ANIMALBOOTH: MULTIMODAL FEATURE ENHANCEMENT FOR ANIMAL SUBJECT PERSONALIZATION

Chen Liu & Haitao Wu
College of Intelligence and Computing
Tianjin University
Tianjin, China
{ttjy, wuhaitao}@tju.edu.cn

Kafeng Wang
Department of Computer Science and Technology
Tsinghua University
Beijing, China
wangkafeng@tsinghua.edu.cn

ABSTRACT

Personalized animal image generation is challenging due to rich appearance cues and large morphological variability. Existing approaches often exhibit feature misalignment across domains, which leads to identity drift. We present *AnimalBooth*, an inference-time tuning-free framework that strengthens identity preservation with an *Animal-Net* and an adaptive attention module, mitigating cross-domain alignment errors. We further introduce a frequency-controlled feature integration module that applies Discrete Cosine Transform filtering in the latent space to guide the diffusion process, enabling a coarse-to-fine progression from global structure to detailed texture. To advance research in this area, we curate *AnimalBench*, a high-resolution dataset for animal personalization. Extensive experiments show that *AnimalBooth* consistently outperforms strong baselines on multiple benchmarks with superior efficiency, improving both identity fidelity and perceptual quality. The code and dataset will be made publicly available in the future.

1 INTRODUCTION

Personalized multimodal generation is a prominent yet challenging subfield that aims to synthesize images conforming to both textual descriptions (text–image consistency) and the intrinsic characteristics of custom concepts (identity consistency) (Deng et al., 2025; Dai et al., 2025; Zhang et al., 2024b; Shen et al., 2025a). This paradigm shows strong potential across diverse applications ranging from creative artistry to product design (Shen et al., 2025d;c). However, accurately capturing the identity of animal subjects, which exhibit complex visual features such as fine-grained fur textures and non-rigid morphologies, remains a significant hurdle (Xu et al., 2025; Shen & Tang, 2024).

Animal personalization is uniquely challenging compared to human subject personalization due to several factors: (1) **Non-rigid deformations**: animals exhibit diverse poses and body configurations that vary dramatically across species; (2) **Anatomical variations**: skeletal structures and body proportions differ significantly between species (e.g., felines vs. bovines); (3) **Intricate texture requirements**: fur patterns, scales, and feathers require fine-grained preservation that general-purpose models often fail to maintain. As demonstrated in recent work (Xu et al., 2025), general-purpose models like DiT-based architectures frequently suffer from identity drift in these challenging scenarios.

Existing personalization methodologies can be broadly categorized by their feature extraction and optimization strategies. The first category comprises optimization-based methods that achieve high fidelity by fine-tuning model components. A representative example is DreamBooth (Ruiz et al., 2022), which fine-tunes the UNet backbone, while Textual Inversion (Gal et al., 2022) instead optimizes word embeddings. This paradigm has been extended to support multiple customized con-

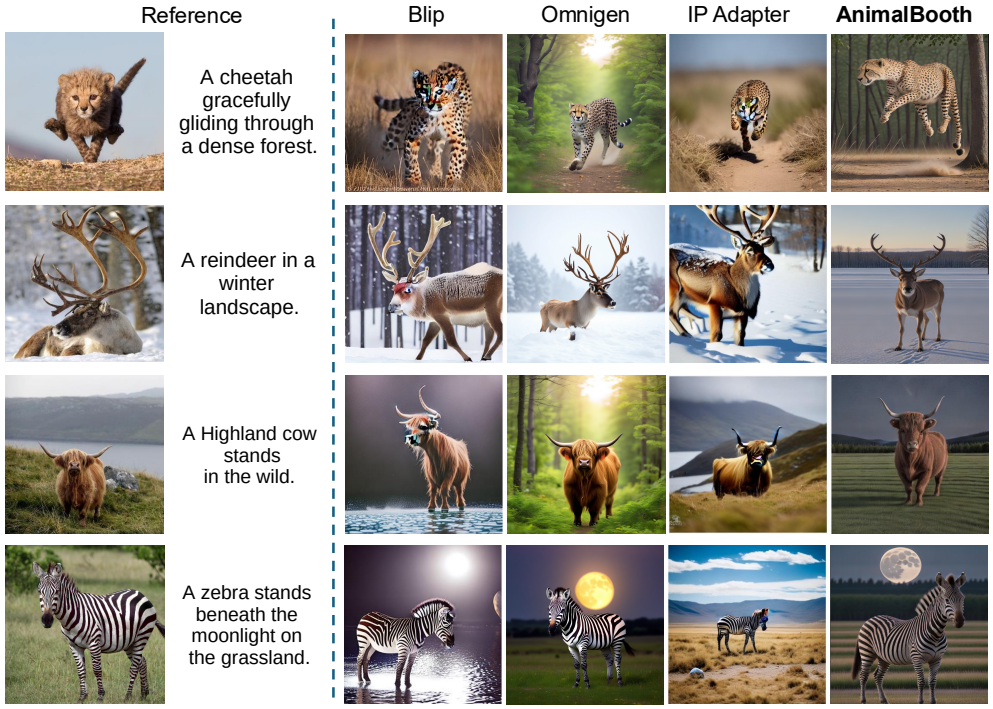

Figure 1: Qualitative comparison with SOTA methods.

cepts (Kumari et al., 2023; Ma et al., 2024; Wang et al., 2025), including applications such as virtual dressing and editing (Shen et al., 2025b). Despite their fidelity, such approaches demand substantial computational resources and often exhibit an inherent trade-off between fidelity and diversity due to overfitting. The second category, encoder-based and tuning-free methods, offers a more efficient alternative. For instance, IP-Adapter (Ye et al., 2023) and its successors (Zhang et al., 2024a) utilize frozen external encoders (e.g., CLIP) to inject identity features. While efficient, their performance is fundamentally limited by the representational capacity of the encoder, particularly when modeling animals with unique identity cues such as body structures and fur patterns. Consequently, identity distortion and detail loss are frequent. Other directions include retrieval-augmented generation (Chen et al., 2022) or LLM-enhanced frameworks (Zeng et al., 2024; Pan et al., 2023; Sun et al., 2024), yet the challenge of inadequate domain-specific representation remains.

Motivated by bridging this domain gap, we propose **AnimalBooth**, a feature-enhanced, **inference-time tuning-free** personalized generation framework specifically tailored for animals. To systematically address these identified challenges, our framework incorporates several targeted mechanisms: (1) To handle *non-rigid deformations*, we introduce a frequency-controlled feature integration module that leverages low-frequency signals to guide coarse structural consistency; (2) To overcome *encoder capacity* limits and *identity drift*, we design a dedicated Animal-Net with a Q-Former bottleneck to filter noise and capture subject-specific semantic features; (3) To maintain *text-control* while injecting identity, we employ a dual-path adaptive attention mechanism.

Furthermore, existing datasets such as AFHQ (Choi et al., 2020) or MS-COCO are primarily designed for classification, translation, or general captioning, and lack the high-resolution subject masks and fine-grained captions necessary for precise animal personalization. To address this, we construct **AnimalBench**, a curated high-definition dataset for animal personalization. Our contributions are threefold: (1) We propose a dedicated branch architecture comprising a lightweight Animal-Net and a novel dual-path adaptive attention mechanism, effectively mitigating identity distortion while preserving generative capacity. (2) We design a frequency-controlled feature integration module leveraging DCT filtering in latent space to enhance attribute manipulation and texture fidelity. (3) We establish and release AnimalBench, a high-definition dataset for animal personalization, on which our method achieves state-of-the-art performance while being significantly more efficient than large-scale DiT-based models.

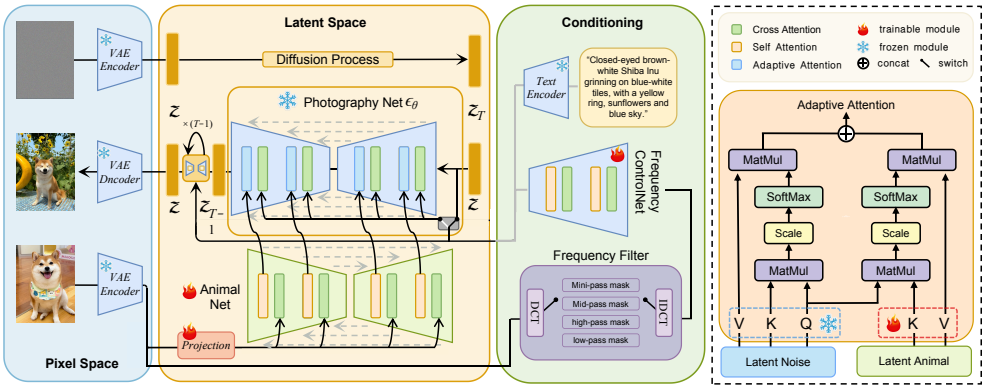

Figure 2: AnimalBooth primarily consists of a trainable Animal-Net and a frozen Photography-Net. The Animal-Net incorporates an Adaptive Attention module for efficient identity feature injection and a Frequency-Controlled module (based on ControlNet (Zhang et al., 2023)) for enhanced control over visual attributes like structure and texture, while the Photography-Net integrates these features with text prompts within the latent space.

Figure 3: Examples of AnimalBench dataset. Each instance provides five components: (a) a high-quality source image, (b) a detailed caption describing the scene and subject, (c) a precise pixel-level semantic segmentation mask of the primary subject, (d) the resulting masked subject for visual verification, and (e) the corresponding paired image.

2 METHODOLOGY

2.1 OVERALL ARCHITECTURE

As depicted in Fig. 2, AnimalBooth effectively integrates a trainable Animal-Net with a frozen Photography-Net. The Animal-Net directly refines the complex features and flexible morphologies of animals within the feature space, effectively bypassing the complexities of cross-domain alignment. Through adaptive attention, identity features are efficiently injected. The incorporation of a frequency-controlled module based on ControlNet (Zhang et al., 2023) further enhances fine-grained control over visual attributes such as the structure and textural details of the generated images. The frozen Photography-Net then merges these extracted fine-grained features with text prompts in the latent space.

2.2 ANIMAL-NET

In the task of personalized animal image generation, precise extraction of fine-grained animal identity features is paramount for maintaining consistency between generated images and reference an-

imals. To achieve this, we propose a specialized Animal-Net capable of simultaneously capturing both semantic information and high-frequency textural features of animals. Specifically, given an animal reference image $\mathcal{X}_a \in \mathbb{R}^{3\times H\times W}$, we first transform it into a latent space representation $\mathbf{L}_a \in \mathbb{R}^{4\times \frac{H}{8}\times \frac{W}{8}}$ using a frozen VAE encoder (Rombach et al., 2022). We then extract token embeddings from $\mathcal{X}_a$ using a frozen CLIP image encoder (Radford et al., 2021) and a trainable projection layer.

Why Q-Former? We employ Q-Former (Li et al., 2023b) as the projection layer due to its superior ability to act as a semantic bottleneck that effectively filters background noise while capturing intricate animal textures via dynamic attention. Unlike static mapping approaches (e.g., feature concatenation or MLP adapters), Q-Former's learnable query tokens dynamically attend to relevant identity features. As shown in Table 1, Q-Former significantly outperforms simpler alternatives in identity fidelity metrics.

Subsequently, animal features within the Animal-Net interact extensively through a cross-attention mechanism, similar to the interaction between text and image features in the original Text-to-Image (T2I) model (Rombach et al., 2022). This interaction ensures a deep fusion of semantic and textural features. Finally, the output of the Animal-Net is aligned in parallel with the Photography-Net, injecting these fine-grained animal identity features into the Photography-Net via an adaptive attention module. Notably, the Animal-Net is solely used for encoding reference images; therefore, during the diffusion process, no noise is added to the reference image, and it undergoes only a single forward pass.

2.3 Adaptive Attention Module

To enable personalized animal image generation, the Photography-Net must possess both its original generative capabilities and the ability to fuse animal identity features. We freeze the core modules of the Photography-Net to preserve the former, and achieve the latter through an adaptive attention module. The architecture of the Photography-Net in AnimalBooth builds upon SD v1.5 (Rombach et al., 2022), with all self-attention modules replaced by adaptive attention modules.

Dual-Path Architecture. Unlike standard single-path cross-attention strategies commonly used in subject-driven generation, our adaptive attention module employs a novel **dual-path architecture**: a frozen self-attention path (to preserve the original generative capacity of Stable Diffusion) and a trainable cross-attention path (to inject identity features), both sharing a single Query matrix $\mathbf{Q}$. This shared-query design enables seamless feature fusion while maintaining the pre-trained model's semantic understanding. As shown in Fig. 2, an adaptive attention module consists of a frozen self-attention module and a learnable cross-attention module. Its self-attention weights are initialized from SD v1.5 to retain generative capacity. Given the query features $\mathbf{Z}_n$ from the Photography-Net and the animal identity features $\mathbf{F}_a$ from the Animal-Net, the output $\mathbf{O}_h$ of the adaptive attention module is defined as:

$$\mathbf{O}_h = \underbrace{\text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right)\mathbf{V}}_{\text{Frozen Self-attention}} + \lambda \underbrace{\text{Softmax}\left(\frac{\mathbf{Q}\left(\mathbf{K}_{ID}\right)^\top}{\sqrt{d}}\right)\mathbf{V}_{ID}}_{\text{Trainable Cross-attention}}, \tag{1}$$

where $\lambda \in [0, 1]$ is a dynamic coefficient that controls the strength of the animal identity feature condition, enabling flexible balance between identity preservation and creative generation. $\mathbf{Q}$, $\mathbf{K}$, $\mathbf{V}$ are derived from $\mathbf{Z}_n$, while $\mathbf{K}_{ID} = \mathbf{F}_a\mathbf{W}_{k_{ID}}$ and $\mathbf{V}_{ID} = \mathbf{F}_a\mathbf{W}_{v_{ID}}$ are derived from $\mathbf{F}_a$. The self-attention part is frozen, while the cross-attention part (i.e., $\mathbf{W}_{k_{ID}}$ and $\mathbf{W}_{v_{ID}}$) is trainable. This design effectively injects animal identity features while preserving the original T2I model's (Rombach et al., 2022) generative capabilities.

2.4 Frequency-Controlled Feature Integration Module

This module guides the diffusion process in the latent space through Discrete Cosine Transform (DCT) filtering, implemented via ControlNet (Zhang et al., 2023) (Fig. 2). First, a channel-wise 2D DCT is applied to the source domain latent features $\mathbf{L}_0$ (obtained from the VAE encoder) to acquire

the frequency domain representation $\mathbf{F}_{DCT}$:

$$\mathbf{F}^{(n)}_{DCT,u,v} = \frac{2}{\sqrt{hw}} m(u)m(v) \sum_{i=0}^{h-1} \sum_{j=0}^{w-1} [(\mathbf{L}_0^{(n)})_{i,j} \cos\left(\frac{(2i+1)u\pi}{2h}\right) \cos\left(\frac{(2j+1)v\pi}{2w}\right)], \tag{2}$$

where $m(0) = \frac{1}{\sqrt{2}}$ and $m(\gamma) = 1$ for $(\gamma > 0)$.

Different frequency bands of the DCT spectrum encode distinct visual attributes: low-frequency components capture global structure and identity-related morphology, while high-frequency components encode fine textures and edges. We design four types of DCT filters (masks) for mini-pass, low-pass, mid-pass, and high-pass filtering to manipulate visual properties from coarse structures to fine textures.

$$\begin{cases} Mask_{mini}(u,v) = 1 \ \ if \ \ u+v \leq 10 \ \ else \ \ 0, \\ Mask_{low}(u,v) = 1 \ \ if \ \ u+v \leq 20 \ \ else \ \ 0, \\ Mask_{mid}(u,v) = 1 \ \ if \ \ 20 < u+v \leq 40 \ \ else \ \ 0, \\ Mask_{high}(u,v) = 1 \ \ if \ \ u+v \geq 50 \ \ else \ \ 0. \end{cases} \tag{3}$$

These filters are multiplied by $\mathbf{F}_{DCT}$ to extract features in specific frequency bands, denoted as $\mathbf{F}_{filtered} = \mathbf{F}_{DCT} \times Mask_*$. Finally, a 2D Inverse DCT (IDCT) is applied to convert $\mathbf{F}_{filtered}$ back to the spatial domain, yielding the control signal $\mathbf{C}_{freq}$:

$$\mathbf{C}^{(n)}_{\text{freq},\,i,j} = \frac{2}{\sqrt{hw}} \sum_{u=0}^{h-1} \sum_{v=0}^{w-1} m(u)\, m(v)\, \left(\mathbf{F}^{(n)}_{\text{filtered}}\right)_{u,v} \times \cos\left(\frac{(2i+1)u\pi}{2h}\right) \cos\left(\frac{(2j+1)v\pi}{2w}\right). \tag{4}$$

The $\mathbf{C}_{freq}$ signals obtained from mini-pass, low-pass, mid-pass, and high-pass filtering, respectively, control the texture, texture and structure, layout, and contour consistency between the generated and reference images. As our experiments demonstrate (Table 4), low-pass filtering achieves optimal identity preservation by retaining the structure and morphology information critical for animal recognition.

2.5 Training and Objectives

AnimalBooth training is bifurcated into two complementary stages that follow a coarse-to-fine progression:

Stage 1: Identity Learning. The Animal-Net and Projection modules are trained to capture global subject identity. The objective function aims to minimize the Mean Squared Error between the predicted noise and the ground truth noise, while incorporating both text conditions $\mathbf{C}_t$ and animal identity features $\mathbf{C}_a$:

$$L_{stage1} = \mathbb{E}_{\mathbf{z}_t, \epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I}), \mathbf{C}_t, \mathbf{C}_a, t} \left\| \epsilon_\theta \left(\mathbf{z}_t, \mathbf{C}_t, \mathbf{C}_a, t\right) - \epsilon_t \right\|^2. \tag{5}$$

Stage 2: Texture Enhancement. The frequency-controlled ControlNet module is trained to enhance fine textures and structural details. This decoupled training strategy avoids convergence difficulties that arise from jointly optimizing identity and texture objectives:

$$L_{stage2} = \mathbb{E}_{\mathbf{z}_0, t, \mathbf{C}_t, \mathbf{C}_{freq}, \epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I})} \left\| \epsilon - \epsilon_\theta(\mathbf{z}_t, t, \mathbf{C}_t, \mathbf{C}_{freq}) \right\|_2^2, \tag{6}$$

where $\mathbf{z}_t$ represents the noisy latent features at time step t, $\mathbf{C}_t$ is the text condition, and $\mathbf{C}_{freq}$ is computed by the frequency-controlled feature integration module based on the source domain latent features $\mathbf{L}_0$. The parameters of ControlNet are trained using four distinct DCT filters $(Mask_{mini}, Mask_{low}, Mask_{mid}, Mask_{high})$.

During inference, classifier-free guidance (Ho & Salimans) is employed to balance conditional and unconditional generation:

$$\hat{\epsilon}_\theta(\mathbf{x}_t, \mathbf{C}_t, \mathbf{C}_a, \mathbf{C}_{freq}, t) = w\epsilon_\theta(\mathbf{x}_t, \mathbf{C}_t, \mathbf{C}_a, \mathbf{C}_{freq}, t) + (1-w)\epsilon_\theta(\mathbf{x}_t, t), \tag{7}$$

Table 1: Ablation on projection layer.

Projection	LPIPS ↓	DINO ↑	CLIP-T ↑	CLIP-I ↑
Concatenation	57.41	31.05	19.85	70.58
MLP Adapter	56.81	31.68	20.02	71.00
Q-Former (Ours)	**49.08**	**75.66**	**20.73**	**90.00**

Table 2: Ablation on guidance scale w.

w	LPIPS ↓	DINO ↑	CLIP-T ↑	CLIP-I ↑
2.0	54.03	43.95	18.92	76.52
5.0	51.26	68.34	20.15	85.67
7.5 (Ours)	**49.08**	**75.66**	**20.73**	**90.00**
10.0	50.12	72.45	20.58	88.34

Table 3: Efficiency comparison on NVIDIA A100.

Method	Time (s) ↓	VRAM (GB) ↓	Params (B)
Omnigen (Xiao et al., 2025)	140.0	10.4	3.8
Flux (DiT) (Chang et al., 2024)	85.0	12.8	12.0
AnimalBooth	**5.5**	**0.7**	**1.2**

where w is the guidance scale. We empirically set $w = 7.5$ in our experiments (see Table 2 for ablation).

3 Experiments

3.1 Experimental Setup

We utilize Stable Diffusion v1.5 (Rombach et al., 2022) as the pre-trained Latent Diffusion Model (LDM) and train it for personalized animal image generation. To comprehensively evaluate the model's capabilities in generating high-definition animal images, we constructed a specialized AnimalBench dataset, comprising 10,958 training images and 1,000 test images. Each entry in the dataset consists of a high-definition, single-subject animal image–text pair, with examples depicted in Fig. 3. Training was conducted on a server equipped with eight NVIDIA A100-SXM4-80GB GPUs. We employed the AdamW optimizer with an initial learning rate of 1e-5 and a batch size of 4. The model was trained at a resolution of 512×512 pixels. During the inference phase, we used the DDIM sampler (Ho et al., 2020) for 50 sampling steps with guidance scale $w = 7.5$.

3.2 Qualitative Results

As shown in Fig. 1, methods like BLIP-Diffusion (Li et al., 2023a), Omnigen (Xiao et al., 2025) and IP-Adapter (Ye et al., 2023) struggle with maintaining fidelity to the original animal's identity. For instance, in the "cheetah" example, BLIP-Diffusion introduces noticeable distortions in the cheetah's facial features and fur patterns. Omnigen, while better, still presents a cheetah that looks somewhat less dynamic and natural compared to the reference. IP-Adapter's cheetah appears to have some color inaccuracies and a less refined texture. Similar issues are observable in the other examples. For the "reindeer," these methods often fail to capture the subtle nuances of its fur, antler structure, or the serene expression seen in the reference, sometimes introducing unnatural postures or color shifts. The "Highland cow" examples highlight a lack of accurate texture and the distinct shagginess that characterizes the breed. Finally, the "zebra" images from these comparative methods often fall short in reproducing the sharp stripe patterns, the reflective quality of its fur under moonlight, or the natural stance, sometimes resulting in blurry details or less vibrant contrasts. AnimalBooth consistently demonstrates superior performance in preserving both coarse structures and fine-grained textures.

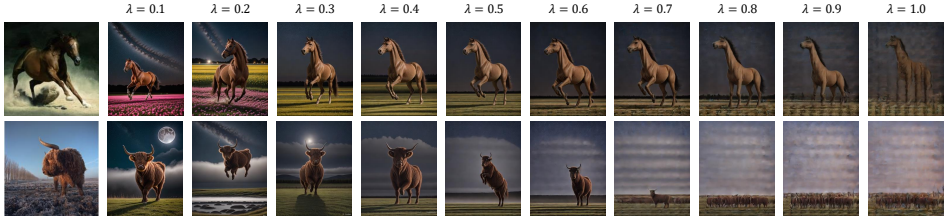

Figure 4: Example results with different animal image strength λ.

Table 4: Ablation study on frequency configurations.

Method	LPIPS ↓	DINO ↑	CLIP-T ↑	CLIP-I ↑
w/o Freq Cond	69.40	58.52	**21.37**	78.17
High-Pass	56.72	64.72	21.21	80.02
Mid-Pass	61.66	66.97	20.84	82.79
Mini-Pass	68.21	59.97	21.20	80.89
Low-Pass (Ours)	**49.08**	**75.66**	20.73	**90.00**

Table 5: Quantitative comparison with state-of-the-art methods.

Method	LPIPS ↓	DINO ↑	CLIP-T ↑	CLIP-I ↑
Textual Inv. (Gal et al., 2022)	72.35	48.92	18.76	71.23
BLIP (Li et al., 2023a)	68.21	62.96	19.68	82.38
Omnigen (Xiao et al., 2025)	71.68	50.05	19.41	72.95
IP-Adapter (Ye et al., 2023)	62.91	72.88	19.39	89.75
Flux (DiT) (Chang et al., 2024)	65.47	55.31	19.52	78.64
AnimalBooth	**49.08**	**75.66**	**20.73**	**90.00**

3.3 Quantitative Results

As presented in Table 5, AnimalBooth obtained a CLIP-T (Radford et al., 2021) score of 20.73, which is notably higher than the next best method, BLIP-Diffusion (Li et al., 2023a), at 19.68. This indicates AnimalBooth's superior ability to comprehend and capture textual semantics, generating animal images that are highly consistent with their descriptions. Concurrently, AnimalBooth achieved the highest DINO (Caron et al., 2021) score of 75.66, compared to IP-Adapter's (Ye et al., 2023) 72.88. This highlights AnimalBooth's exceptional capability in capturing image details and realism, effectively preserving the intricate textures and fine structures of animals. Furthermore, AnimalBooth also achieved leading scores of 90.00 and 49.08 for CLIP-I and LPIPS, respectively, comprehensively outperforming other methods.

Comparison with Optimization-based and DiT Methods. Table 5 also compares AnimalBooth with Textual Inversion (Gal et al., 2022) (a representative optimization-based method) and Flux (Chang et al., 2024) (a state-of-the-art DiT-based model). Textual Inversion, despite requiring per-subject optimization, achieves suboptimal results due to limited embedding capacity for complex animal features. Flux, while generating high-quality images, suffers from identity drift when personalizing animal subjects without specialized adaptation. AnimalBooth outperforms both methods across all metrics while requiring no per-subject optimization at inference time.

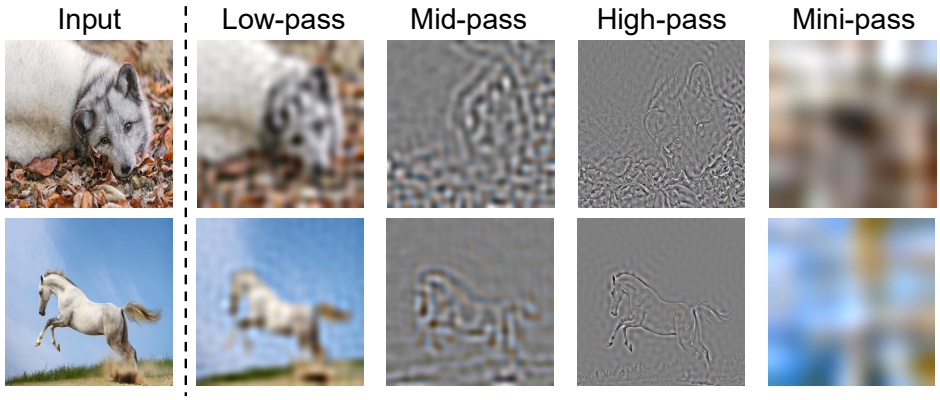

Figure 5: Output results of different frequency filter masks.

3.4 Ablation Study

Frequency Configuration. We evaluate the impact of five distinct frequency configuration modes on generated image quality: Low-Pass, Mini-Pass, Mid-Pass, High-Pass, and without frequency conditioning. As presented in Table 4, the Low-Pass frequency configuration significantly outperforms other configurations on CLIP-I, DINO and LPIPS metrics, indicating its superior performance in preserving animal identity details such as fur textures, patterns, and coat colors. This is because low-frequency components capture the global structure and morphology essential for identity recognition, while high-frequency details can be effectively reconstructed by the diffusion model. In contrast, while High-Pass shows a slight advantage in semantic structure alignment (CLIP-T), its DINO and CLIP-I scores are notably lower, which is detrimental to the fine-grained restoration of individual animal characteristics.

Projection Layer Design. Table 1 compares different projection layer designs. Q-Former significantly outperforms alternatives, improving CLIP-I by 19.42 points and DINO by 44.61 points. This demonstrates Q-Former's effectiveness as a semantic bottleneck for identity-relevant features.

Guidance Scale. Table 2 presents the ablation on guidance scale w. Low guidance ($w = 2.0$) produces poor identity preservation. The optimal $w = 7.5$ achieves the best balance between identity fidelity and generation quality.

Hyper-parameter λ. Figure 4 demonstrates the effects of the hyper-parameter λ on generated samples with a fixed random seed. As λ increases to 1.0, the generated animal gradually loses structural integrity. A smaller λ ensures the generated results adhere more closely to the input animal's identity. Consequently, we empirically set λ to 0.4 in our experiments.

Visualization. Fig. 5 demonstrates the impact of different frequency filter masks. The "Mini-pass" filter significantly blurs the image, whereas "Low-pass" preserves general shapes and smooth color transitions, which is critical for identity preservation. "Mid-pass" and "High-pass" filters accentuate textural details and sharp edges respectively.

3.5 Limitations

Despite these advantages, our method has limitations. The reliance on low-frequency structural guidance can occasionally constrain the diversity of generated poses, particularly for highly dynamic actions that differ significantly from the reference. Additionally, while the Q-Former effectively filters background noise, it may occasionally overlook extremely subtle identity cues that are not semantically salient. Future work will explore extending this framework to animal consistency generation tasks in video modalities.

4 Conclusion

This paper introduces AnimalBooth, an inference-time tuning-free personalized generation framework specifically designed for animal subjects. Experiments were conducted on our self-constructed AnimalBench dataset, comprising 10,958 training images and 1,000 test images. By integrating the Animal-Net with Q-Former projection and the novel dual-path adaptive attention module, AnimalBooth achieves state-of-the-art performance across all metrics. Furthermore, through the low-pass configuration of the DCT frequency-controlled feature integration module (based on ControlNet), we have enhanced the LPIPS metric by 20 percentage points over baselines. Notably, AnimalBooth achieves 25$\times$ faster inference than DiT-based alternatives while using only 0.7GB VRAM, making it highly practical for real-world deployment. Future work will explore extending this framework to animal consistency generation tasks in video modalities.

REFERENCES

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

Li-Wen Chang, Wenlei Bao, Qi Hou, Chengquan Jiang, Ningxin Zheng, Yinmin Zhong, Xuanrun Zhang, Zuquan Song, Chengji Yao, Ziheng Jiang, Haibin Lin, Xin Jin, and Xin Liu. Flux: Fast software-based communication overlap on gpus through kernel fusion, 2024. URL `https://arxiv.org/abs/2406.06858`.

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

Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8188–8197, 2020.

Miaomiao Dai, Qianyu Zhou, Ran Yi, and Lizhuang Ma. Diffusefist: A fast image-guided style transfer method for adapting large-scale diffusion models. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Fan Deng, Yaguang Wu, Xinyang Yu, Xiangjun Huang, Jian Yang, Guangyu Yan, and Qiang Xu. Locref-diffusion: Tuning-free layout and appearance-guided generation. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

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 *arXiv preprint arXiv:2208.01618*, 2022.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Custom-diffusion: Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8183–8194, 2023.

Dongxu Li, Junnan Li, and Steven C. H. Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. In *NeurIPS*, 2023a.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023b.

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*, pp. 1–12, 2024.

Xichen Pan, Weizhi Wang, Xinyu Geng, Wen Song, Jian-Fu Li, Wenhu Chen, William W Cohen, and Sébastien Bubeck. Kosmos-g: Generating images in context with multimodal large language models. *arXiv preprint arXiv:2310.02992*, 2023.

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*, pp. 8748–8763. PMLR, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

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. 2022.

Fei Shen and Jinhui Tang. Imagpose: A unified conditional framework for pose-guided person generation. *Advances in neural information processing systems*, 37:6246–6266, 2024.

Fei Shen, Xiaoyu Du, Yutong Gao, Jian Yu, Yushe Cao, Xing Lei, and Jinhui Tang. Imagharmony: Controllable image editing with consistent object quantity and layout. *arXiv preprint arXiv:2506.01949*, 2025a.

Fei Shen, Xin Jiang, Xin He, Hu Ye, Cong Wang, Xiaoyu Du, Zechao Li, and Jinhui Tang. Imagdressing-v1: Customizable virtual dressing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 6795–6804, 2025b.

Fei Shen, Cong Wang, Junyao Gao, Qin Guo, Jisheng Dang, Jinhui Tang, and Tat-Seng Chua. Long-term talkingface generation via motion-prior conditional diffusion model. *arXiv preprint arXiv:2502.09533*, 2025c.

Fei Shen, Jian Yu, Cong Wang, Xin Jiang, Xiaoyu Du, and Jinhui Tang. Imaggarment-1: Fine-grained garment generation for controllable fashion design. *arXiv preprint arXiv:2504.13176*, 2025d.

Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14398–14409, June 2024.

Xierui Wang, Siming Fu, Qihan Huang, Wanggui He, and Hao Jiang. Ms-diffusion: Multi-subject zero-shot image personalization with layout guidance, 2025. URL https://arxiv.org/abs/2406.07209.

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 *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13294–13304, 2025.

Yuanfeng Xu, Yuhao Chen, Zhongzhan Huang, Zijian He, Guangrun Wang, and Liang Lin. Anima2: Cross-species animal animation through image-to-video synthesis with subject alignment. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2025. doi: 10.1109/ICASSP49660.2025.10889218.

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.

Yu Zeng, Vishal M Patel, Haochen Wang, Xun Huang, Ting-Chun Wang, Ming-Yu Liu, and Yogesh Balaji. Jedi: Joint-image diffusion models for finetuning-free personalized text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6786–6795, 2024.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023.

Yuxuan Zhang, Jiaming Liu, Yiren Song, Rui Wang, Hao Tang, Jinpeng Yu, Huaxia Li, Han Pan, and Zhongliang Jing. Ssr-encoder: Encoding selective subject representation for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8069–8079, 2024a.

Yuxuan Zhang, Yiren Song, Jinpeng Yu, Han Pan, and Zhongliang Jing. Fast personalized text to image synthesis with attention injection. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6195–6199, 2024b. doi: 10.1109/ICASSP48485.2024.10447042.