

OminiControl: Minimal and Universal Control for Diffusion Transformer

Zhenxiong Tan Songhua Liu Xingyi Yang Qiaochu Xue Xinchao Wang
 National University of Singapore
 {zhenxiong, songhua.liu, xyang, e1352520}@u.nus.edu xinchao@nus.edu.sg

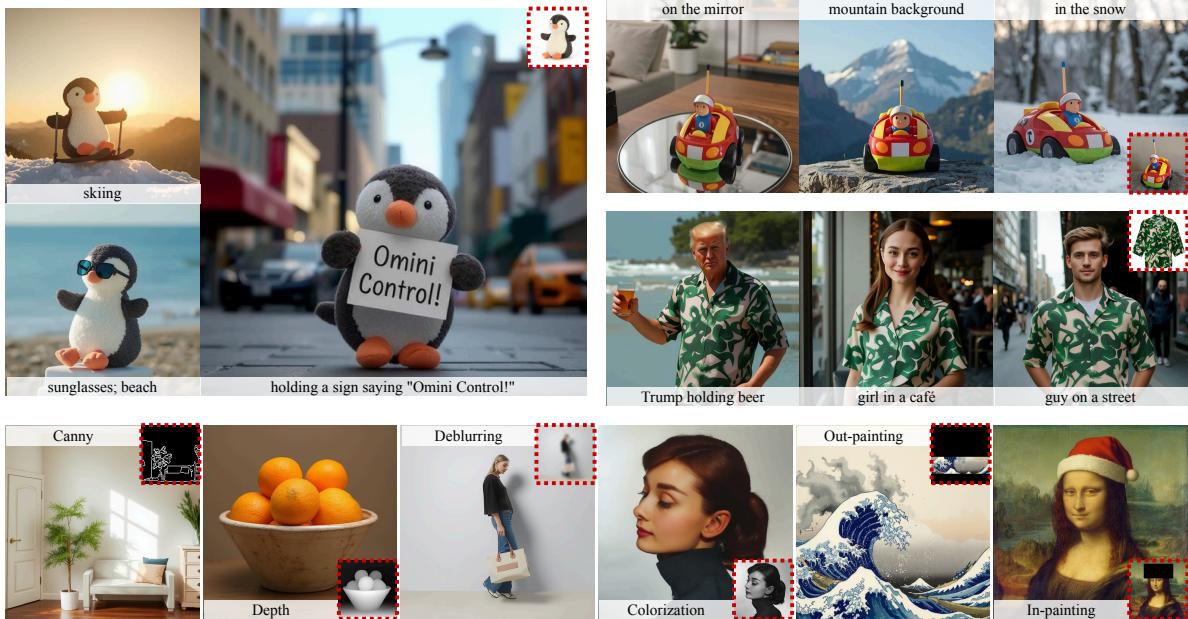


Figure 1. Results of our OminiControl on both subject-driven generation (top) and spatially-aligned tasks (bottom). The small images in red boxes show the input conditions.

Abstract

In this paper, we introduce OminiControl, a highly versatile and parameter-efficient framework that integrates image conditions into pre-trained Diffusion Transformer (DiT) models. At its core, OminiControl leverages a parameter reuse mechanism, enabling the DiT to encode image conditions using itself as a powerful backbone and process them with its flexible multi-modal attention processors. Unlike existing methods, which rely heavily on additional encoder modules with complex architectures, OminiControl (1) effectively and efficiently incorporates injected image conditions with only 0.1% additional parameters, and (2) addresses a wide range of image conditioning tasks in a unified manner, including subject-driven generation and spatially-aligned conditions such as edges, depth, and more. Remarkably, these capabilities are achieved by training on images generated by the DiT itself, which is particularly beneficial for subject-driven generation. Extensive

evaluations demonstrate that OminiControl outperforms existing UNet-based and DiT-adapted models in both subject-driven and spatially-aligned conditional generation. Additionally, we release our training dataset, Subjects200K, a diverse collection of over 200,000 identity-consistent images, along with an efficient data synthesis pipeline to advance research in subject-consistent generation.¹

1. Introduction

Diffusion models[9, 25, 28] have revolutionized the field of visual generation, demonstrating remarkable capabilities that significantly outperform traditional approaches like Generative Adversarial Networks (GANs)[6] in terms of image quality and diversity. While these models excel at generating highly realistic imagery, a critical challenge persists: enabling precise and flexible control over the genera-

¹Code and dataset are available at <https://github.com/YuanShi9815/OminiControl>

tion process to accommodate diverse and complex user requirements.

Text-based conditioning has been a cornerstone in advancing controllable generation[2, 13, 23, 25, 28, 35], offering an intuitive interface for users to specify their desired outputs. However, text prompts alone often fail to convey precise spatial details and structural attributes that users wish to control. Consequently, recent research has explored complementary conditioning modalities for guiding diffusion models, with image-based control emerging as a particularly effective approach[15, 22, 39, 41, 43]. This multi-modal conditioning strategy enables more detailed and accurate control over the generation process, addressing the limitations inherent in purely text-based interfaces.

Current image conditioning methods can be broadly categorized into spatially aligned and non-spatially aligned approaches. Spatially aligned tasks such as sketch-to-image and inpainting require direct correspondence between conditioning and output images, typically achieved through methods like ControlNet[41] that inject conditioning features in a spatially-preserving manner. In contrast, non-spatially aligned applications including subject driven generation and style transfer, as demonstrated by IP-Adapter[39], often employ pre-trained encoders like CLIP[27] to extract global features for integration via cross-attention mechanisms.

Despite the effectiveness of existing image-conditioned approaches, they present several limitations that hinder their efficiency and flexibility [22, 39, 41]. First, most existing methods are designed specifically for UNet-based architectures [16, 22, 24, 29, 31, 39–43], as seen in Stable Diffusion models [25, 28]. While these approaches work well with UNet’s encoder-decoder structure, they may not translate effectively to the more advanced Diffusion Transformer (DiT) models[23] that have demonstrated superior image generation quality[2, 13]. Additionally, current approaches typically specialize in either spatially aligned [22, 41, 43] or non-spatially aligned tasks [12, 15, 17, 39, 42], lacking a unified architecture to handle both control types effectively. This specialization often requires practitioners to employ different methods for different control scenarios, increasing system complexity and implementation overhead. Furthermore, these methods rely heavily on additional network structures [17, 22, 39, 41–43], which introduce substantial parameter overhead.

To address these limitations, we propose a parameter-efficient approach for incorporating image-based control into DiT architectures[23]. Our method reuse the model’s existing VAE encoder[28] to process conditioning images. Following the same token processing pipeline as noisy image tokens, we augment the encoded features with learnable position embeddings[34], and seamlessly integrate them alongside latent noise in the denoising network. This design

enables direct multi-modal attention interactions[23, 30] between condition and generation tokens throughout the DiT’s transformer blocks, facilitating efficient information exchange and control signal propagation.

We implemented our method on the high-performing DiT-structured diffusion model, FLUX.1-dev[13], a large-scale model containing 12 billion parameters. Extensive experiments on edge-guided generation, depth-aware synthesis, region-specific editing, and identity-preserving generation indicate that our DiT-based approach yields better results compared to both UNet-based implementations[7, 39, 41] and their community adaptations on the FLUX.1 model [14, 37].

For identity-preserving generation, we developed a novel data synthesis pipeline that generates high-quality, identity-consistent image pairs. Using this pipeline, we created a comprehensive dataset comprising over 200,000 diverse images. To facilitate future research in this direction, we will release both our dataset and the complete pipeline implementation as open-source resources².

In summary, we highlight our contributions as follows:

1. We present a parameter-efficient method for enabling image-conditioned control in Diffusion Transformer (DiT) models, achieving both spatially aligned and non-spatially aligned control within a unified framework.
2. We demonstrate the effectiveness of our approach through extensive experiments across diverse control tasks, including edge-guided generation, depth-aware synthesis, region-specific editing, and identity-preserving generation, consistently outperforming existing methods on both UNet implementations and their DiT adaptations.
3. We develop and release Subjects200K, a high-quality dataset of over 200,000 subject-consistent images, along with an efficient data synthesis pipeline, providing valuable resources to the research community for further exploration of subject-consistent generation tasks.

2. Related works

2.1. Diffusion-based models

Diffusion-based methods have emerged as a powerful framework for image generation[9, 28], demonstrating success across diverse tasks including text-to-image synthesis [2, 28, 33], image-to-image translation [32], and image editing [1, 20]. Recent advances have led to significant improvements in both quality and efficiency, notably through the introduction of latent diffusion models [28]. To further enhance generative capabilities, large-scale transformer architectures have been integrated into these frameworks, leading to advanced models like DiT[2, 3, 13, 23]. Building on these architectural innovations, FLUX[13] incor-

²More details are available in the supplementary material.

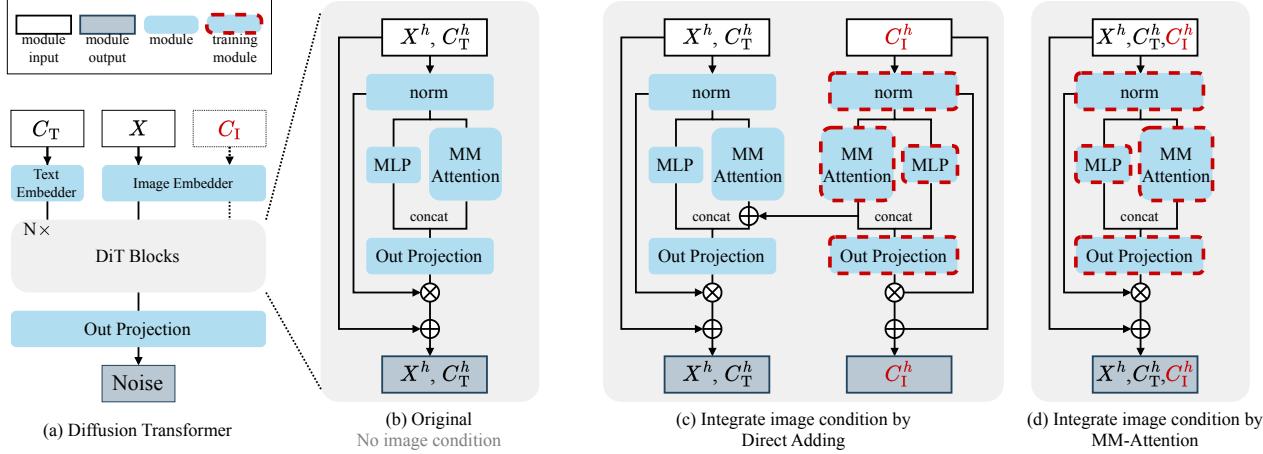


Figure 2. Overview of the Diffusion Transformer (DiT) architecture and integration methods for image conditioning.

porates transformer-based design with flow matching objectives [18], achieving state-of-the-art generation performance.

2.2. Controllable generation with diffusion models

Controllable generation has been extensively studied in the context of diffusion models. Text-to-image models[25, 28] have established a foundation for conditional generation, while various approaches have been developed to incorporate additional control signals such as image. Notable methods include ControlNet [41], enabling spatially aligned control in diffusion models, and T2I-Adapter [22], which improves efficiency with lightweight adapters. UniControl [26] uses Mixture-of-Experts (MoE) to unify different spatial conditions, further reducing model size. However, these methods rely on spatially adding condition information to the denoising network’s hidden states, inherently limiting their effectiveness for non-spatial tasks like subject-driven generation. IP-Adapter [39] addresses this by introducing cross-attention through an additional encoder, and SSR-Encoder [42] further enhances identity preservation in image-conditioned tasks. Despite these advances [5, 15, 19], a unified solution for both spatially aligned and non-aligned tasks remains elusive.

3. Methods

3.1. Preliminary

The Diffusion Transformer (DiT) model [23], employed in architectures like FLUX.1 [13], Stable Diffusion 3 [28], and PixArt [2], uses a denoising network of transformer blocks to refine noisy image tokens iteratively.

Each transformer block processes two types of tokens: noisy image tokens $\mathbf{X} \in \mathbb{R}^{N \times d}$ and text condition tokens $\mathbf{C}_T \in \mathbb{R}^{M \times d}$, where d is the embedding dimension, N and

M are the number of image and text tokens respectively (Figure 2). These tokens are embedded into hidden states X and C_T , which maintain consistent shapes throughout the transformer blocks.

In each DiT block, after normalizing X and C_T , they are processed by the core MM-Attention module [30], which employs Rotary Position Embedding (RoPE) [34] to incorporate positional dependencies across tokens. For a token at position (i, j) in the 2D grid, RoPE applies rotation matrices to the query and key projections:

$$Q_X(i, j) = W_Q(X_{i,j} \cdot R(i, j)), \quad (1)$$

$$K_X(i, j) = W_K(X_{i,j} \cdot R(i, j)), \quad (2)$$

where $R(i, j)$ is the rotation matrix at position (i, j) . Similarly, the text condition tokens C_T have their query and key projections defined in the same way, with all text token positions set to $(0, 0)$ in FLUX.1.

After applying RoPE, the queries, keys, and values from both token types are concatenated to form unified matrices Q_Z , K_Z , and V_Z , representing the combined token set $Z = [X; C_T]$. The MM-Attention operation is then computed as:

$$\text{MMAttention}(Z) = \text{softmax} \left(\frac{Q_Z K_Z^\top}{\sqrt{d}} \right) V_Z, \quad (3)$$

enabling interactions between image and text condition tokens through the attention mechanism.

3.2. Image condition integration

Our approach first encodes the condition image through the model’s VAE, projecting it into the same latent space as the noisy image tokens to form $\mathbf{C}_I \in \mathbb{R}^{N \times d}$.

Previous methods like ControlNet [41] and T2I-Adapter [22] incorporate the condition image by spatially aligning and adding its hidden states directly to those of the



Figure 3. Comparison of results using two methods for integrating image conditions. The multi-modal approach demonstrates better condition following compared to direct addition.

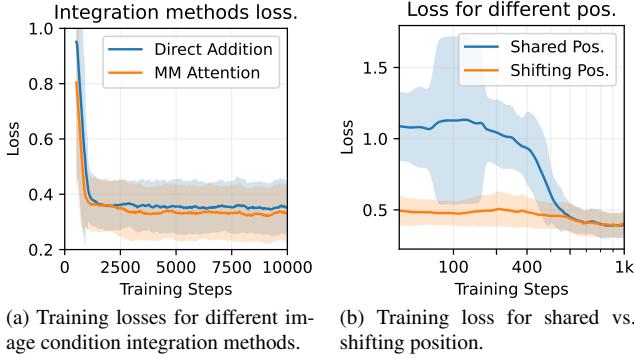


Figure 4. Training loss comparisons.

noisy image tokens:

$$H_X = H_X + H_{C_I}, \quad (4)$$

where H_X represents the combined hidden states for further processing, with H_{C_I} being the hidden states from the condition image. While this approach proves effective for spatially aligned tasks, it faces two key limitations: (1) it lacks flexibility when handling non-aligned scenarios, and (2) even in spatially aligned cases, the direct addition of hidden states constrains token interactions, potentially limiting the model’s performance.

In contrast, to enable non-aligned control tasks and provide greater conditioning flexibility, our method processes condition image tokens uniformly with text and noisy image tokens, integrating them into a unified sequence:

$$Z = [X; C_T; C_I], \quad (5)$$

where Z represents the concatenated sequence of noisy image tokens X , text tokens C_T , and condition image tokens C_I . This unified approach enables direct participation in multi-modal attention [30] without specialized processing pathways (illustrated in Figure 2).

The comparative result shows that our approach achieves higher generation quality and better alignment with the conditions compared to the direct adding method, as illustrated in Figure 3. Moreover, the training curves in Figure 4a demonstrate that multi-modal attention method consistently achieves lower loss values than the direct adding approach.

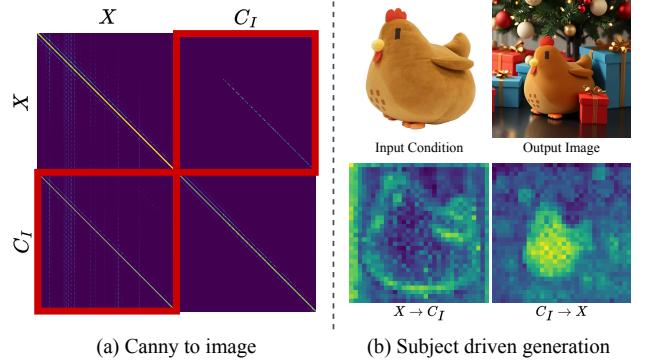


Figure 5. (a) Attention maps for the Canny-to-image task, showing interactions between noisy image tokens X and image condition tokens C_I . Strong diagonal patterns indicate effective spatial alignment. (b) Subject-driven generation task, with input condition and output image. Attention maps for $X \rightarrow C_i$ and $C_i \rightarrow X$ illustrate accurate subject-focused attention.

Besides, the effectiveness of this unified sequence approach is demonstrated across both spatially aligned and non-spatially aligned tasks (Figure 5), highlighting its versatility in handling diverse conditional generation scenarios.

3.3. Adaptive position embedding

Our unified sequence design allows for flexible integration of condition image tokens, but this requires incorporating positional information to ensure effective interaction with noisy image tokens. The relative positioning of these tokens is critical, as it directly affects the model’s learning efficiency and generalization capability.

In FLUX.1’s Transformers, each token is assigned a corresponding position index to encode spatial information. For a 512×512 target image, the VAE [11] encoder first projects it into the latent space, then the latent representation is divided into a 32×32 grid of tokens, where each token is assigned a unique two-dimensional position index (i, j) with $i, j \in [0, 31]$. This indexing scheme preserves the spatial structure of the original image in the latent space, while text tokens maintain a fixed position index of $(0, 0)$.

For spatially aligned tasks, our initial approach was to assign condition tokens the same position embeddings as their corresponding tokens in the noisy image. However, for non-spatially aligned tasks such as subject-driven generation, our experiments revealed that shifting the position indices of condition tokens leads to faster convergence (Figure 4b). Specifically, we shift the condition image tokens to indices (i, j) where $i \in [0, 31]$ and $j \in [32, 64]$, ensuring no spatial overlap with the original image tokens X .

3.4. Condition strength factor

The unified attention mechanism we adopt not only enables flexible token interaction but also allows us to precisely

control the influence of condition images. Specifically, we designed a method that allows for manual adjustment of the condition image’s effect during inference. For a given strength factor γ , setting $\gamma = 0$ removes the condition image’s influence, resulting in an output based purely on the original input. At $\gamma = 1$, the output fully reflects the condition image, and as γ increases beyond 1, the condition’s effect becomes even more pronounced.

To achieve this controllability, we introduce a bias term into the original MM-Attention operation. Specifically, we modify Equation 3 to:

$$\text{BiasedAttention}(Z) = \text{softmax} \left(\frac{Q_Z K_Z^\top}{\sqrt{d}} + \text{bias}(\gamma) \right) V_Z, \quad (6)$$

where $\text{bias}\gamma$ is designed to adjust the attention weights between condition tokens and other tokens based on the strength factor γ . The bias term is constructed as a $(M + 2N) \times (M + 2N)$ matrix, where M is the number of text tokens, and N is the number of noisy image tokens and condition image tokens each. The matrix has the following structure:

$$\text{bias}(\gamma) = \begin{bmatrix} \mathbf{0}_{M \times M} & \mathbf{0}_{M \times N} & \mathbf{0}_{M \times N} \\ \mathbf{0}_{N \times M} & \mathbf{0}_{N \times N} & \log(\gamma) \mathbf{1}_{N \times N} \\ \mathbf{0}_{N \times M} & \log(\gamma) \mathbf{1}_{N \times N} & \mathbf{0}_{N \times N} \end{bmatrix}. \quad (7)$$

This design ensures that the strength factor γ only affects the attention weights between noisy image tokens and condition image tokens, while maintaining the original attention patterns for text tokens and within-modality interactions.

3.5. Subjects200K datasets

Training models for subject-consistent generation typically requires paired images that maintain identity consistency while exhibiting variations in pose, lighting, and other attributes. Previous methods like IP-Adapter [39] use identical images for conditioning and target pairs, which proves effective for their approaches. However, in our framework, this setup leads to overfitting, causing the model to generate outputs nearly identical to the inputs.

To overcome these limitations, we developed a dataset featuring images that preserve subject identity while incorporating natural variations. While existing datasets [12, 15, 17, 31] address similar needs, they often face constraints in either quality or scale. We therefore propose a novel synthesis pipeline leveraging FLUX’s inherent capability to generate pairs of visually related images from carefully crafted prompts.

Our pipeline utilizes ChatGPT-4o to generate over 20,000 diverse image descriptions, which guide FLUX in producing more than 200,000 images (Figure 6). The generated images undergo quality assessment using ChatGPT-

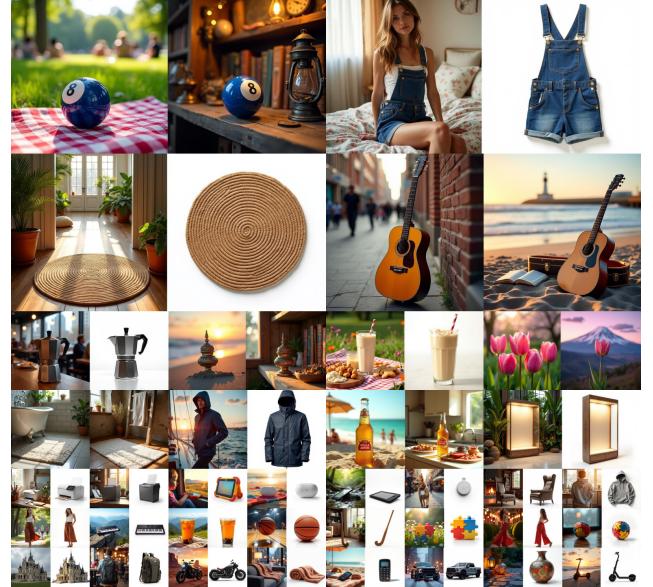


Figure 6. Examples from our Subjects200Kdataset. Each pair of images shows the same object in varying positions, angles, and lighting conditions. The dataset includes diverse objects such as clothing, furniture, vehicles, and animals, totaling over 200,000 images. This dataset, along with the generation pipeline, will be publicly released.

4o’s vision capabilities, ensuring both consistency and diversity in the final dataset. The curated dataset and complete generation pipeline are publicly available³.

4. Experiment

4.1. Setup

Tasks and base model. We evaluate our method on two categories of conditional generation tasks: spatially aligned tasks (including Canny-to-image, depth-to-image, masked-based inpainting, and colorization) and subject-driven generation. We build our method upon FLUX.1 [13], a latent rectified flow transformer model for image generation. By default, we use FLUX.1-dev to generate images for spatially aligned tasks. In subject-driven generation tasks, we switch to FLUX.1-schnell as we observed it tends to produce better visual quality.

Implement details. Our method utilizes LoRA [4] for fine-tuning the base model with a default rank of 4. To preserve the model’s original capabilities and achieve flexibility, the LoRA scale is set to 0 when processing non-condition tokens by default.

Training. Our model is trained with a batch size of 1 and gradient accumulation over 8 steps (effective batch size of

³More details are provided in the supplementary material. Dataset and code are available at <https://github.com/YuanShi9815/Subjects200K>.

Condition	Model	Method	Controllability F1 ↑ /MSE ↓	General Quality				Text Consistency CLIP-Score ↑
				FID ↓	SSIM ↑	MAN-IQA ↑	MUSIQ ↑	
Canny	SD1.5	ControlNet	0.34	18.74	0.35	0.45	67.81	0.75
		T2I-Adapter	0.22	20.06	0.35	0.39	67.88	0.74
	FLUX.1	ControlNet	0.21	98.68	0.25	0.37	56.90	0.53
		Ours	0.38	20.63	0.40	0.61	75.91	0.76
Depth	SD1.5	ControlNet	923	23.02	0.34	0.47	70.73	0.726
		T2I-Adapter	1560	24.72	0.27	0.39	69.99	0.72
	FLUX.1	ControlNet	2958	62.20	0.26	0.38	66.84	0.54
		Ours	903	27.26	0.39	0.55	75.06	0.728
Deblur	FLUX.1	ControlNet	572	30.38	0.74	0.31	54.37	0.78
Colorization	FLUX.1	ControlNet	351	16.27	0.64	0.43	70.95	0.85
Mask	SD1.5	ControlNet	7588	13.14	0.40	0.41	67.22	0.84
	FLUX.1	Ours	6248	15.66	0.48	0.45	72.61	0.80

Table 1. Quantitative comparison with baseline methods on five spatially aligned tasks. We evaluate methods based on Controllability (F1-Score for Canny, MSE for others), General Quality (FID, SSIM, MAN-IQA, MUSIQ), and Text Consistency (CLIP-Score). For F1-Score, higher is better; for MSE, lower is better. Best results are shown in **bold**.

8). We employ the Prodigy optimizer [21] with safeguard warmup and bias correction enabled, setting the weight decay to 0.01. The experiments are conducted on 2 NVIDIA H100 GPUs (80GB each). For spatially aligned tasks, models are trained for 50,000 iterations, while subject-driven generation models are trained for 15,000 iterations.

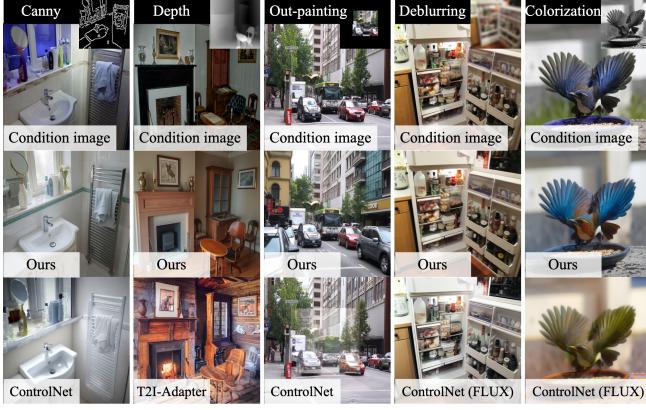
Baselines. For spatially-aligned tasks, we compare our method with both the original ControlNet [41] and T2I-Adapter [22] on Stable Diffusion 1.5, as well as ControlNetPro [14], the FLUX.1 implementation of ControlNet. For subject-driven generation, we compare with IP-Adapter [39], evaluating its implementations FLUX.1 [37].

Evaluation metrics. We evaluate our model on both spatially aligned tasks and subject-driven generation. For spatially aligned tasks, we assess two aspects: generation quality and controllability. Generation quality is measured using FID [8], SSIM, MAN-IQA [38], and MUSIQ [10] for visual fidelity, along with CLIP Score [27] for semantic consistency. For controllability, we compute F1 Score between extracted and input edge maps in edge-conditioned generation, and MSE between extracted and original condition maps for other tasks (using Depth Anything for depth, color channel separation for colorization, etc.). For subject-driven generation, we propose a five-criteria framework evaluating both preservation of subject characteristics (identity preservation, material quality, color fidelity, natural appearance) and accuracy of requested modifications, with all assessments conducted through GPT-4o’s vision capabilities to ensure systematic evaluation. Detailed evaluation methodology is presented in Appendix B.1.

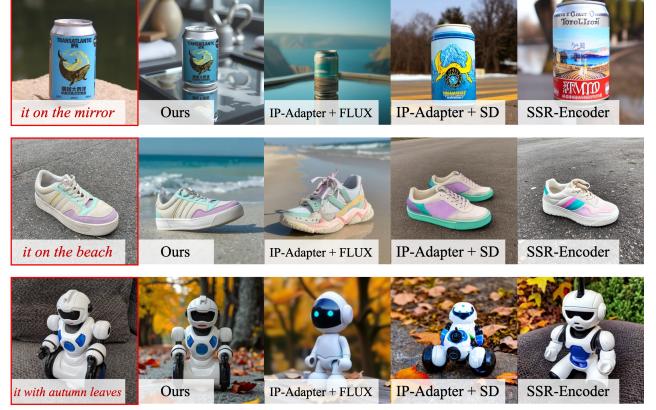
Evaluation protocol. We conducted evaluations on two datasets. For spatially-aligned tasks, we use COCO 2017 validation set (5,000 images) resized to 512×512 , using task-specific conditions and associated captions as prompts with a fixed seed of 42. For subject-driven generation, we test on 750 text-condition pairs (30 subjects \times 25 prompts) from DreamBooth[31] dataset with 5 different seeds, using one selected image per subject as the condition.

4.2. Main result

Spatially aligned tasks As shown in Table 1, we comprehensively evaluate our method against existing approaches on five spatially aligned tasks. Our method achieves the highest F1-Score of 0.38 on depth-to-image generation, significantly outperforming both SD1.5-based methods ControlNet [41] and T2I-Adapter [22], as well as FLUX.1-based ControlNetPro [14]. In terms of general quality metrics, our approach demonstrates consistent superiority across most tasks, showing notably better performance in SSIM [36], MAN-IQA [38], and MUSIQ [10] scores. For challenging tasks like deblurring and colorization, our method achieves substantial improvements: the MSE is reduced by 77% and 93% respectively compared to ControlNetPro, while the FID scores [8] improve from 30.38 to 11.49 for deblurring. The CLIP-Score metrics [27] indicate that our method maintains high text-image consistency across all tasks, suggesting effective preservation of semantic alignment while achieving better control and visual quality. As shown in Figure 7, our method produces sharper details and more faithful color reproduction in colorization



(a) Spatially aligned tasks



(b) Subject driven generation

Figure 7. Qualitative results comparing different methods. Left: Spatially aligned tasks across Canny, depth, out-painting, deblurring, colorization. Right: Subject-driven generation with beverage can, shoes and robot toy. Our method demonstrates superior controllability and visual quality across all tasks.

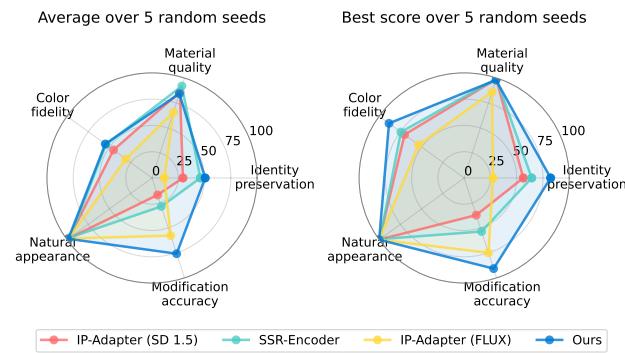


Figure 8. Radar charts visualization comparing our method (blue) with baselines across five evaluation metrics.

tasks, while maintaining better structural fidelity in edge-guided generation and deblurring scenarios.

Subject driven generation Figure 8 presents a comprehensive comparison against existing baselines. Our method demonstrates superior performance, particularly in identity preservation and modification accuracy. Averaging over random seeds, we achieve 75.8% modification accuracy compared to IP-Adapter (FLUX)'s 57.7%, while maintaining 50.6% identity preservation against IP-Adapter (SD 1.5)'s 29.4%. The advantage amplifies in best-seed scenarios, achieving 90.7% modification accuracy and 82.3% identity preservation - surpassing the strongest baselines by 15.8 and 18.0 percentage points, demonstrating effective subject-fidelity editing. These quantitative results are further corroborated by user studies presented in Appendix B.1.

Comparative parameter efficiency. As shown in Table 2, our approach achieves remarkable parameter effi-

Methods	Base model	Parameters	Ratio
ControlNet		361M	~42%
T2I-Adapter	SD1.5 / 860M	77M	~9.0%
IP-Adapter		449M	~52.2%
ControlNet	FLUX.1 / 12B	3.3B	~27.5%
IP-Adapter		918M	~7.6%
Ours	FLUX.1 / 12B	14.5M / 48.7M w/ Encoder	~0.1% / ~0.4% w/ Encoder

Table 2. Additional parameters introduced by different image conditioning methods. For IP-Adapter, the parameter count includes the CLIP Image encoder. For our method, we also report results when using the original VAE encoder from FLUX.1.

ciency compared to existing methods. For the 12B parameter FLUX.1 model, our method requires only 14.5M trainable parameters (approximately 0.1%), which is significantly lower than ControlNet (27.5%) and IP-Adapter (7.6%). Even when utilizing the original VAE encoder from FLUX.1, our method still maintains high efficiency with just 0.4% additional parameters, demonstrating the effectiveness of our parameter-efficient design.

4.3. Empirical studies

Effect of training data. For subject-driven generation, our model takes a reference image of a subject (e.g., a plush toy or an object) and a text description as input, aiming to generate novel images of the same subject following the text guidance while preserving its key characteristics.

To validate the effectiveness of our Subjects200K dataset described in Section 3.5, we compare two training strategies for this task. The first approach relies on traditional data augmentation, where we apply random cropping, rotation, scaling, and adjustments to contrast, saturation, and color to the original images. The second approach utilizes

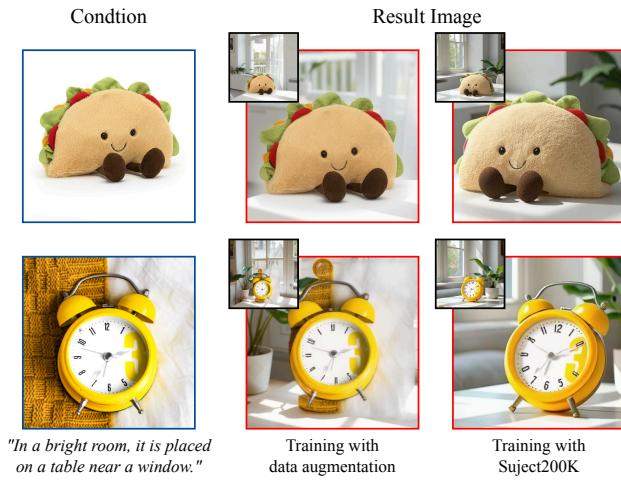


Figure 9. Comparison of models trained with different data. The model trained by data augmentation tends to copy inputs directly, while model trained by our Subjects200K generates novel views while preserving identity.

our Subjects200K dataset. As shown in Figure 9, the model trained with data augmentation only learns to replicate the input conditions with minimal changes. In the first row, it simply places the taco plush toy in a bright room setting while maintaining its exact appearance and pose. Similarly, in the second row, the yellow alarm clock is reproduced with nearly identical details despite the window-side placement instruction. In contrast, our Subjects200K-trained model demonstrates the ability to generate diverse yet consistent views of the subjects while faithfully following the text prompts.

Condition strength analysis. We evaluate our condition strength control through qualitative experiments. Figure 10 shows generated results with varying strength factor $\gamma \in \{0.25, 0.5, 0.75, 1.0\}$. Results show that γ effectively controls the generation process for both spatially-aligned tasks like depth-to-image generation and non-spatially-aligned tasks like subject-driven generation, enabling flexible control over the condition’s influence.

Impact of LoRA rank. We conducted extensive experiments with different LoRA ranks (1, 2, 4, 8, and 16) for the Canny-to-image task. As shown in Table 3, our experiments show that increasing the LoRA rank generally improves model performance, with rank 16 achieving the best results across multiple aspects: image quality (measured by FID and SSIM), condition control capability (measured by F1 Score), while maintaining competitive text-image consistency (measured by CLIP-Score). Notably, even with smaller ranks (e.g., rank 1), the model demonstrates competitive performance, especially in text-image alignment where it achieves the highest CLIP-Score of 0.765, showing



Figure 10. Comparison of models trained with different data. The model trained by data augmentation tends to copy inputs directly, while model trained by our Subjects200K generates novel views while preserving identity.

Study	Setting	FID ↓	SSIM ↑	F1 Score ↑	CLIP Score ↑
LoRA Rank	1	21.09	0.412	0.385	0.765
	2	21.28	0.411	0.377	0.751
	4	20.63	0.407	0.380	0.761
	8	21.40	0.404	0.3881	0.761
	16	19.71	0.425	0.407	0.764
Condition Blocks	Early	25.66	0.369	0.23	0.72
	Full	20.63	0.407	0.38	0.76

Table 3. Ablation studies on (1) LoRA rank for the *Canny-to-image* task and (2) condition signal integration approaches. Results show that LoRA rank of 16 and full-depth integration achieve the best performance. Rows with blue background indicate our default settings (LoRA rank=4, Full condition integration). Best results are in **bold**.

the efficiency of our approach even with limited parameters.

Conditioning depth. FLUX.1’s transformer architecture features two distinct types of blocks: early blocks that employ separate normalization modules for different modalities tokens (text and image) and later blocks that share unified normalization across all tokens. As shown in Table 3, experiments reveal that restricting condition signal integration to only these early blocks results in insufficient controllability over the generation process. This suggests that allowing the condition signals to influence the entire transformer stack is crucial for achieving the desired levels of control over the output. Notably, this finding indicates that the previous approaches[14, 22, 37, 39, 41] of inserting condition signals primarily in early blocks, which were effective in UNet-based architectures, may not fully translate to DiT-based models like FLUX.1.

5. Conclusion

OminiControl offers parameter-efficient image-conditioned control for Diffusion Transformers across diverse tasks using a unified token approach, without extra modules. Our method outperforms traditional approaches, and the new Subjects200K dataset—featuring over 200,000 high-

quality, subject-consistent images—supports advancements in subject-consistent generation. Results confirm Omini-Control’s scalability and effectiveness in diffusion models.

6. Acknowledgment

We would like to acknowledge that the computational work involved in this research work is partially supported by NUS IT’s Research Computing group using grant numbers NUSREC-HPC-00001.

References

- [1] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18208–18218, 2022. [2](#)
- [2] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart- α : Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023. [2, 3](#)
- [3] Junsong Chen, Yue Wu, Simian Luo, Enze Xie, Sayak Paul, Ping Luo, Hang Zhao, and Zhenguo Li. Pixart- $\{\backslash\delta\}$: Fast and controllable image generation with latent consistency models. *arXiv preprint arXiv:2401.05252*, 2024. [2](#)
- [4] Shilpa Devalal and A Karthikeyan. Lora technology—an overview. In *2018 second international conference on electronics, communication and aerospace technology (ICECA)*, pages 284–290. IEEE, 2018. [5](#)
- [5] 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. *arXiv preprint arXiv:2208.01618*, 2022. [3](#)
- [6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. [1](#)
- [7] Jacky Hate. Text-to-image-2m dataset. <https://huggingface.co/datasets/jackyhate/text-to-image-2M>, 2024. [2](#)
- [8] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. [6](#)
- [9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. [1, 2](#)
- [10] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5148–5157, 2021. [6](#)
- [11] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. [4](#)
- [12] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1931–1941, 2023. [2, 5, 1](#)
- [13] Black Forest Labs. Flux: Official inference repository for flux.1 models, 2024. Accessed: 2024-11-12. [2, 3, 5, 1](#)
- [14] Shakker Labs. Flux.1-dev-controlnet-union-pro. <https://huggingface.co/Shakker-Labs/FLUX.1-dev-ControlNet-Union-Pro>, 2024. [2, 6, 8](#)
- [15] Dongxu Li, Junnan Li, and Steven Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *Advances in Neural Information Processing Systems*, 36, 2024. [2, 3, 5, 1](#)
- [16] Ming Li, Taojiannan Yang, Huafeng Kuang, Jie Wu, Zhaoning Wang, Xuefeng Xiao, and Chen Chen. Controlnet++: Improving conditional controls with efficient consistency feedback. In *European Conference on Computer Vision*, pages 129–147. Springer, 2025. [2](#)
- [17] Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8640–8650, 2024. [2, 5, 1](#)
- [18] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. [3](#)
- [19] Jian Ma, Junhao Liang, Chen Chen, and Haonan Lu. Subject-diffusion: Open domain personalized text-to-image generation without test-time fine-tuning. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–12, 2024. [3](#)
- [20] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. [2](#)
- [21] Konstantin Mishchenko and Aaron Defazio. Prodigy: An expeditiously adaptive parameter-free learner. In *Forty-first International Conference on Machine Learning*, 2024. [6](#)
- [22] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4296–4304, 2024. [2, 3, 6, 8](#)
- [23] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. [2, 3](#)
- [24] Bohao Peng, Jian Wang, Yuechen Zhang, Wenbo Li, Ming-Chang Yang, and Jiaya Jia. Controlnext: Powerful and efficient control for image and video generation. *arXiv preprint arXiv:2408.06070*, 2024. [2](#)
- [25] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. [1, 2, 3](#)
- [26] Can Qin, Shu Zhang, Ning Yu, Yihao Feng, Xinyi Yang, Yingbo Zhou, Huan Wang, Juan Carlos Niebles, Caiming

- Xiong, Silvio Savarese, et al. Unicontrol: A unified diffusion model for controllable visual generation in the wild. *arXiv preprint arXiv:2305.11147*, 2023. 3
- [27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 6
- [28] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 1, 2, 3
- [29] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5–9, 2015, proceedings, part III* 18, pages 234–241. Springer, 2015. 2
- [30] Ludan Ruan, Yiyang Ma, Huan Yang, Huigu He, Bei Liu, Jianlong Fu, Nicholas Jing Yuan, Qin Jin, and Baining Guo. Mm-diffusion: Learning multi-modal diffusion models for joint audio and video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10219–10228, 2023. 2, 3, 4
- [31] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023. 2, 5, 6, 1
- [32] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 conference proceedings*, pages 1–10, 2022. 2
- [33] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 2
- [34] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. 2, 3
- [35] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *arXiv preprint arXiv:2404.02905*, 2024. 2
- [36] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 6
- [37] XLabs-AI. Flux-ip-adapter. <https://huggingface.co/XLabs-AI/flux-ip-adapter>, 2024. 2, 6, 8
- [38] Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. Maniq: Multi-dimension attention network for no-reference image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1191–1200, 2022. 6
- [39] 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. 2, 3, 5, 6, 8
- [40] Denis Zavadski, Johann-Friedrich Feiden, and Carsten Rother. Controlnet-xs: Designing an efficient and effective architecture for controlling text-to-image diffusion models. *arXiv preprint arXiv:2312.06573*, 2023.
- [41] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 2, 3, 6, 8
- [42] Yuxuan Zhang, Yiren Song, Jiaming Liu, Rui Wang, Jinpeng Yu, Hao Tang, Huaxia Li, Xu Tang, Yao Hu, Han Pan, et al. Ssr-encoder: Encoding selective subject representation for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8069–8078, 2024. 2, 3
- [43] Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K Wong. Uni-controlnet: All-in-one control to text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024. 2

OminiControl: Minimal and Universal Control for Diffusion Transformer

Supplementary Material

A. Details of Subjects200K datasets

We present a comprehensive synthetic dataset constructed to address the limitations in scale and image quality found in previous datasets [12, 15, 17, 31]. Our approach leverages FLUX.1-dev [13] to generate high-quality, consistent images of the same subject under various conditions.

Subjects200K dataset currently consists of two splits, both generated using similar pipelines. Split-1 contains paired images of objects in different scenes, while Split-2 pairs each object's scene image with its corresponding studio photograph. Due to their methodological similarities, we primarily focus on describing the synthesis process and details of Split-2, although both splits are publicly available. Our complete Subjects200K dataset can be fully accessed via this [link](#).

A.1. Generation pipeline

Our dataset generation process consists of three main stages: description generation, image synthesis, and quality assessment.

Description Generation We employed ChatGPT-4o to create a hierarchical structure of descriptions: We first generated 42 diverse object categories, including furniture, vehicles, electronics, clothing, and others. For each category, we created multiple object instances, totaling 4,696 unique objects. Each object entry consists of: (1) A brief description, (2) Eight diverse scene descriptions, (3) One studio photo description. Figure S2 shows a representative example of our structured description format.

Image Synthesis We designed a prompt template to leverage FLUX's capability of generating paired images containing the same subject. Our template synthesizes a comprehensive prompt by combining a brief object description with two distinct scene descriptions, ensuring subject consistency while introducing environmental variations.

The detailed prompt structure is illustrated in Figure S3. For each prompt, we set the image dimensions to 1056×528 pixels and generated five images using different random seeds to ensure diversity in our dataset. During the training process, we first split the paired images horizontally, then performed central cropping to obtain 512×512 pixel image pairs. This padding strategy was implemented to address cases where the generated images were not precisely bisected, preventing potential artifacts from appearing in the wrong half of the split images.

Quality assessment We leveraged ChatGPT-4o's vision capabilities to rigorously evaluate the quality of images generated by FLUX.1-dev. The assessment focused on multiple

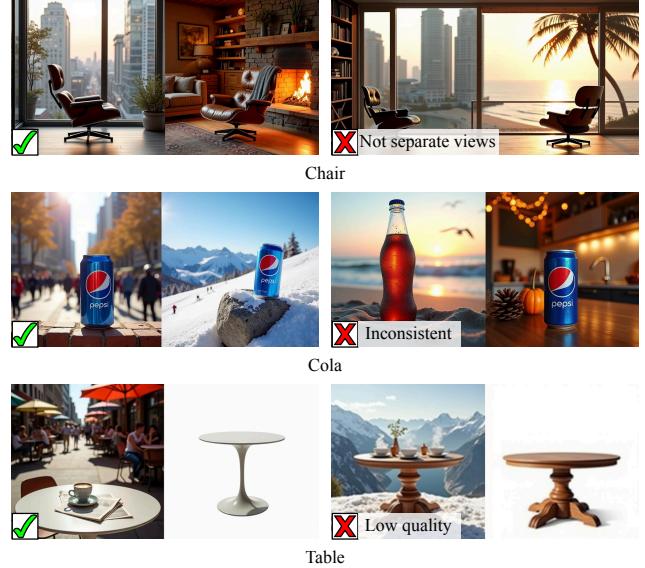


Figure S1. Examples of successful and failed generation results from Subjects200K dataset. Green checks indicate successful cases where subject identity and characteristics are well preserved, while red crosses show failure cases.

critical aspects:

- Image composition: Verifying that each image properly contains two side-by-side views.
- Subject consistency: Ensuring the subject maintains identity across both views.
- Image quality: Confirming high resolution and visual fidelity.

To maintain stringent quality standards, each image underwent five independent evaluations by ChatGPT-4o. Only images that passed all five evaluations were included in our training dataset. Figure S1 presents representative examples from our quality-controlled dataset.

A.2. Dataset Statistics

In Split-2, we first generated 42 distinct object categories, from which we created and curated a set of 4,696 detailed object instances. Then we combine these descriptions to generate 211,320 subject-consistent image pairs. Through rigorous quality control using GPT-4o, we selected 111,767 high-quality image pairs for our final dataset. This extensive filtering process ensured the highest standards of image quality and subject consistency, resulting in a collection of 223,534 high-quality training images.

```
{
    "brief_description": 
        "A finely-crafted wooden seating piece.",
    "scene_descriptions": [
        "Set on a sandy shore at dusk, it faces the ocean with a gentle breeze rustling nearby palms, bathed in soft, warm twilight.",
        "Positioned in a bustling urban cafe, it stands out against exposed brick walls, capturing the midday sun through a wide bay window."
        // Additional six scene descriptions omitted
    ],
    "studio_photo_description": 
        "In a professional studio against a plain white backdrop, it is captured in three-quarter view under uniform high-key lighting, showcasing the delicate grain and smooth of its finely-crafted surfaces."
}
```

Figure S2. An example of our structured description format for dataset generation.

```
prompt_1 = f"Two side-by-side images of the same object: {brief_description}"
prompt_2 = f"Left: {scene_description1}"
prompt_3 = f"Right: {scene_description2}"
prompt_image = f"{prompt_1}; {prompt_2}; {prompt_3}"
```

Figure S3. Our prompt template for paired image generation. The template combines a brief object description with two distinct scene descriptions to maintain subject consistency while varying environmental conditions.

B. Additional experimental results

B.1. Evaluation for subject-driven generation

Framework and criteria. To systematically evaluate subject-driven generation quality, we establish a framework with five criteria assessing both preservation of subject characteristics and accuracy of requested modifications:

- **Identity Preservation:** Evaluates preservation of essential identifying features (e.g., logos, brand marks, distinctive patterns)
- **Material Quality:** Assesses if material properties and surface characteristics are accurately represented
- **Color Fidelity:** Evaluates if colors remain consistent in regions not specified for modification
- **Natural Appearance:** Assesses if the generated image appears realistic and coherent
- **Modification Accuracy:** Verifies if the changes specified in the text prompt are properly executed

User studies. To further validate our approach, we conducted user studies collecting 375 valid responses. Participants evaluated the generated images across three key dimensions: identity consistency, text-image alignment, and visual coherence between subjects and backgrounds. The results shown in Figure S4 corroborate our quantitative findings, with our method achieving superior performance across all evaluation criteria.

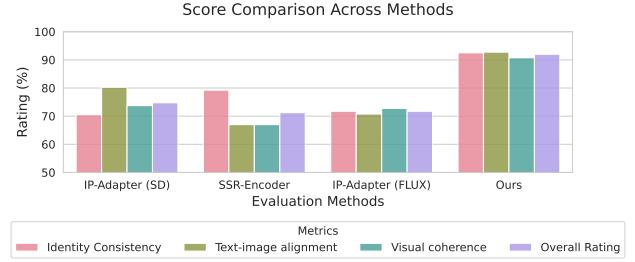


Figure S4. User study results comparing different methods across three metrics: identity consistency, text-image alignment, and visual coherence.

B.2. Additional generation results

We showcase more generation results from our method. Figure S5 presents additional results on the DreamBooth dataset, while Figure S6 demonstrates our method's effectiveness on other subject-driven generation tasks.

Method	Identity preservation	Material quality	Color fidelity	Natural appearance	Modification accuracy	Average score
Average over 5 random seeds						
IP-Adapter (SD 1.5)	29.4	86.1	45.3	97.9	17.0	55.1
SSR-Encoder	46.0	92.0	54.2	96.3	28.5	63.4
IP-Adapter (FLUX)	11.8	65.8	30.8	98.1	57.7	52.8
Ours	50.6	84.3	55.0	98.5	75.8	72.8
Best score over 5 random seeds						
IP-Adapter (SD 1.5)	56.3	98.9	70.1	99.7	37.2	72.5
SSR-Encoder	64.3	99.2	74.4	99.1	53.6	78.1
IP-Adapter (FLUX)	27.5	86.1	53.6	99.9	74.9	68.4
Ours	82.3	98.0	88.4	100.0	90.7	91.9

Table S1. Quantitative evaluation results (in percentage) across different evaluation criteria. Higher values indicate better performance.

Cases from Dreambooth dataset

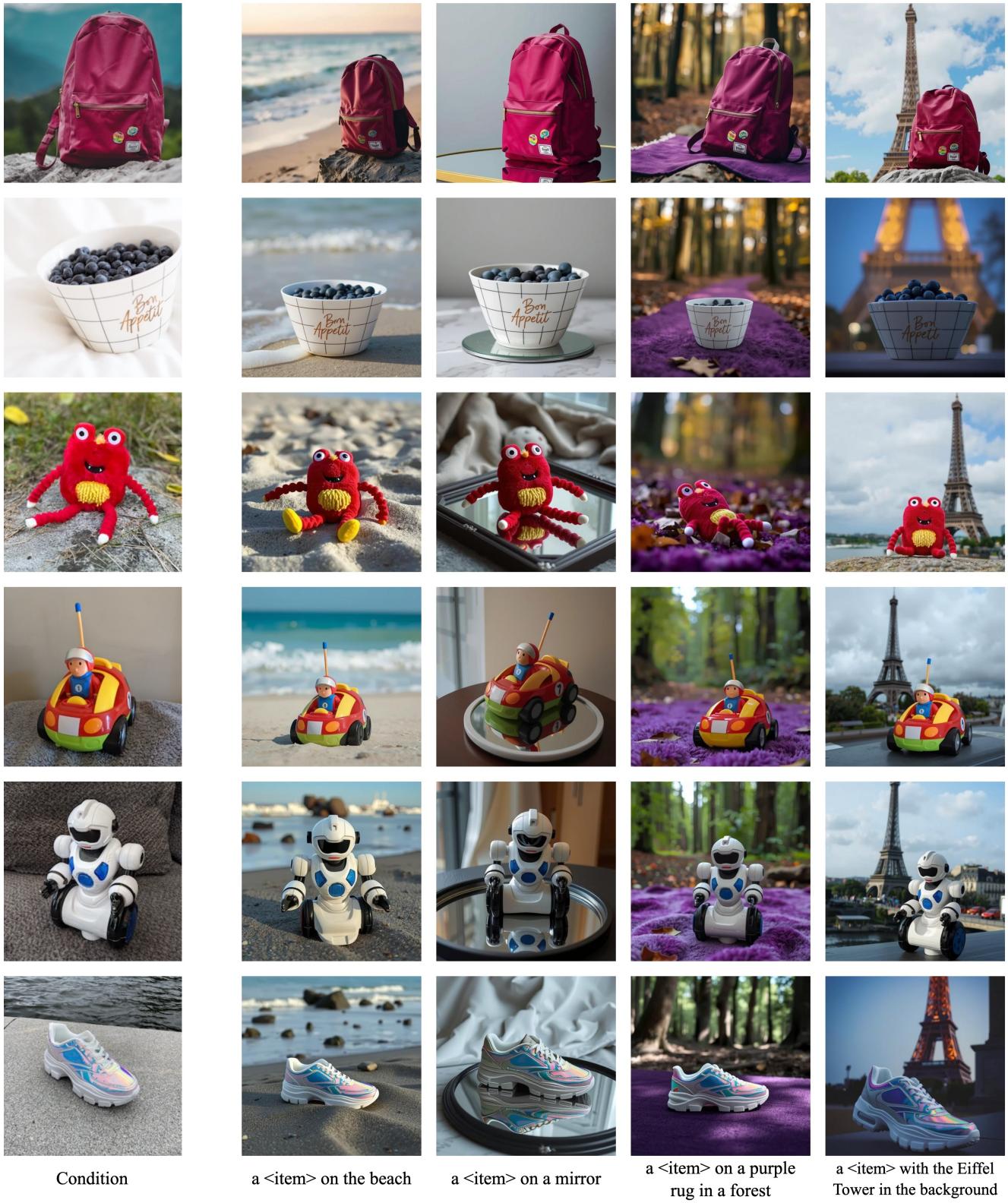


Figure S5. More results on Dreambooth dataset.

Scene Variation



Try On

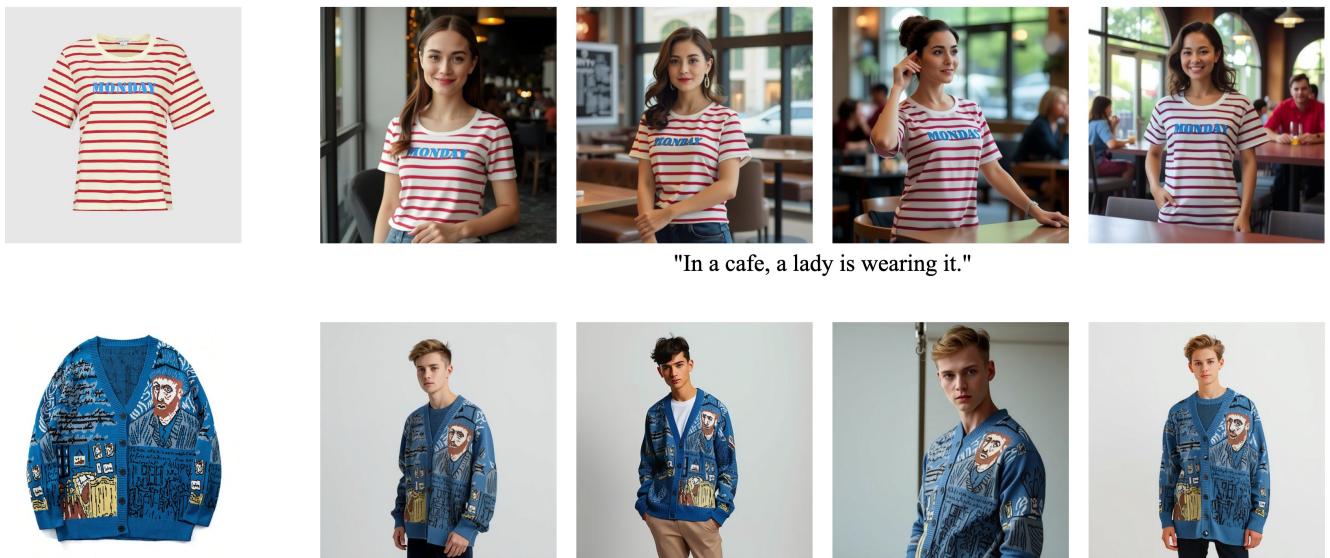


Figure S6. More results on other subject-driven generation tasks.