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 [23, 27, 28, 47]. Our approach leverages FLUX.1-dev [24] 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 A2 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 A3. 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 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 A1 presents representative examples from our quality-controlled dataset.

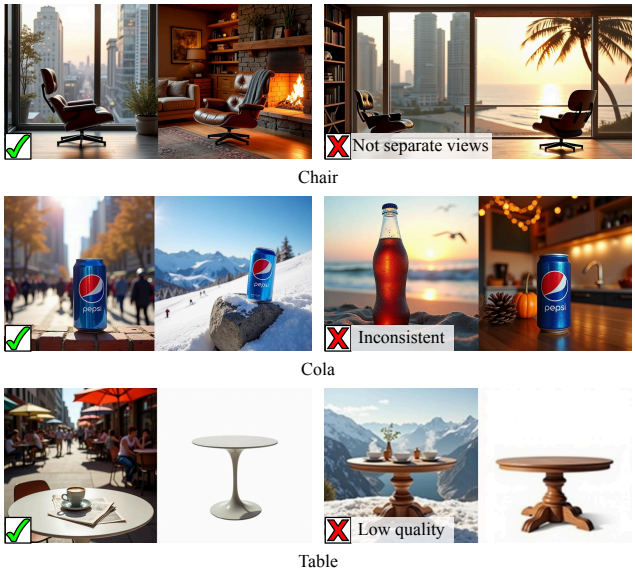

Figure A1. Examples of successful and failed generation results from Subjects200K dataset. Green checks indicate successful cases where subject identity and characteristics are well preserved, while red crosses show failure cases.

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 A2. 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 A3. 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.

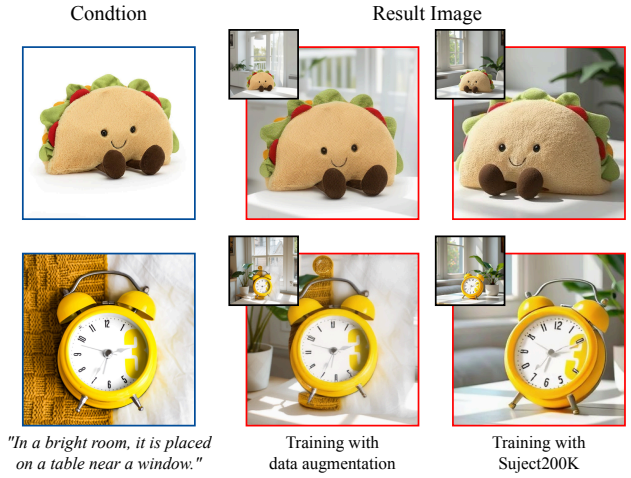

Figure A4. Comparison of models trained with different data. The model trained by data augmentation tends to copy inputs directly, while model trained by our Subjects200K generates novel views while preserving identity.

B. Additional experimental results

B.1. 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 3.3, 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 our Subjects200K dataset. As shown in Figure A4, 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.

B.2. 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 essen-

Setting	Controllability	Image quality						Alignment	
	F1↑	SSIM↑	CLIP-IQA↑	MAN-IQA↑	MUSICQ↑	PSNR↑	FID↓	CLIP-Text↑	CLIP-Image↑
OminiControl (FLUX.dev)	0.502	0.454	0.663	0.616	74.9	11.3	24.2	0.305	0.785
+ Shifted Position Encoding	0.488	0.408	0.675	0.615	75.9	11.8	23.6	0.302	0.768
+ Derect Addition	0.212	0.371	0.615	0.495	71.8	11.1	20.1	0.306	0.734
+ Derect Addition (Zero-gate)	0.224	0.384	0.624	0.501	73.3	11.2	21.1	0.304	0.746
OminiControl (SD3.5 medium)	0.386	0.377	0.656	0.527	73.8	9.6	21.5	0.312	0.770

Table A1. Evaluation on Canny-to-Image task.

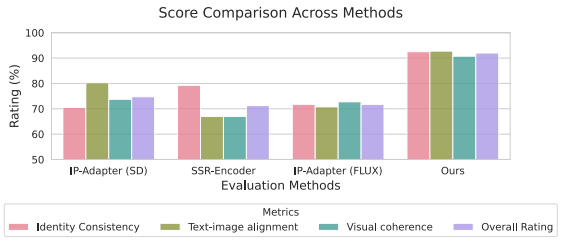

Figure A5. User study results comparing methods across identity consistency, text-image alignment, and visual coherence.

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 A2. Quantitative evaluation results (in percentage). Higher is better.

tial 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

The results of these evaluations are summarized in Table A2. We report both average scores across five random seeds and the best scores achieved by any seed. Our method outperforms baselines in most criteria, achieving the highest average score of 72.8% and the best score of 91.9% across all criteria.

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 A5 corroborate our quantitative findings, with our method achieving superior performance across all evaluation criteria.

B.3. Generality to other DiT models

We further demonstrate the general applicability of our approach by applying it to SD3.5-medium (2.6B), a representative DiT variant. The qualitative and quantitative evaluations, shown in Figure A6 and Table A1 respectively, confirm consistent improvements.

Figure A6. SD3.5 with OminiControl

B.4. Ablation on shifted position encoding

To investigate whether shifted Rotary Position Embedding (RoPE) influences model performance, we compare shifted and unshifted encoding variants in an ablation study (Table A1). Results show minimal differences between these two settings, suggesting that employing shifted RoPE does not substantially impact the quality of generated images in spatially aligned tasks.

B.5. Computational cost

To quantify the computational overhead introduced by unified sequence processing, we compare inference times against representative baselines (see Table A3). Our results indicate that although the unified approach does incur additional computational costs, these can be substantially mitigated through a key-value (KV) caching mechanism, which caches condition-token key/value pairs to avoid redundant computations. Further optimizations and refinements of this KV-cache strategy will be investigated in future research.

B.6. Evaluation of inpainting task

To better characterize our model's performance in the inpainting scenario, we separately evaluate reconstruction quality in the inpainted and non-inpainted regions. Table A4 provides a detailed breakdown, reporting not only

Setting	Extra parameters	Inference time (second)
IP-Adapter	918M / +7.6%	8.32
ControlNet	3.3B / +27.5%	9.02
OminiControl	14.5M / +0.1%	13.93
+ KV-cache	14.5M / +0.1%	8.62

Table A3. Efficiency comparison

overall SSIM and PSNR but also MSE focused explicitly on non-inpainted areas. Results confirm that OminiControl achieves 45–70% lower error compared to baselines, indicating enhanced fidelity for reconstructed details and superior preservation of unaltered regions.

B.7. Additional generation results

We showcase more generation results from our method. Figure A8 presents additional results on the DreamBooth dataset, while Figure A9 demonstrates our method's effectiveness on other subject-driven generation tasks.

Method	Overall			Inpainted region			Non-inpainted region
	MSE ↓	SSIM ↑	PSNR ↑	MSE ↓	SSIM ↑	PSNR ↑	MSE ↓
ControlNet	907	0.6798	18.9584	7588	0.2482	9.7815	219
Flux Tool	1087	0.7282	18.3723	6610	0.2267	10.409	122
OminiControl	860	0.7808	19.5898	6351	0.2554	10.697	66

Table A4. Breakdown of Inpainting Results.

(a) Spatially aligned tasks

(b) Subject driven-generation

(c) ID-reservation

Figure A7. More comparative results.

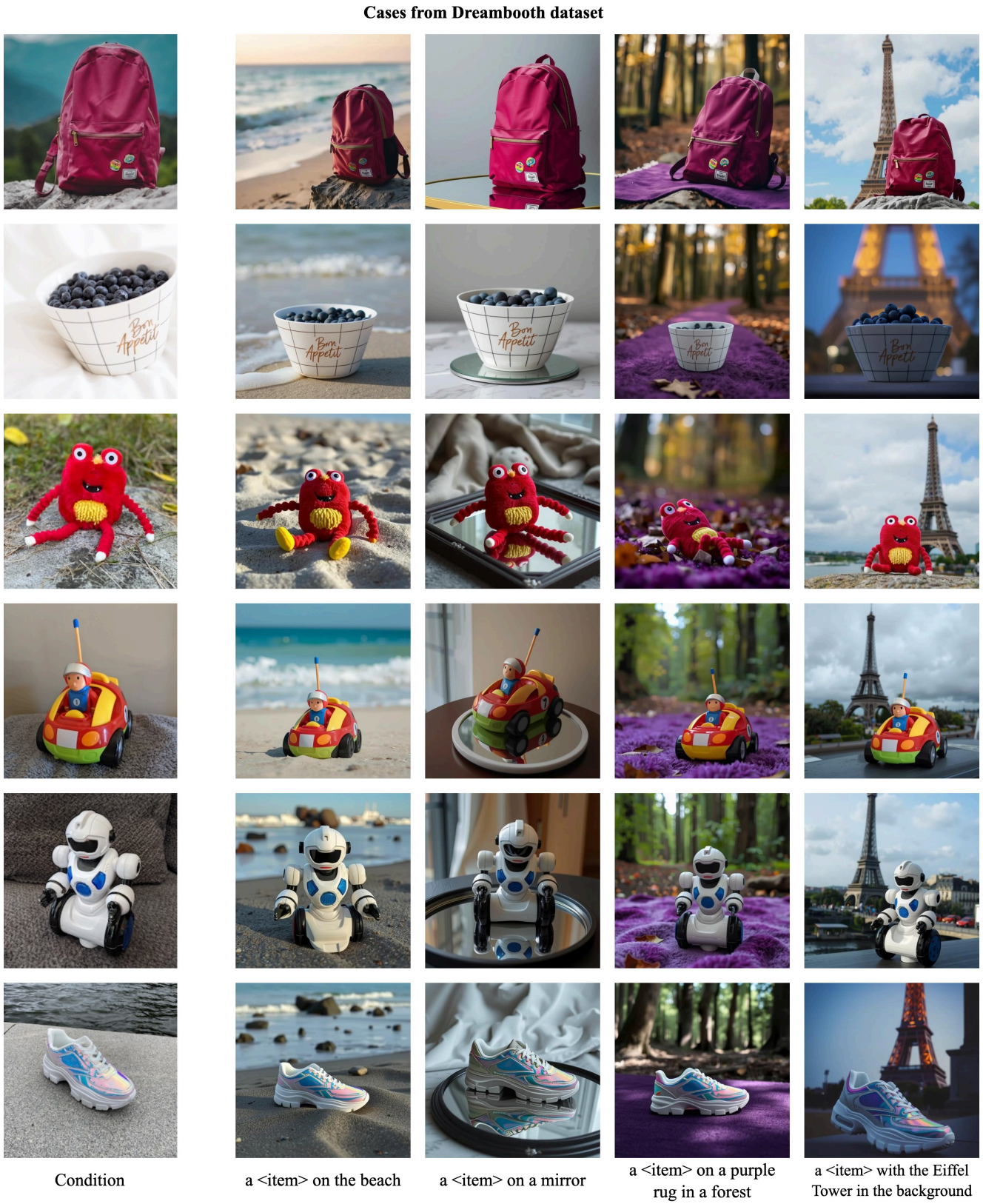

Figure A8. More results on Dreambooth dataset.

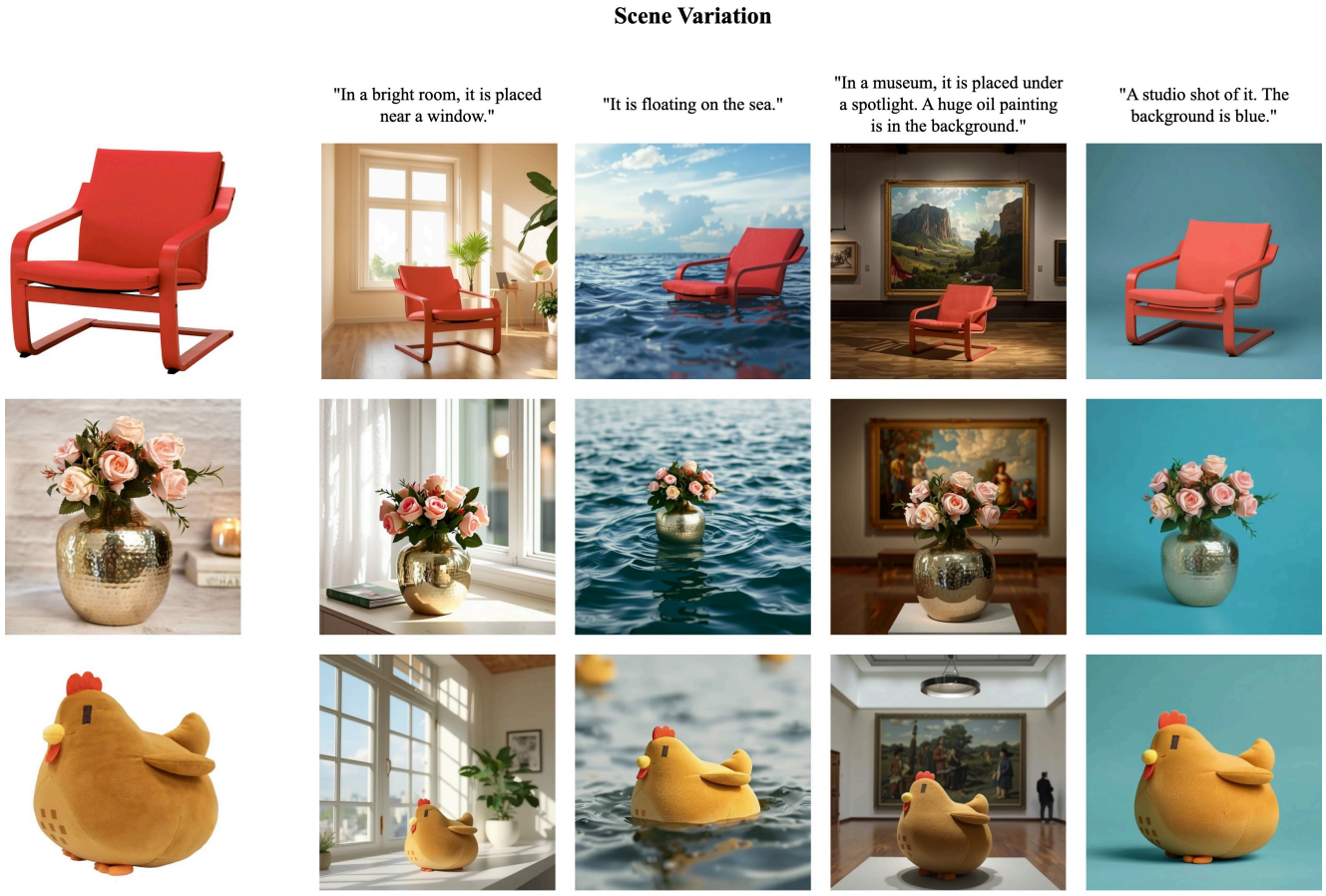

Figure A9. More results on other subject-driven generation tasks.