FreeCus: Free Lunch Subject-driven Customization in Diffusion Transformers

Supplementary Material

6. Experiments

Ablations on vital layer selection. We investigate: Does the benefit arise from simply reducing layers or specifically using vital layers? Do non-vital layers impact generation? Does attention-dropout [58] suffice?

Two ablations address this: 1) sharing attention in 10 random non-vital layers (ours-N; 10=N_v vital layers), and 2) sharing with random dropout in all 57 layers, dropping 5/6 to approximate $1-N_v/57$ (ours-D'). Other components remain unchanged. Results (Fig. 9) show key detail loss in both settings: ours-N alters hairstyle and removes leg features, while ours-D' shifts clothing color (purple $\rightarrow$ red). This confirms vital layers carry critical information. Non-vital layers also influence generation but contain excessive unimportant information—sharing all layers creates a copy-paste effect (fifth column in the Fig. 9).

Will stronger MLLMs improve our method? With ongoing advances in MLLMs, our method continues to improve. For example, upgrading from Qwen2-VL to Qwen2.5-VL reduces errors (highlighted in red) for rare subjects, as illustrated in Fig. 10.

Deeper discussion of artifact mitigation. We explored two spatial-level strategies: spatial masking (ours-M) and position index shifting of shared attention (ours-S). Both reduce artifacts but introduce trade-offs, as shown in Fig. 11, ours-S loses details and ours-M misaligns with reference subject's body geometry, lowering quality. We also tried randomly dropping half the shared attention, achieving the best balance and allowing adjustable dropout rates for controlled artifact reduction. Future work will explore adaptive dropout strategies to enhance generalization.

More qualitative samples. Fig. 12 illustrates that our method handles both human subjects (e.g., basketball player) and complex objects (e.g., camera with distinct features), as well as multiple and rare subjects (see in Fig. 10). While FreeCus is designed for single-subject customization, it can be extended to multi-subject scenes by tailoring the prompts fed into MLLMs.

Detailed quantitative results on each class. As shown in Tab. 3, our genuinely training-free method achieves state-of-the-art or comparable performance across all classes when benchmarked against approaches requiring additional training.

Prompt for detailed subject caption. The detailed subject descriptions, discussed in "Designs for captions" of Sec. 4.3, are generated by Qwen2-VL with specialized prompts as shown in Fig. 14.

Prompt for style transfer. For the style transfer task, the prompt fed to Qwen2-VL is "Describe this style briefly and precisely in max 20 words, focusing on its aesthetic qualities, visual elements, and distinctive artistic characteristics.".

Subsequently, the prompt fed to Qwen2.5 is "Please extract only the stylistic and artistic characteristics of the style from this description, removing any information about physical objects, specific subjects, narrative elements, or factual content. Focus solely on the aesthetic qualities, visual techniques, artistic movements, and distinctive style el-

Figure 9. **Ablations on vital layer selection.**

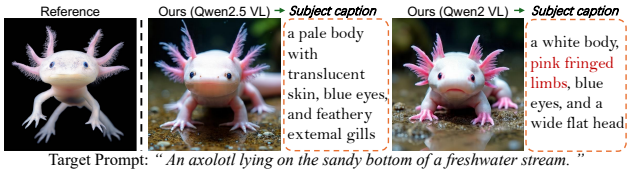

Figure 10. **Stronger MLLMs would yield better results.**

Figure 11. **Strategies to eliminate artifacts.**

Figure 12. **More qualitative samples.**

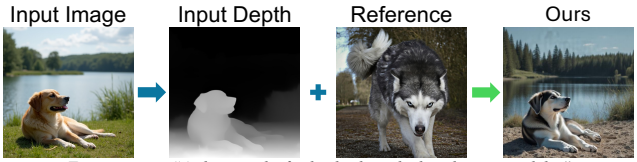

Figure 13. **Harmonizing with the control model to stabilize target structure.**

Method	CLIP-T ↑	CLIP-I ↑	DINO ↑
Qwen2VL-Flux	0.267	0.841	0.664
Ours+Qwen2VL-Flux	**0.274**	**0.853**	0.658

Table 2. **Quantitative results with and without our method integration in Qwen2VL-Flux framework.**

Method	BaseModel	Animal			Human			Object			Averaged		
		CLIP-T ↑	CLIP-I ↑	DINO ↑	CLIP-T ↑	CLIP-I ↑	DINO ↑	CLIP-T ↑	CLIP-I ↑	DINO ↑	CLIP-T ↑	CLIP-I ↑	DINO ↑
Textual Inversion†	SD v1.5	0.314	0.784	0.537	0.281	0.645	0.322	0.297	0.709	0.412	0.298	0.713	0.430
DreamBooth†	SD v1.5	0.322	0.817	0.655	0.322	0.561	0.253	0.323	0.770	0.568	0.322	0.716	0.505
DreamBooth-L†	SDXL v1.0	0.342	0.840	0.724	0.339	0.623	0.316	0.343	0.791	0.602	0.341	0.751	0.547
BLIP-Diffusion	SD v1.5	0.304	0.857	0.692	0.236	0.763	0.567	0.286	0.827	0.658	0.276	0.815	0.639
Emu2	SDXL v1.0	0.315	0.812	0.621	0.284	0.736	0.476	0.316	0.742	0.490	0.305	0.763	0.529
IP-Adapter	SDXL v1.0	0.314	0.892	0.719	0.292	0.784	0.479	0.307	0.859	0.665	0.305	0.845	0.621
IP-Adapter-Plus	SDXL v1.0	0.293	0.939	0.840	0.236	0.890	0.747	0.283	0.919	0.834	0.271	0.916	0.807
MS-Diffusion	SDXL v1.0	0.344	0.925	0.816	0.322	0.810	0.629	0.342	0.885	0.741	0.336	0.873	0.729
Qwen2VL-Flux	FLUX.1	0.287	0.902	0.704	0.232	0.779	0.669	0.283	0.842	0.619	0.267	0.841	0.664
IP-Adapter	FLUX.1	0.325	0.898	0.700	0.285	0.786	0.633	0.332	0.836	0.581	0.314	0.840	0.638
OminiControl	FLUX.1	0.336	0.869	0.656	0.323	0.693	0.439	0.331	0.829	0.615	0.330	0.797	0.570
Ours	FLUX.1	**0.328**	**0.902**	**0.738**	**0.276**	**0.788**	**0.675**	**0.321**	**0.869**	**0.677**	**0.308**	**0.853**	**0.696**

Table 3. **Quantitative evaluation results for each class.** Blue indicates scores higher than ours, and † denotes optimization-based methods.

ements. Return only the extracted style description without any additional commentary. The description is: { [output from Qwen2-VL] }".

Quantitative results with and without our method integration in DiT-based framework. As shown in Tab. 2, compared to the original Qwen2VL-Flux, our method combined with it achieves higher scores on two metrics, further demonstrating the compatibility and orthogonality of *FreeCus* with other DiT-based models.

Subject-driven layout-guidance generation. As illustrated in Fig. 13, our method also supports layout-guided synthesis when integrated with the Flux.1-Depth-dev model.

7. Compared Methods and Implementation Details

IP-Adapter (IPA) [65] IPA introduces a lightweight adapter that decouples image and text features, addressing limitations in fine-grained control when merging these features in cross-attention layers. For IPA (Flux.1) implementation, we use the third-party code from XLabs-AI.

MS-Diffusion (MS-D) [62] MS-D incorporates grounding tokens with feature resampling to preserve subject detail fidelity. It requires inputting a bounding box for layout guidance; we set the default box values to [0.25, 0.25, 0.75, 0.75].

Qwen2VL-Flux (QVL-Flux) [37] QVL-Flux replaces Flux's conventional T5-XXL text encoder with a vision-language model, enabling image-to-image generation. We utilize the official repository and weights to generate 1024 × 1024 images.

Textual Inveresion (TI) [17] TI updates only the new token embedding representing the novel subject while keeping all other parameters frozen. Experimental results are from the DreamBench++ [44] implementation.

DreamBooth [50] DreamBooth updates all layers of the T2I model to maintain visual fidelity and employs prior preservation loss to prevent language drift. DreamBooth-Lora only updates additional lora adapters. Experimental results are from the DreamBench++ [44] implementation.

BLIP-Diffusion (BLIP-D) [34] BLIP-D leverages the pretrained BLIP-2 multimodal encoder to create multiple learnable embeddings representing input subject features, then fine-tunes the base model to adapt these embeddings for personalization. Experimental results are from the DreamBench++ [44] implementation.

Emu2 [54] Emu2 employs an autoregressive approach to process multimodal information with a predict-the-next-element objective. Images are tokenized via a visual encoder and interleaved with text tokens, enabling straightforward customization with target text. Experimental results are from the DreamBench++ [44] implementation.

OminiControl [55] OminiControl performs multiple image-to-image tasks using a unified sequence processing strategy and dynamic position encoding, introducing only lightweight trainable LoRA parameters. We reproduced results using the official repository.

Prompt for Detailed Subject Caption

```
[Task Description]
As an experienced image analyst, your
    task is to provide a detailed
    description of the main features
    and characteristics of the given
    {} in this image according to the
    following criteria.

[Feature Analysis Criteria]
Analyze and describe the following
    visual elements:
1. Shape
- Main body outline
- Overall structure
- Proportions and composition
- Spatial organization

2. Color
- Color palette and schemes
- Saturation levels
- Brightness/contrast
- Color distribution patterns
```

```
3. Texture
- Surface qualities
- Detail clarity
- Visual patterns
- Material appearance

4. Subject-Specific Features
- If human/animal: facial features,
    expressions, poses
- If object: distinctive
    characteristics, condition
- If landscape: environmental elements
    , atmosphere

[Description Quality Levels]
Your description should aim for the
    highest level of detail:
Level 1: Basic identification of main
    elements
Level 2: Description of obvious
    features
Level 3: Detailed analysis of multiple
     characteristics
Level 4: Comprehensive analysis with
    subtle details

[Output Format]
Please provide your analysis in the
    following structure:

Main Subject: [Brief identifier]
Primary Features:
- Shape: [Description]
- Color: [Description]
- Texture: [Description]
- Subject-Specific Details: [
    Description]
Overall Composition: [Brief summary]
```

Figure 14. Prompt for Detailed Subject Caption.