ICLR*, 2024. 18

[76] Senthil Purushwalkam, Akash Gokul, Shafiq Joty, and Nikhil Naik. BootPIG: Bootstrapping zero-shot personalized image generation capabilities in pretrained diffusion models. In *ECCV Workshops*, 2024. 1

[77] Guocheng Gordon Qian, Ruihang Zhang, Tsai-Shien Chen, Yusuf Dalva, Anujraaj Argo Goyal, Willi Menapace, Ivan Skorokhodov, Meng Dong, Arpit Sahni, Daniil Ostashev, Ju Hu, Sergey Tulyakov, and Kuan-Chieh Jackson Wang. LayerComposer: Multi-human personalized generation via layered canvas. *arXiv preprint arXiv:2510.20820*, 2025. 2

[78] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 7

[79] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 2020. 24

[80] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 18

[81] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 18

[82] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *MICCAI*, 2015. 18

[83] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, 2023. 1, 2, 7, 21, 24

[84] Oindrila Saha, Vojtech Krs, Radomir Mech, Subhransu Maji, Kevin James Blackburn-Matzen, and Matheus Gadelha. SIGMA-Gen: Structure and identity guided multi-subject assembly for image generation. In *ICLR*, 2026. 2

[85] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022. 18

[86] Dong She, Siming Fu, Mushui Liu, Qiaoqiao Jin, Hualiang Wang, Mu Liu, and Jidong Jiang. MOSAIC: Multi-subject personalized generation via correspondence-aware alignment and disentanglement. In *ICLR*, 2026. 2

[87] Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. InstantBooth: Personalized text-to-image generation without test-time finetuning. In *CVPR*, 2024. 1

[88] Xuanke Shi, Boxuan Li, Xiaoyang Han, Zhongang Cai, Lei Yang, Dahua Lin, and Quan Wang. ConsistCompose: Unified multimodal layout control for image composition. *arXiv preprint arXiv:2511.18333*, 2025. 2

[89] Aditi Singhania, Arushi Jain, Krutik Malani, Riddhi Dhawan, Souymodip Chakraborty, Vineet Batra, and Ankit Phogat. Taming identity consistency and prompt diversity in diffusion models via latent concatenation and masked conditional flow matching. *arXiv preprint arXiv:2511.08061*, 2025. 1

[90] Aditi Singhania, Krutik Malani, Riddhi Dhawan, Arushi Jain, Garv Tandon, Nippun Sharma, Souymodip Chakraborty, Vineet Batra, and Ankit Phogat. Beyond the Pixels: VLM-based evaluation of identity preservation in reference-guided synthesis. *arXiv preprint arXiv:2511.08087*, 2025. 1

[91] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015. 18

[92] Xinyang Song, Libin Wang, Weining Wang, Zhiwei Li, Jianxin Sun, Dandan Zheng, Jingdong Chen, Qi Li, and Zhenan Sun. 3SGen: Unified subject, style, and structure-driven image generation with adaptive task-specific memory. *arXiv preprint arXiv:2512.19271*, 2025. 2

[93] Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Emu: Generative pretraining in multimodality. In *ICLR*, 2024. 19

[94] Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. OminiControl: Minimal and universal control for diffusion transformer. In *ICCV*, 2025. 7, 23, 32, 33

[95] Jiale Tao, Yanbing Zhang, Qixun Wang, Yiji Cheng, Haofan Wang, Xu Bai, Zhengguang Zhou, Ruihuang Li, Linqing Wang, Chunyu Wang, et al. InstantCharacter: Personalize any characters with a scalable diffusion transformer framework. *arXiv preprint arXiv:2504.12395*, 2025. 2

[96] Gemma Canet Tarrés, Manel Baradad, Francesc Moreno-Noguer, and Yumeng Li. PLACID: Identity-preserving multi-object compositing via video diffusion with synthetic trajectories. *arXiv preprint arXiv:2602.00267*, 2026. 2

[97] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 18

[98] Andrey Voynov, Qinghao Chu, Daniel Cohen-Or, and Kfir Aberman. P+: Extended textual conditioning in text-to-image generation. *arXiv preprint arXiv:2303.09522*, 2023. 1

[99] Shulei Wang, Longhui Wei, Xin He, Jianbo Ouyang, Hui Lu, Zhou Zhao, and Qi Tian. PSR: Scaling multi-subject personalized image generation with pairwise subject-consistency rewards. *arXiv preprint arXiv:2512.01236*, 2025. 2

[100] Xierui Wang, Siming Fu, Qihan Huang, Wanggui He, and Hao Jiang. MS-Diffusion: Multi-subject zero-shot image personalization with layout guidance. In *ICLR*, 2025. 2

[101] Yuran Wang, Bohan Zeng, Chengzhuo Tong, Wenxuan Liu, Yang Shi, Xiaochen Ma, Hao Liang, Yuanxing Zhang, and Wentao Zhang. Scone: Bridging composition and distinction in subject-driven image generation via unified understanding-generation modeling. *arXiv preprint arXiv:2512.12675*, 2025. 2

[102] Zhizhong Wang, Tianyi Chu, Zeyi Huang, Nanyang Wang, and Kehan Li. DynaIP: Dynamic image prompt adapter for scalable zero-shot personalized text-to-image generation. *arXiv preprint arXiv:2512.09814*, 2025. 2

[103] Zhongsheng Wang, Ming Lin, Zhedong Lin, Yaser Shakib, Qian Liu, and Jiamou Liu. CharCom: Composable identity control for multi-character story illustration. In *ACM Multimedia Asia*, 2025. 2

[104] Zehan Wang, Ziang Zhang, Tianyu Pang, Chao Du, Hengshuang Zhao, and Zhou Zhao. Orient Anything: Learning robust object orientation estimation from rendering 3D models. In *ICML*, 2025. 7

[105] Hongyang Wei, Bin Wen, Yancheng Long, Yankai Yang, Yuhang Hu, Tianke Zhang, Wei Chen, Haonan Fan, Kaiyu Jiang, Jiankang Chen, Changyi Liu, Kaiyu Tang, Haojie Ding, Xiao Yang, Jia Sun, Huaiqing Wang, Zhenyu Yang, Xinyu Wei, Xianglong He, Yangguang Li, Fan Yang, Tingting Gao, Lei Zhang, Guorui Zhou, and Han Li. UniRef-Image-Edit: Towards scalable and consistent multi-reference image editing. *arXiv preprint arXiv:2602.14186*, 2026. 2

[106] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. ELITE: Encoding visual concepts into textual embeddings for customized text-to-image generation. In *ICCV*, 2023. 1

[107] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-Image technical report. *arXiv preprint arXiv:2508.02324*, 2025. 1, 4, 7, 9, 19, 23, 32, 33

[108] Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, Ze Liu, Ziyi Xia, Chaofan Li, Haoge Deng, Jiahao Wang, Kun Luo, Bo Zhang, Defu Lian, Xinlong Wang, Zhongyuan Wang, Tiejun Huang, and Zheng Liu. OmniGen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025. 7, 23, 32, 33

[109] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. NExT-GPT: Any-to-any multimodal LLM. In *ICML*, 2024. 19

[110] Shaojin Wu, Mengqi Huang, Wenxu Wu, Yufeng Cheng, Fei Ding, and Qian He. Less-to-more generalization: Unlocking more controllability by in-context generation. In *ICCV*, 2025. 1, 2, 5, 6, 7, 20, 21, 23, 28, 32, 33

[111] Shaojin Wu, Mengqi Huang, Yufeng Cheng, Wenxu Wu, Jiahe Tian, Yiming Luo, Fei Ding, and Qian He. USO: Unified style and subject-driven generation via disentangled and reward learning. In *CVPR*, 2026. 2, 7, 23, 32, 33

[112] Tao Wu, Yibo Jiang, Yehao Lu, Zhizhong Wang, Zeyi Huang, Zequn Qin, and Xi Li. MultiCrafter: High-fidelity multi-subject generation via disentangled attention and identity-aware preference alignment. In *CVPR*, 2026. 2

[113] Bin Xia, Bohao Peng, Yuechen Zhang, Junjia Huang, Jiyang Liu, Jingyao Li, Haoru Tan, Sitong Wu, Chengyao Wang, Yitong Wang, Xinglong Wu, Bei Yu, and Jiaya Jia. DreamOmni2: Multimodal instruction-based editing and generation. *arXiv preprint arXiv:2510.06679*, 2025. 2

[114] Bin Xia, Yuechen Zhang, Jingyao Li, Chengyao Wang, Yitong Wang, Xinglong Wu, Bei Yu, and Jiaya Jia. DreamOmni: Unified image generation and editing. In *CVPR*, 2025.

[115] Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. OmniGen: Unified image generation. In *CVPR*, 2025. 2

[116] Jiazheng Xing, Fei Du, Hangjie Yuan, Pengwei Liu, Hongbin Xu, Hai Ci, Ruigang Niu, Weihua Chen, Fan Wang, and Yong Liu. LumosX: Relate any identities with their attributes for personalized video generation. In *ICLR*, 2026. 2

[117] Hengyuan Xu, Wei Cheng, Peng Xing, Yixiao Fang, Shuhan Wu, Rui Wang, Xianfang Zeng, Daxin Jiang, Gang YU, Xingjun Ma, and Yu-Gang Jiang. WithAnyone: Toward controllable and ID consistent image generation. In *ICLR*, 2026. 2

[118] Ruihang Xu, Dewei Zhou, Fan Ma, and Yi Yang. ContextGen: Contextual layout anchoring for identity-consistent multi-instance generation. In *ICLR*, 2026. 2

[119] Yijia Xu, Zihao Wang, and Jinshi Cui. Hierarchical concept-to-appearance guidance for multi-subject image generation. *arXiv preprint arXiv:2602.03448*, 2026.

[120] Hongji Yang, Yucheng Zhou, Wencheng Han, Runzhou Tao, Zhongying Qiu, Jianfei Yang, and Jianbing Shen. HiCoGen: Hierarchical compositional text-to-image generation in diffusion models via reinforcement learning. *arXiv preprint arXiv:2511.19965*, 2025. 2

[121] Yixiong Yang, Tao Wu, Senmao Li, Shiqi Yang, Yaxing Wang, Joost van de Weijer, and Kai Wang. EchoDistill: Bidirectional concept distillation for one-step diffusion personalization. *arXiv preprint arXiv:2510.20512*, 2025. 2

[122] Zebin Yao, Lei Ren, Huixing Jiang, Chen Wei, Xiaojie Wang, Ruifan Li, and Fangxiang Feng. FreeGraftor: Training-free cross-image feature grafting for subject-driven text-to-image generation. *arXiv preprint arXiv:2504.15958*, 2025. 2

[123] 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. 1, 2

[124] Zixuan Ye, Quande Liu, Cong Wei, Yuanxing Zhang, Xintao Wang, Pengfei Wan, Kun Gai, and Wenhan Luo. Visual-Aware CoT: Achieving high-fidelity visual consistency in unified models. In *CVPR*, 2026. 1

[125] Jun Zhan, Junqi Dai, Jiasheng Ye, Yunhua Zhou, Dong Zhang, Zhigeng Liu, Xin Zhang, Ruibin Yuan, Ge Zhang, Linyang Li, et al. AnyGPT: Unified multimodal LLM with discrete sequence modeling. In *ACL*, 2024. 19

[126] Hui Zhang, Dexiang Hong, Maoke Yang, Yutao Cheng, Zhao Zhang, Weidong Chen, Jie Shao, Xinglong Wu, Zuxuan Wu, and Yu-Gang Jiang. CreatiDesign: A unified multi-conditional diffusion transformer for creative graphic design. In *ICLR*, 2026. 2

[127] Yuxuan Zhang, Yiren Song, Jiaming Liu, Rui Wang, Jinpeng Yu, Hao Tang, Huaxia Li, Xu Tang, Yao Hu, Han Pan, and Zhongliang Jing. SSR-Encoder: Encoding selective subject representation for subject-driven generation. In *CVPR*, 2024. 1

[128] Peng Zheng, Ye Wang, Rui Ma, and Zuxuan Wu. FreeLoRA: Enabling training-free LoRA fusion for autoregressive multi-subject personalization. *arXiv preprint arXiv:2507.01792*, 2025. 2

[129] Shuhong Zheng, Ashkan Mirzaei, and Igor Gilitschenski. Track, Inpaint, Resplat: Subject-driven 3D and 4D generation with progressive texture infilling. In *NeurIPS*, 2025. 8, 21

[130] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. 6, 20, 23, 27

[131] Zhuofan Zong, Dongzhi Jiang, Bingqi Ma, Guanglu Song, Hao Shao, Dazhong Shen, Yu Liu, and Hongsheng Li. EasyRef: Omni-generalized group image reference for diffusion models via multimodal LLM. In *ICML*, 2025. 1, 4, 7, 19, 23, 32, 33

Squeezing Capacity from Multimodal Large Language Models for Subject-driven Generation

Technical Appendices and Supplementary Material

In this appendix, we provide additional analyses and experiments to further validate the effectiveness of our method along with more discussions. First, in Section A, we conduct a layer-wise analysis of the DLA module during inference to understand its contribution across different layers. Section B explores various layer selection strategies for training DLA, analyzing both efficiency and performance trade-offs. Section C discusses on extended related works including the development of text-to-image generation, and existing approaches that bridge large multimodal models and diffusion models. Section D conducts the sensitivity analysis to demonstrate the robustness of hyperparameter choices and flexible user control from our method. In Section E, we perform an ablation study on different MLLM backbones, demonstrating the robustness and generalizability of our approach. We further show the adaptability of our method to multi-subject generation in Section F with small-scale finetuning. More implementation details about the construction of the multimodal reasoning benchmark, the instruction and interface used for the user study, and the prompts used for MLLM-based evaluation are explained in Sections G, H, I, respectively. More quantitative and qualitative evaluations are displayed in Sections J and K to highlight the visual advantages of our approach. Section L include the license and terms of use for the models and data used in the paper. Finally, Sections M and N present the limitations, discussions, and potential societal impact of our method.

A Layer Analysis for DLA at Inference Time

In this section, we analyze how different layers in the dual text LAP and image LAP of our DLA contribute to performance. Specifically, we take the fully trained model and selectively zero out certain layers during inference to examine their impact. The results in Table A reveal three key observations. First, the image modality is highly sensitive to the removal of the early MLLM layers, suggesting that these layers are essential for preserving fine-grained visual details. Second, the text modality shows more robustness to layer removal. This indicates that, although the text modality primarily relies on the later layers, the model can still retrieve similar textual information from the preceding layers when later layers are removed. Third, we find that disabling one modality can occasionally bring slight improvements when the other modality is partially dropped, implying that the model may rely more heavily on a single modality in such cases. This further supports our claim that balancing the two modalities is crucial for optimal performance.

We further include qualitative comparisons in Figure B and Figure C, which visually support the observations drawn from Table A. Specifically, we observe two consistent trends. First, when early layers of the image modality are dropped (e.g., zeroing out layers 0-19), the model struggles to preserve identity, whereas using only early layers achieves comparable ID consistency. Second, for the text modality, the later layers appear more critical as using only layers 0-9 significantly weakens the model's ability to understand and follow the prompt.

B Layer Selection for Training DLA

In the main paper, we primarily ablate layer selection strategies for the single LAP. Here, we extend the investigation with a more comprehensive analysis of dual LAP layer strategies within our DLA module. It is important to note the key difference between this experiment and the previous one in Table A. In Section A, the study was conducted during inference using a fully trained model where individual layers were selectively zeroed out. In contrast, the experiments presented here involve training the entire pipeline from scratch while using only a subset of MLLM layers for either the text LAP or the image LAP.

Thus, the analysis in Section A emphasizes the contribution of each individual group of layers within DLA, whereas this section focuses on evaluating different strategies for connecting MLLM features to DiT. The results are shown in Table B, and our key observations are summarized as follows. First, pre-selecting early layers (0-9) yields noticeable performance gains for identity metrics, likely because the model leans more toward copy-paste behavior, as text-following performance drops.

Table A: Layer analysis of our DLA during inference. We assess the contribution of each MLLM layer in our full DLA by selectively zeroing out text and image modalities during inference. For simplicity, the numerical results of difference are rounded to two digits after the decimal point.

Image Layers	Text Layers	DINO-I (↑)	CLIP-I (↑)	CLIP-T (↑)
0-28	0-28	0.7482	0.8443	0.3010
~~0-9~~ 10-19 20-28	0-28	0.6368 (-0.11)	0.7959 (-0.05)	0.3111 (+0.01)
0-9 ~~10-19~~ 20-28	0-28	0.7472 (-0.00)	0.8439 (-0.00)	0.3011 (+0.00)
0-9 10-19 ~~20-28~~	0-28	0.7344 (-0.01)	0.8353 (-0.01)	0.3041 (+0.00)
~~0-9~~ 10-19 ~~20-28~~	0-28	0.6058 (-0.14)	0.7837 (-0.06)	0.3117 (+0.01)
0-9 ~~10-19~~ ~~20-28~~	0-28	0.7093 (-0.04)	0.8251 (-0.02)	0.3067 (+0.01)
~~0-9~~ ~~10-19~~ 20-28	0-28	0.5898 (-0.16)	0.7773 (-0.07)	0.3129 (+0.01)
~~0-12~~ 13-16 ~~17-28~~	0-28	0.5962 (-0.15)	0.7769 (-0.07)	0.3134 (+0.01)
~~0-24~~ 25-28	0-28	0.5560 (-0.19)	0.7618 (-0.08)	0.3129 (+0.01)
0-3 ~~4-28~~	0-28	0.5493 (-0.20)	0.7609 (-0.08)	0.3126 (+0.01)
0-28	~~0-9~~ 10-19 20-28	0.7557 (+0.01)	0.8474 (+0.00)	0.3006 (-0.00)
0-28	0-9 ~~10-19~~ 20-28	0.7399 (-0.01)	0.8405 (-0.00)	0.2990 (-0.00)
0-28	0-9 10-19 ~~20-28~~	0.7267 (-0.02)	0.8354 (-0.01)	0.2991 (-0.00)
0-28	~~0-9~~ ~~10-19~~ 20-28	0.7560 (+0.01)	0.8498 (+0.01)	0.2966 (-0.00)
0-28	~~0-9~~ 10-19 ~~20-28~~	0.7363 (-0.01)	0.8402 (-0.00)	0.3018 (+0.00)
0-28	0-9 ~~10-19~~ ~~20-28~~	0.7473 (-0.00)	0.8492 (+0.00)	0.2545 (-0.05)
0-28	~~0-12~~ 13-16 ~~17-28~~	0.7661 (+0.02)	0.8631 (+0.02)	0.2657 (-0.04)
0-28	~~0-24~~ 25-28	0.8274 (+0.08)	0.9068 (+0.06)	0.2480 (-0.05)
0-28	0-3 ~~4-28~~	0.7823 (+0.03)	0.8724 (+0.03)	0.2590 (-0.04)
0-9 ~~10-19~~ ~~20-28~~	0-9 ~~10-19~~ ~~20-28~~	0.6950 (-0.05)	0.8153 (-0.03)	0.2586 (-0.04)
0-9 ~~10-19~~ ~~20-28~~	~~0-9~~ 10-19 ~~20-28~~	0.6906 (-0.06)	0.8173 (-0.03)	0.3080 (+0.01)
0-9 ~~10-19~~ ~~20-28~~	~~0-9~~ ~~10-19~~ 20-28	0.7135 (-0.03)	0.8301 (-0.01)	0.3042 (+0.00)
~~0-9~~ 10-19 ~~20-28~~	0-9 ~~10-19~~ ~~20-28~~	0.4852 (-0.26)	0.7129 (-0.13)	0.2582 (-0.04)
~~0-9~~ 10-19 ~~20-28~~	~~0-9~~ 10-19 ~~20-28~~	0.5743 (-0.17)	0.7700 (-0.07)	0.3135 (+0.01)
~~0-9~~ 10-19 ~~20-28~~	~~0-9~~ ~~10-19~~ 20-28	0.6139 (-0.13)	0.7878 (-0.06)	0.3085 (+0.01)
~~0-9~~ ~~10-19~~ 20-28	0-9 ~~10-19~~ ~~20-28~~	0.4144 (-0.33)	0.6820 (-0.16)	0.2500 (-0.05)
~~0-9~~ ~~10-19~~ 20-28	~~0-9~~ 10-19 ~~20-28~~	0.5562 (-0.19)	0.7644 (-0.08)	0.3139 (+0.01)
~~0-9~~ ~~10-19~~ 20-28	~~0-9~~ ~~10-19~~ 20-28	0.6038 (-0.14)	0.7848 (-0.06)	0.3099 (+0.01)

Second, almost all layer-preselection strategies for the text modality lead to degraded performance, aligning with the finding from Section A that textual information is distributed across all MLLM layers. Third, despite using only subset layers, the diffusion model can still attain comparable or even improved performance for both modalities, suggesting that the current layer aggregation design may not fully exploit the representational efficiency of different layers. Some layers may be redundant while others are more informative, potentially depending on the specific context.

We present a qualitative comparison of different layer selection strategies for the DLA module in Figure D. The figure shows 16 combinations of layers for the text and image modalities, including ranges 0-9, 10-19, 20-28, and all layers (0-28). Each row corresponds to the text modality layer setting, while each column represents the image modality layer setting. Importantly, each combination represents a separately re-trained model, allowing us to isolate the effect of specific layer selections on the final generation. Compared with the inference-time zero-out analysis in Figure B and Figure C, we observe that pre-selecting layers during training can lead to serious relaying on identity image and worse text-following capability as shown in Figure D, especially when not using all 28 layers for the text modality. This visualization highlights the trade-offs between constraining layers for efficiency and maintaining balanced identity preservation and multimodal understanding.

C Additional Related Work

Text-to-image (T2I) generation has rapidly advanced in recent years, with successful systems adopting denoising diffusion frameworks [34, 91]. Early studies validated diffusion models for T2I and demonstrated advantages over GAN and autoregressive-based approaches [70, 80, 85]. Latent diffusion—training the diffusion process in a compact latent space—proved especially effective at efficiency improvement and output resolution, and has become a *de facto* standard in large-scale T2I systems (*e.g.*, LDM and the Stable Diffusion family) [20, 75, 81]. Recent works replace the traditional UNet [82] backbone with transformer-based [97] decoders, *e.g.,* Diffusion Transformer

Table B: Layer selection of our DLA during training. We re-train each of our variants of DLA by selecting different parts of layers for text and image modalities.

Selected Image LAP Layers	Selected Text LAP Layers	DINO-I (↑)	CLIP-I (↑)	CLIP-T (↑)
0-28	0-28	0.7482	0.8443	0.3010
0-9	0-28	0.7781 (+0.03)	0.8567 (+0.01)	0.2932 (-0.01)
10-19	0-28	0.7519 (+0.00)	0.8424 (-0.00)	0.2972 (-0.00)
20-28	0-28	0.7189 (-0.03)	0.8292 (-0.02)	0.2990 (-0.00)
0-28	0-9	0.7464 (-0.00)	0.8439 (-0.00)	0.2960 (-0.01)
0-28	10-19	0.7493 (+0.00)	0.8464 (+0.00)	0.2969 (-0.00)
0-28	20-28	0.7530 (+0.00)	0.8473 (+0.00)	0.2984 (-0.00)
0-9	0-9	0.7730 (+0.02)	0.8522 (+0.01)	0.2840 (-0.02)
0-9	10-19	0.7620 (+0.01)	0.8517 (+0.01)	0.2925 (-0.01)
0-9	20-28	0.7788 (+0.03)	0.8584 (+0.01)	0.2888 (-0.01)
10-19	0-9	0.7520 (+0.00)	0.8386 (-0.01)	0.2865 (-0.01)
10-19	10-19	0.7466 (-0.00)	0.8426 (-0.00)	0.2936 (-0.01)
10-19	20-28	0.7025 (-0.05)	0.8261 (-0.02)	0.2992 (-0.00)
20-28	0-9	0.7327 (-0.02)	0.8298 (-0.01)	0.2864 (-0.01)
20-28	10-19	0.7515 (+0.00)	0.8534 (+0.01)	0.2919 (-0.01)
20-28	20-28	0.6742 (-0.07)	0.8200 (-0.02)	0.3030 (+0.00)

(DiT) architectures [73], showing significant gains in image quality and scalability. These transformer decoders can better model long-range structure and complex compositions, which is central to our goal of conditioning high-fidelity generation on rich multimodal embeddings.

Bridging large multimodal models and diffusion decoders. Recently, integrating large language and multimodal models (LMMs/MLLMs) with diffusion decoders has attracted growing interest, enabling rich, structured, and interleaved text–image generation. Some approaches [57] translate complex multimodal instructions into textual or latent control codes that diffusion models can directly consume. Others introduce discrete or continuous visual tokenizers (*e.g.*, Seed-Tokenizer [25], Seed-LLaMA [26]) that encode compact visual semantics to align language and vision token spaces for diffusion decoding. Jointly trained systems [42] such as GILL [46], Emu [93], NExT-GPT [109], and Any-GPT [125] further strengthen semantic alignment between multimodal embeddings and diffusion backbones. Methods like BLIP-Diffusion [50] extend this idea by projecting unified image–text representations into diffusion conditioning spaces to handle complex interleaved prompts. More recent pipelines—including UniFusion [55], DreamEngine [11], Qwen-Image [107], and EasyRef [131]—leverage pretrained MLLM or VLM features as conditioning signals for downstream diffusion transformers, enabling flexible text–image interleaving. However, these approaches typically rely only on the final-layer features of the MLLMs (*e.g.,* Qwen-Image), or blend the ViT features from MLLMs with final-layer outputs via simple scalar mixing (*e.g.,* DreamEngine). As a result, they often overlook fine-grained visual cues without relying on ID-relevant features (*e.g.,* VAE), or provide only suboptimal identity preservation in subject-driven generation.

D Additional Sensitivity Analysis

Our framework, which leverages both MLLM and VAE for identity preservation, supports a multi-stage, timestep-aware denoising process: early steps rely on MLLM features for high-level reasoning, while later steps use VAE features for fine-grained detail. We ablate different configurations of this process to guide users in selecting denoising thresholds (τ_1, τ_2) that balance identity fidelity and pose variation. As shown in Figure A, higher thresholds (b) improve identity preservation but reduce pose diversity, whereas lower thresholds (c) allow more creative poses, but with slightly compromised identity. Sample (d) illustrates that extreme CFG values can degrade image quality. This design offers users flexibility to control the trade-off between subject fidelity and creativity, and Table C shows that the overall performance remains robust across a range of parameter choices.

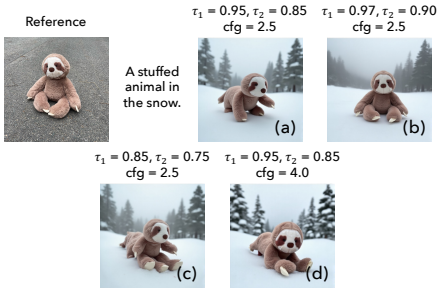

Figure A: Ablations on different configurations of the timestep-aware denoising process. The thresholds τ_1 and τ_2, which partition the denoising stages, along with the CFG value, can be adjusted freely by users to balance high-fidelity identity preservation with diverse pose generation.

Table C: Quantitative results for different multi-stage denoising configurations, with the gray row indicating our current parameter choice. The thresholds τ_1 and τ_2 along with the CFG value control the trade-off between identity preservation and text following, while the model remains robust across a range of parameter settings.

τ_1	τ_2	CFG	DINO-I (↑)	CLIP-I (↑)	CLIP-T (↑)
0.00	0.00	2.5	0.6905	0.8225	**0.3044**
1.00	1.00	2.5	0.7351	0.8462	0.2554
0.97	0.90	2.5	**0.7490**	**0.8466**	0.2963
0.85	0.70	2.5	0.7282	0.8376	0.3034
0.95	0.85	1.5	0.7268	0.8385	0.2990
0.95	0.85	2.0	0.7418	0.8430	0.3004
0.95	0.85	2.5	0.7482	0.8443	0.3010
0.95	0.85	3.0	0.7481	0.8428	0.3008
0.95	0.85	4.0	0.7431	0.8381	0.3012

Table D: Ablation on different MLLM backbones. We compare InternVL3-8B—used as our main model—with alternative architectures of varying sizes, including InternVL3-2B, Qwen2.5-VL-3B, and Qwen2.5-VL-7B.

Model	DINO-I (↑)	CLIP-I (↑)	CLIP-T (↑)
InternVL3-8B	0.7482	0.8443	0.3010
InternVL3-2B	0.7415	0.8380	0.2987
Qwen2.5-VL-3B	0.7194	0.8300	0.3027
Qwen2.5-VL-7B	0.7282	0.8241	0.3031

E Ablation on Different MLLM Selection

For our main evaluation, we use InternVL3-8B [130] as it provides a good balance between model capacity and efficiency. To study the impact of different MLLM backbones, we further ablate a range of alternatives with varying sizes and architectures, including InternVL3-2B [130], Qwen2.5-VL-3B [4], and Qwen2.5-VL-7B [4]. As shown in Table D, all models follow a similar performance trend across the metrics, with only minor variations. Qwen2.5-VL exhibits slightly weaker visual context understanding but marginally better text alignment. Overall, the difference is not substantial. Notably, InternVL3-2B achieves comparable results to the 8B model while using significantly fewer parameters, offering a promising lightweight alternative.

We present a qualitative comparison of different MLLM backbones in Figure E. Although these models differ in type and parameter size, the results do not reveal any major visual differences across the backbones. All four models, including Qwen2.5-VL-3B, Qwen2.5-VL-7B, InternVL3-2B, and InternVL3-8B, produce results with comparable image fidelity, identity preservation, and text-following ability. While larger models show slightly stronger grounding and semantic alignment, the overall performance trend remains consistent, indicating that our DLA framework generalizes well across different MLLM architectures. This further supports the conclusion from the quantitative analysis that the choice of backbone has limited impact on the final generation quality.

F Adaptation to Multi-subject Generation

In the main paper, we primarily focus on single-subject generation. We intentionally focus on single-subject training for two reasons: (1) our primary goal is to explore and analyze how to optimally leverage MLLM features for subject-driven generation; and (2) high-quality multi-subject datasets are often private and difficult to obtain at scale. However, our framework can also be extended to handle multi-subject generation with minimal adaptation. Hence, we fine-tune the model using the public two-subject dataset MUSAR-Gen [31], which contains fewer than 30K image pairs. During training, after completing the MLLM-only stage on UNO-1M [110] for single-subject learning, we continue to fine-tune the model on MUSAR-Gen—still within the MLLM-only framework. In the subsequent

stage involving both MLLM and VAE, we jointly train on a mixture of UNO-1M and MUSAR-Gen to establish the full multi-subject pipeline. We compare the resulting multi-subject model against UNO [110], DreamO [69], and UMO [14]. As shown in Figure F, our method achieves superior results on the multi-subject DreamBench [83] samples, excelling in both identity preservation and text compliance.

G Multimodal Reasoning Benchmark Construction

In the main paper, to quantify the multimodal reasoning capability, we propose to evaluate on a constructed benchmark consisting of complex prompts that require the model to perform concept binding and cross-modal reasoning. The key idea for constructing this benchmark is to collect images where a primary subject appears together with additional visible objects that function as accessories. The corresponding text prompts, however, deliberately refer to a non-salient object in the image rather than the main subject. Under this setting, the models cannot simply presume that the most salient object in the reference image is the subject whose identity needs to be preserved. Instead, they are expected to reason about the prompt and correctly locate the concept mentioned in the text within the image.

To curate such images, we collect generated samples from state-of-the-art subject-driven methods like USO on DreamBench, and manually verify their content and quality. The associated prompts are then modified by replacing the subject category in the original prompts. For example, the prompt "A *cat* wearing a shirt" can be changed to "An *elephant* wearing a shirt". Furthermore, we construct variants that include two reference images. Each sample is formed by combining a generated composite image, an original reference image from DreamBench, and a modified prompt in which the subject category matches that of the selected DreamBench image. Samples in the curated benchmark can be seen in Figure G. In total, the benchmark contains 170 single-reference samples and 180 two-reference samples, resulting in 350 test samples overall.

H User Study

Below, we provide further details of the user study setup. Participants are given the following instructions at the beginning of the study.

In subject-driven generation, the user gives a reference image along with a text prompt, and the goal is to generate an image that ***aligns with the text description while preserving the identity in the reference image****.*

For each of the following cases, we use 5 different methods to perform subject-driven generation. We would like to invite you to give an overall score from 1 (worst quality) to 10 (best quality) to measure the quality of the generated results.

There are a few points to consider when providing the scores:

- *Whether the generated images preserve the identity of the reference image*
- *Whether the generated images follow the text prompt*
- *The visual quality of the generated images (e.g., whether they look realistic, whether they follow the physical rules)*

We also include screenshots of the user study interface in Figures I and J.

I MLLM-based Evaluation Details

We follow the concept preservation evaluation protocol in DreamBench++ [74] and its follow-ups [47, 129] to construct prompts for MLLM-based scoring of identity preservation between the reference image and the generated image. The prompts are provided below.

Task Definition

You will be provided with an image generated based on reference image.

As an experienced evaluator, your task is to evaluate the semantic consistency between the subject of the generated image and the reference image, according to the scoring criteria.

Scoring Criteria

It is often compared whether two subjects are consistent based on four basic visual features:

1. *Shape: Evaluate whether the main body outline, structure, and proportions of the generated image match those of the reference image. This includes the geometric shape of the main body, clarity of edges, relative sizes, and spatial relationships between various parts composing the main body.*
2. *Color: Comparing the accuracy and consistency of the main colors generated in the image with those of the reference image. This includes saturation, hue, brightness, and whether the distribution of colors is similar to that of the subject in the reference image.*
3. *Texture: Focus on the local parts of the RGB image, whether the generated image effectively captures fine details without appearing blurry, and whether it possesses the required realism, clarity, and aesthetic appeal. Please note that unless specifically mentioned in the text prompt, excessive abstraction and formalization of texture are not necessary.*
4. *Facial Features: If the evaluation is of a person or animal, facial features will greatly affect the judgment of image consistency, and you also need to focus on judging whether the facial area looks very similar visually.*

Scoring Range

You need to give a specific integer score based on the comprehensive performance of the visual features above, ranging from 0 to 4:

- *Very Poor (0): No resemblance. The generated image's subject has no relation to the reference.*
- *Poor (1): Minimal resemblance. The subject falls within the same broad category but differs significantly.*
- *Fair (2): Moderate resemblance. The subject shows likeness to the reference with notable variances.*
- *Good (3): Strong resemblance. The subject closely matches the reference with only minor discrepancies.*
- *Excellent (4): Near-identical. The subject of the generated image is virtually indistinguishable from the reference.*

Input Format

Every time you will receive two images, the first image is a reference image, and the second image is the generated image.

Please carefully review each image of the subject.

Output Format

Score: [Your Score]

You must adhere to the specified output format, which means that only the scores need to be output, excluding your analysis process.

Table E: Quantitative results on additional benchmarks of XVerseBench [8] and LAMICBench [13], both on the IP-Sim metric, where higher value means better performance.

Dataset	DreamO	UNO	USO	Ours
XVerseBench	76.08	**80.36**	78.90	<u>79.10</u>
LAMICBench (two-subject)	<u>65.25</u>	64.93	Not Applicable	**66.46**

J Evaluation on More Benchmarks

We include evaluation on additional benchmarks of XVerseBench [8] and LAMICBench [13]. Specifically, we conduct experiments on the single-subject set of XVerseBench and the two-reference subset of LAMICBench with a slightly finetuned version of our model trained on two-subject data as described in Section F, both on the subject categories. The results, shown in Table E, demonstrate the decent performance of our method. Please note that we refrained from including human face and identity in our evaluation because our model is not trained on human-related data due to ethical considerations.

K More Qualitative Results

K.1 Stress Testing

To further evaluate the robustness of our method under more challenging conditions, we provide additional stress test samples involving complex instructions. In particular, we consider two scenarios: (1) prompts containing multiple instances of the subject, and (2) attribute binding in long context prompts. These settings require the model to correctly preserve subject identity while simultaneously satisfying multiple constraints specified in the text. As shown in Figure H, our method has the ability of handling multiple instances and bind the concepts in complex prompts, showing the robustness and the strong reasoning capability within our MLLM-DiT system.

K.2 Additional Qualitative Comparisons

In this section, we provide additional qualitative comparisons with state-of-the-art methods, supplementing the limited space in the main paper. As shown in Figure K and Figure L, we compare our approach with XVerse [8], EasyRef [131], DreamO [69], UMO [14], OminiControl [94], UNO [110], Qwen-Image [107], OmniGen2 [108], and USO [111]. Our approach achieves competitive subject identity with diverse subject pose variations, alleviating the copy-paste issue from other VAE-based models.

L Licenses for Existing Assets

The following list contains licenses for data and model used in the paper:

- Flux [5]: Apache License 2.0
  https://github.com/black-forest-labs/flux
- InternVL-3 [130]: MIT License
  https://github.com/opengvlab/internvl
- UNO-1M [110]: Apache License 2.0
  https://github.com/bytedance/UNO

- DreamBench [83]: CC-BY-4.0 License
  https://github.com/google/dreambooth
- XVerseBench [8]: Apache License 2.0
  https://github.com/bytedance/XVerse
- LAMICBench [13]: Apache License 2.0
  https://github.com/Suchenl/LAMIC
- DreamBench++ [74]: Apache License 2.0
  https://github.com/yuangpeng/dreambench_plus

M Discussions and Limitations

One of the limitations of the current framework lies in the alignment between the MLLM text representation space and the DiT text conditioning space, which was originally designed to operate with the T5 encoder [79]. Pretrained diffusion models such as Flux [5] require substantial computational resources and massive text-to-image data to achieve effective alignment between the T5 text encoding space and the DiT conditioning space. In the current case, notably, even with no dedicated text-to-image alignment stage, our approach can already achieve comparable text alignment performance on standard benchmarks, and evidently superior prompt adherence on multimodal understanding. This suggests that, given sufficient computational budget and high-quality text-to-image data, our MLLM-DiT system would likely exhibit improved text-following capabilities.

Another limitation lies in the scope of multi-subject evaluation, although we manifest the model's capability of easily adapting to multi-subject scenarios in Section F. This is also owing in part to the scarcity nature of high-quality multi-subject data collections. Moreover, as the primary objective of this work is to investigate optimal strategies for squeezing MLLM capacity for subject-driven generation, multi-subject scenarios are less discussed to prevent from getting distracted from our central focus. Nevertheless, studying whether MLLMs can provide benefits with their internal knowledge for multi-subject harmonization and interaction—including, for instance, physical interactions between subjects—can be a promising direction for future research.

N Societal Impact

We expect our work to have a meaningful and positive societal impact by enabling more flexible and accessible personalized image generation. In particular, we sincerely wish that our method can help users express their creativity by generating personalized visual content for versatile applications. Moreover, we hope that our work serves as *the hitchhiker's guide* for future research to be aware of the great benefits of leveraging MLLMs for subject-driven generation, and to explore even more effective solutions to further squeeze capacity from MLLMs for various subject-driven tasks.

Potential negative societal impact. Our work is likely to be similar as other research on data generation regarding potential negative societal impact with the risk of digital forgery. In addition, unintended or inappropriate use of the technique may raise copyright and ethical concerns.

Figure B: Qualitative results of zero-out layers for text and image modalities in our DLA module.

Figure C: Qualitative results of zero-out layers for text and image modalities in our DLA module.

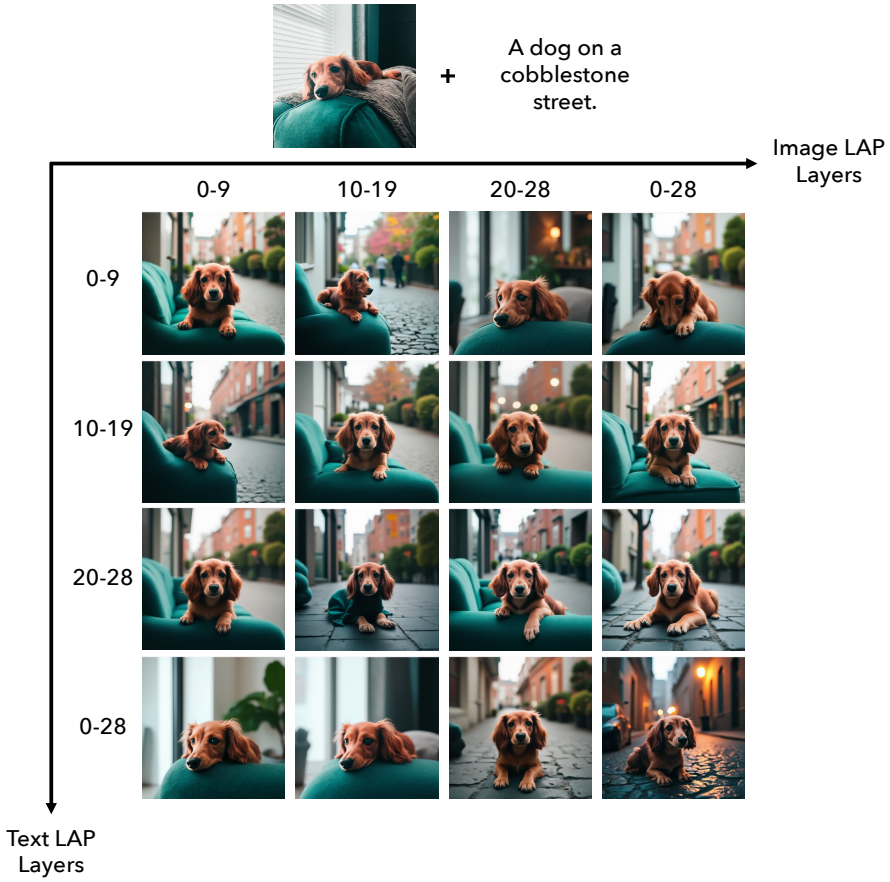

Figure D: Qualitative results of different layer selections for DLA. Rows correspond to text modality layer ranges, and columns correspond to image modality layer ranges, with layer groups 0–9, 10–19, 20–28, and all layers (0–28). Each subplot shows the output of a separately re-trained model for the given text–image layer combination.

Figure E: Qualitative comparison of different MLLM backbones, including Qwen2.5-VL-3B [4], Qwen2.5-VL-7B [4], InternVL3-2B [130], and InternVL3-8B [130].

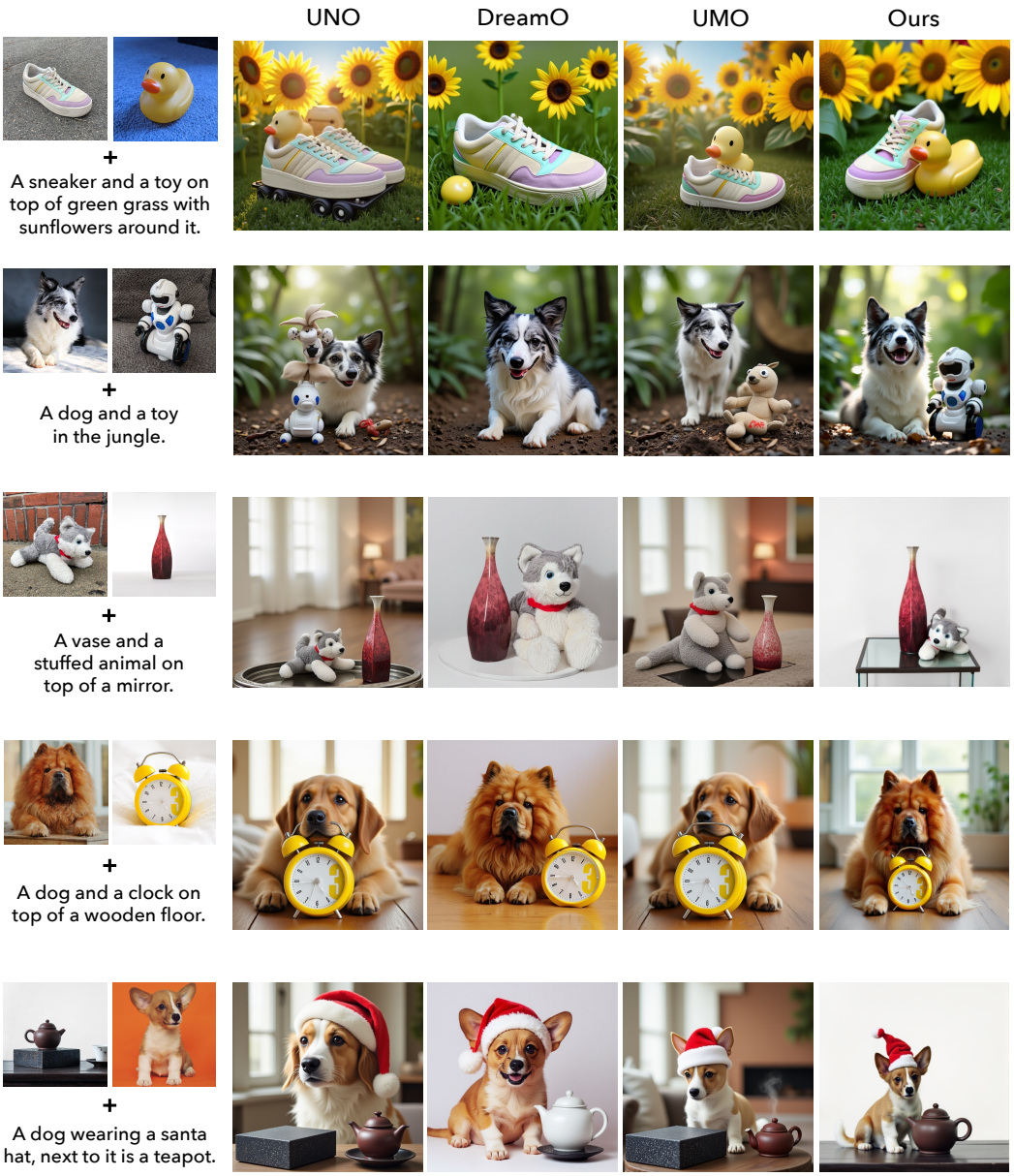

Figure F: Multi-reference generation results. Although our method is originally designed for single-subject generation, it adapts effectively to multi-subject scenarios after lightweight fine-tuning on MUSAR-Gen [31]. Compared to UNO [110], DreamO [69], and UMO [14], our model achieves clearer identity separation, consistent posture, and more reliable concept binding across subjects.

Figure G: Test samples from the constructed multimodal reasoning benchmark.

Figure H: Stress testing performance of our method on challenging scenarios, including attribute binding in long context prompts, and test cases that contain multiple instances of the subject. The results demonstrate the robustness of the strong reasoning capability within our MLLM-DiT system.

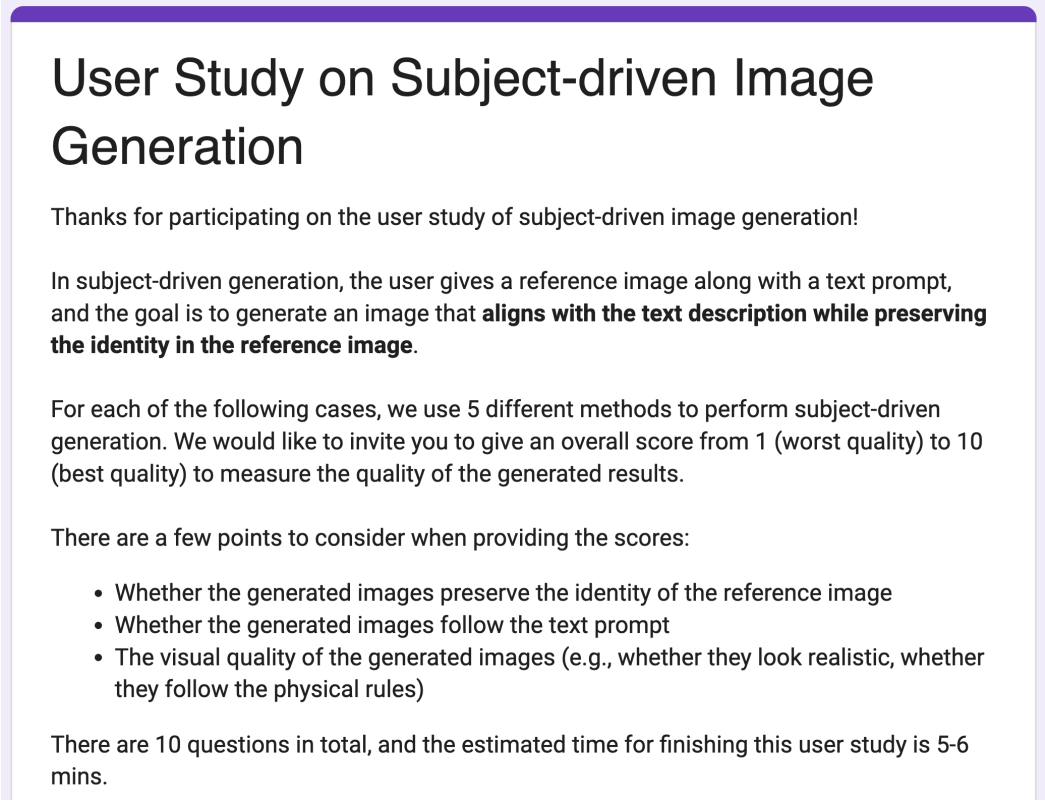

Figure I: Screenshot of the instructions of our user study.

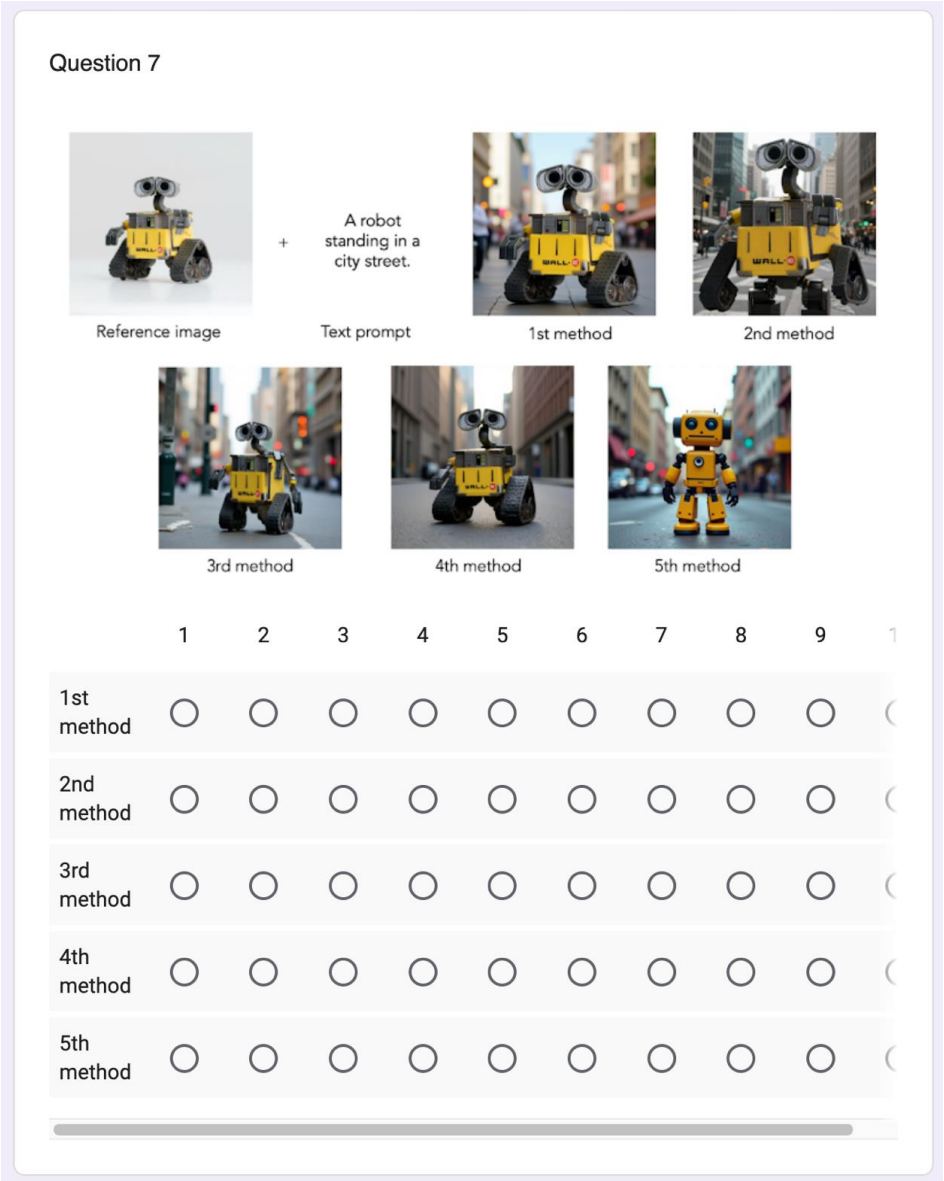

Figure J: Screenshot of the user study interface with an example question, containing the reference image, text prompts, and images generated by five methods placed side by side *in random order*, and the questions to score the five generated results.

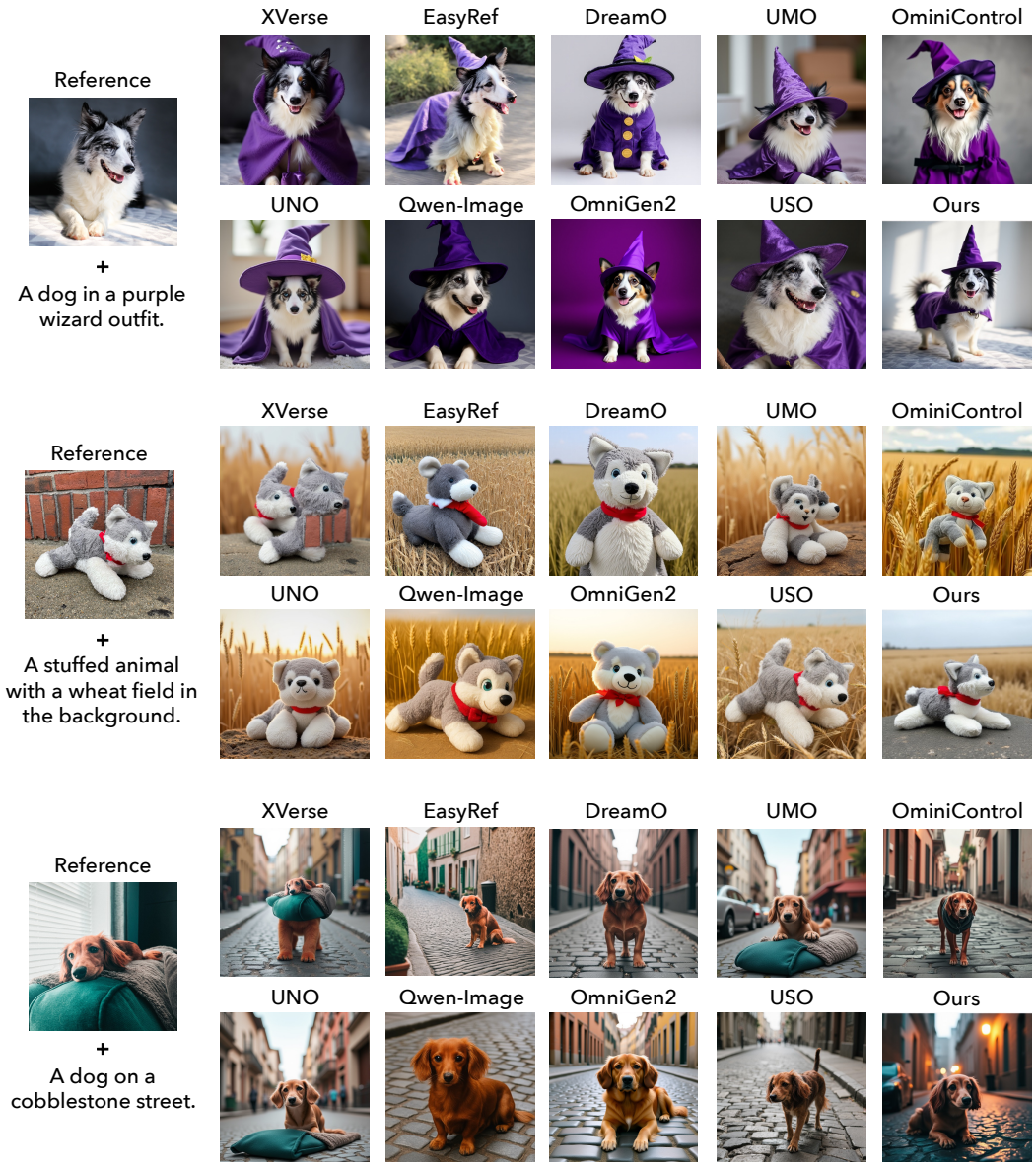

Figure K: Additional qualitative comparisons with state-of-the-art subject-driven generation methods. Our method consistently achieves better identity preservation and text alignment across diverse prompts. We compare against XVerse [8], EasyRef [131], DreamO [69], UMO [14], OminiControl [94], UNO [110], Qwen-Image [107], OmniGen2 [108], and USO [111].

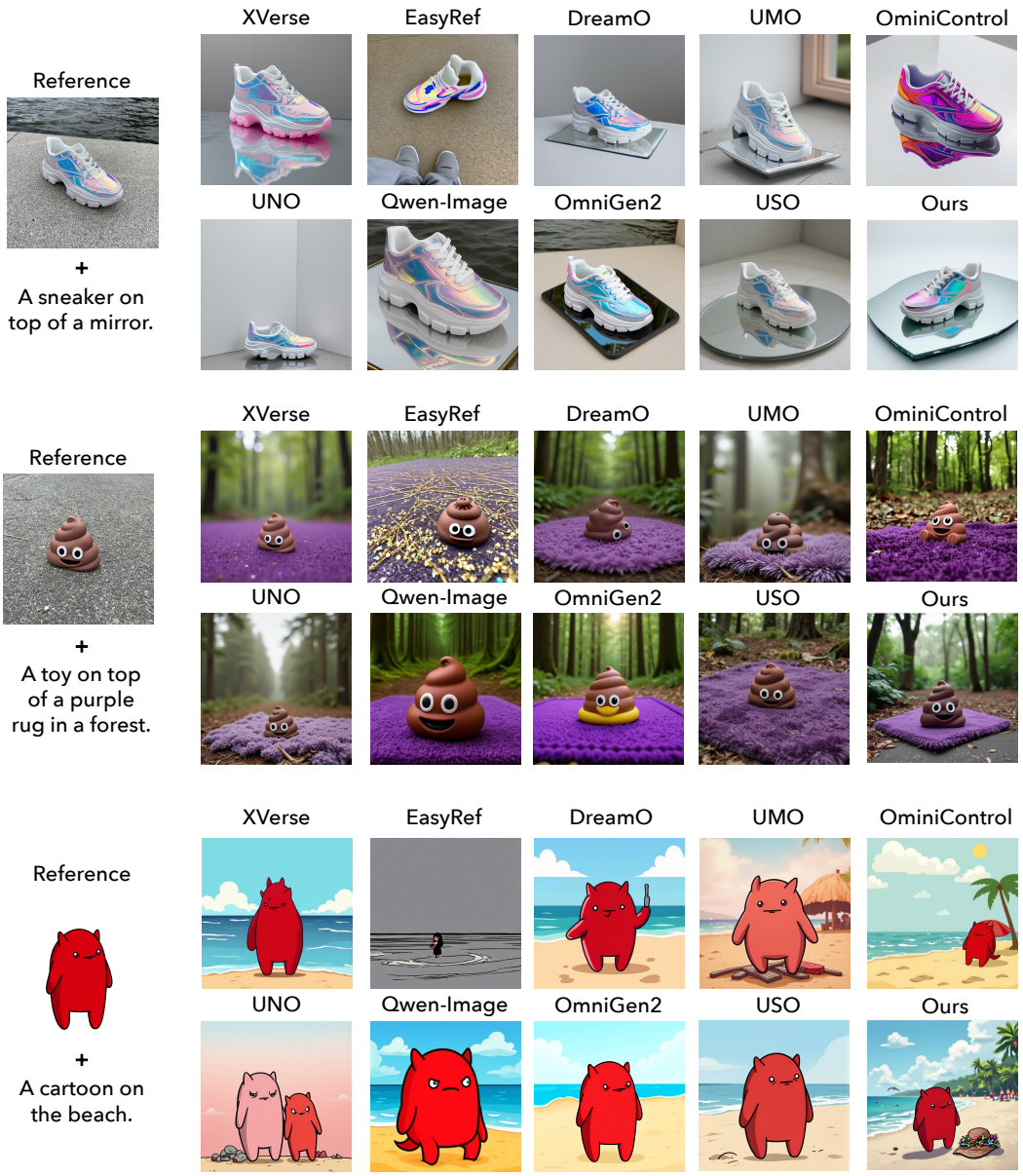

Figure L: Additional qualitative comparisons with state-of-the-art subject-driven generation methods. Our method consistently achieves better identity preservation and text alignment across diverse prompts. We compare against XVerse [8], EasyRef [131], DreamO [69], UMO [14], OminiControl [94], UNO [110], Qwen-Image [107], OmniGen2 [108], and USO [111].