

1 Problem & Motivation

Personalized animal image generation is challenging due to **non-rigid deformations**, **anatomical variation across species**, and **intricate textures like fur, feathers, and stripes**. Existing methods suffer from **cross-domain feature misalignment** and **identity drift**.

Pose Variation

Large pose changes and non-rigid deformation.

Species Morphology

Anatomical diversity across species.

Fine Texture Preservation

Intricate textures: fur, feathers, stripes.

2 Key Idea / Contributions

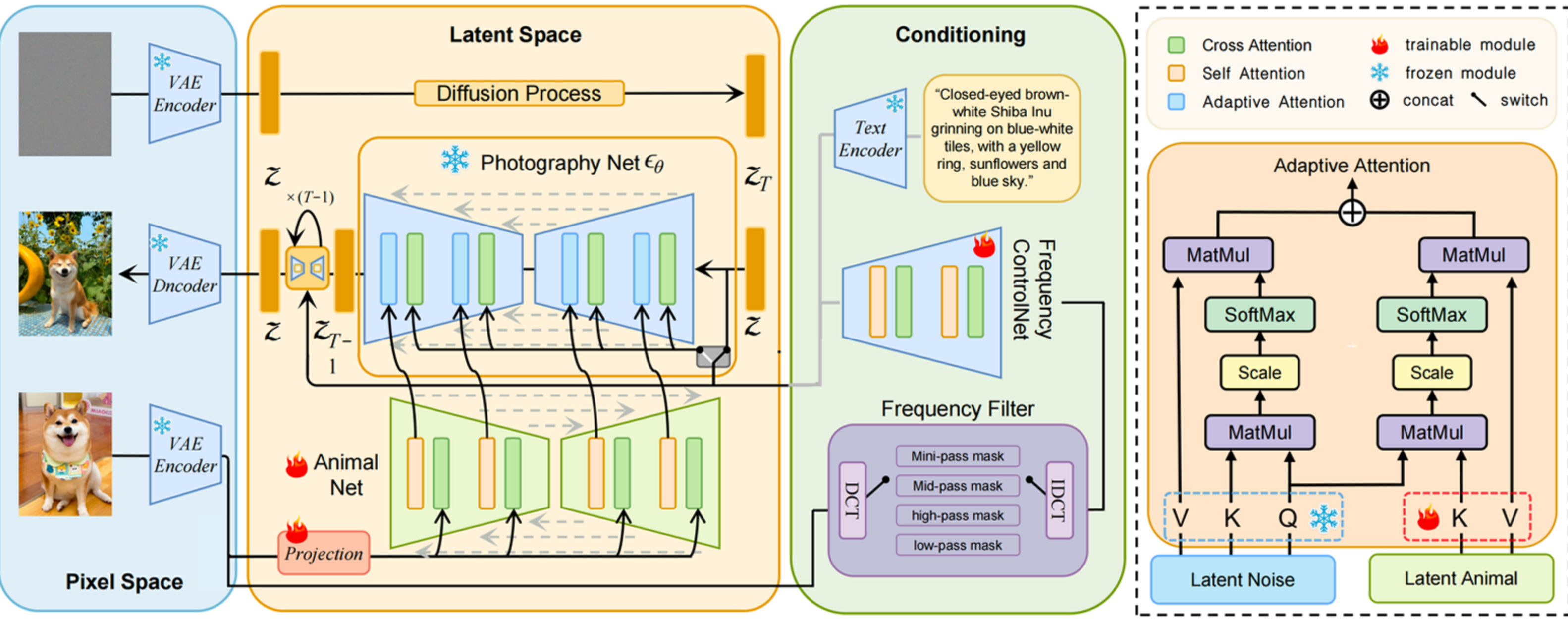
- Inference-time tuning-free animal personalization framework.
- Animal-Net with Q-Former bottleneck for identity-preserving semantic filtering.
- Dual-path adaptive attention for text control + identity injection.
- Frequency-controlled latent feature integration using DCT for coarse-to-fine structure/texture guidance.
- AnimalBench high-definition dataset for animal personalization.

4 AnimalBench Dataset

Source Image	Detailed Caption	Segmentation Mask	Masked Subject	Paired Image
	A playful Arctic fox pouncing in a snowy landscape under golden sunlight.			
	A cat lounging on a stack of vintage books in a dimly lit library aisle.			
	A cow strolling a bustling city street with neon reflections on its coat.			
	A zebra grazing peacefully under the rain, with sparkling beads on its coat.			

10,958 training images | 1,000 test images

3 Method Overview



7 Ablations

$\lambda = 0.4$ achieves the best balance between structure preservation and texture fidelity.

Figure 7 shows a grid of images generated with different λ values (0.1 to 1.0). As λ increases, the generated images become more abstract and lose fine texture details. $\lambda = 0.4$ is highlighted as the optimal balance.

(a) Projection Layer

Methods	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

(b) 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

(c) Frequency Ablation (Control Signal)

Filter Type	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

5 Results (Qualitative Comparison)

	Reference	BLIP-Diffusion	Omnigen	IP-Adapter	AnimalBooth (Ours)
A cheetah gracefully gliding through a dense forest.					
A reindeer in a winter landscape.					
A Highland cow stands in the wild.					
A zebra stands beneath the moonlight on the grassland.					

AnimalBooth preserves both coarse structures and fine-grained textures.

6 Quantitative Performance

Methods	LPIPS↓	DINO↑	CLIP-T↑	CLIP-I↑
Textual Inv.	72.35	48.92	18.76	71.23
BLIP	68.21	62.96	19.68	82.38
Omnigen	71.68	50.05	19.41	72.95
IP-Adapter	62.91	72.88	19.39	89.75
Flux (DiT)	65.47	55.31	19.52	78.64
AnimalBooth (Ours)	49.08	75.66	20.73	90.00

8 Efficiency & Takeaways

Methods	Time per Image (50 steps)	Peak VRAM	Params
Omnigen	140.0 s	10.4 GB	3.8 B
Flux (DiT)	85.0 s	12.8 GB	12.0 B
AnimalBooth (Ours)	5.5 s	0.7 GB	1.2 B

25x Faster than DiT alternatives

0.7GB VRAM

SOTA Identity Fidelity

Limitations: Low-frequency guidance may constrain extreme pose diversity and fine local details in rare cases.

