Disentangling Dual Image References in Frequency Aware Diffusion Models for Personalized Generation

Haipeng Liu, Yang Wang*, Meng Wang

School of Computer Science and Information Engineering
Hefei University of Technology, China

* Yang Wang is the corresponding author

Advances in Neural Information Processing Systems (NeurIPS), 2026

Background

Personalized image generation aims to synthesize text-driven images conditioned on reference images

Image Customization (Foreground)



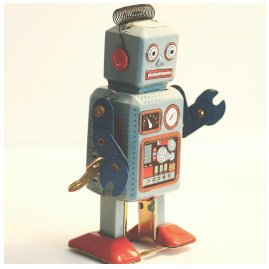
A photograph of a butterfly fluttering around a lush green garden





A photograph of Spider-Man rescuing a cat from a tall tree





A photograph of a robot helping an elderly woman cross the street



Style Transfer (Background)



A blue giraffe and a brown vase





Seven women





The colorful balloons floated near the shiny silver airplane



Personalized Image Generation -- Dual Image References



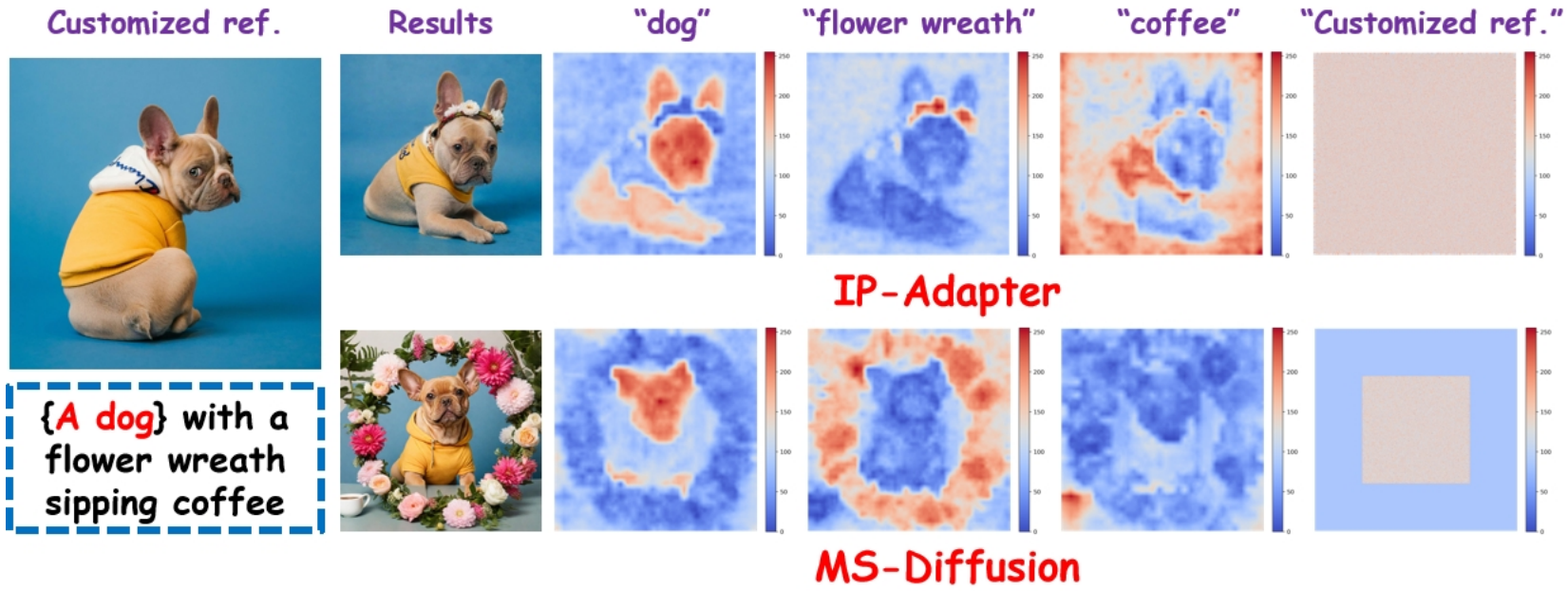
(a) Customization Style Transfer



(b) Color Style Transfer

Motivation

- Neither image customization nor style transfer achieves satisfactory alignment between the generated image and the text prompt for diffusion models.



(a) Image customization



(b) Style Transfer

Motivation



Color Transfer



Image Customization



Style Transfer



Color

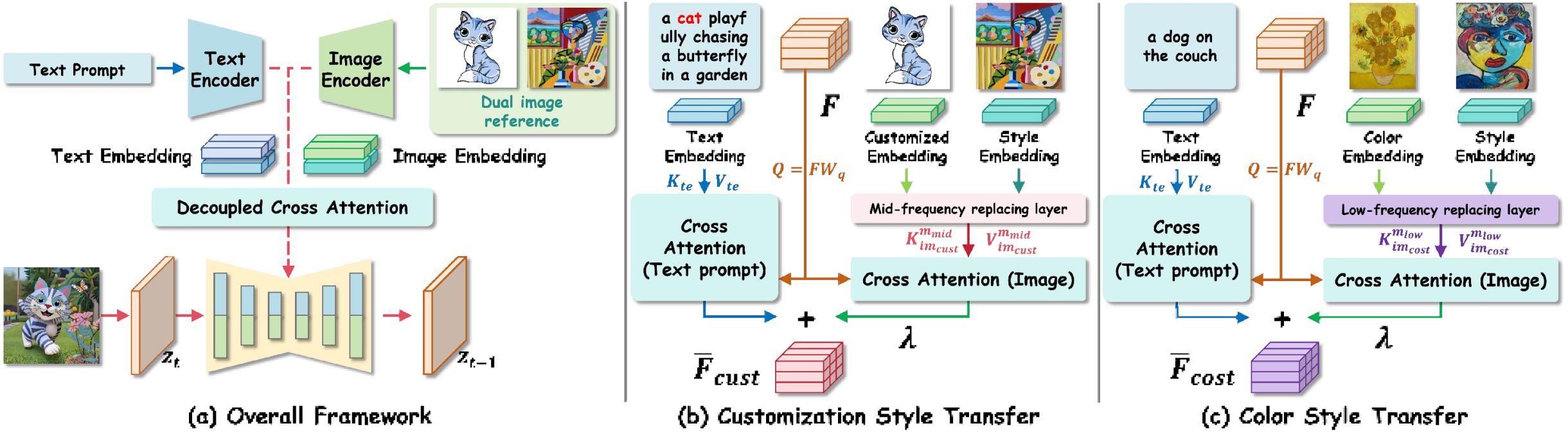
Global Structure

Local Texture

The early (high-noise) denoising stage mainly focuses on the low-and-mid frequency bands, while the late (low-noise) denoising stage progressively exhibits the high-frequency band, which, all together, overwhelm the text prompt guidance, suffering from the heavy reliance on reference image

Methodology

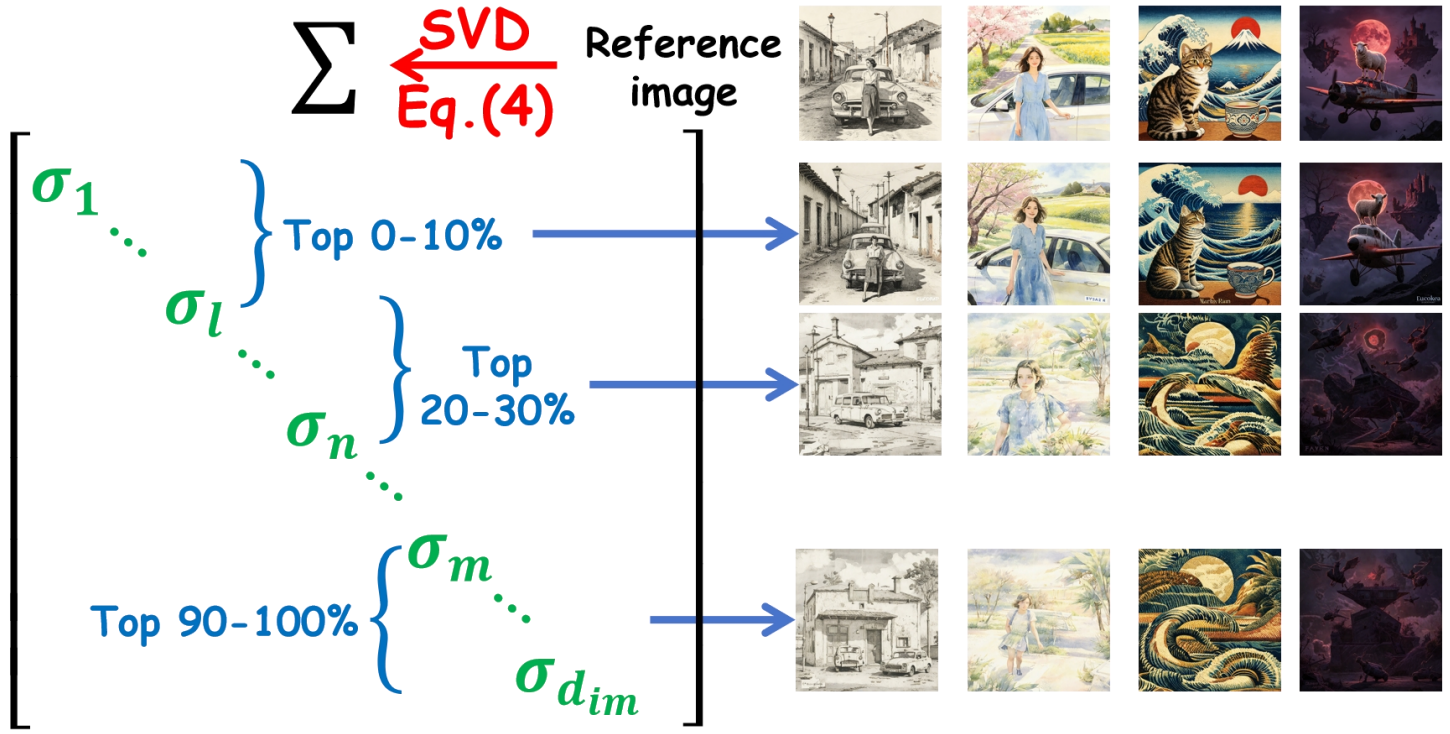
- Developing a framwork to disentangle these **Dual image references** within Frequency-aware Diffusion Models, dubbed **Dual-FDM**, for two classical personalized image generation tasks in a row: customization style transfer and color style transfer, by disentangling different frequency bands via mask strategy within frequency domain.



Methodology

- How to pop up the Background of Image Reference (Style)

$$W_k^{im} = U\Sigma V^T$$

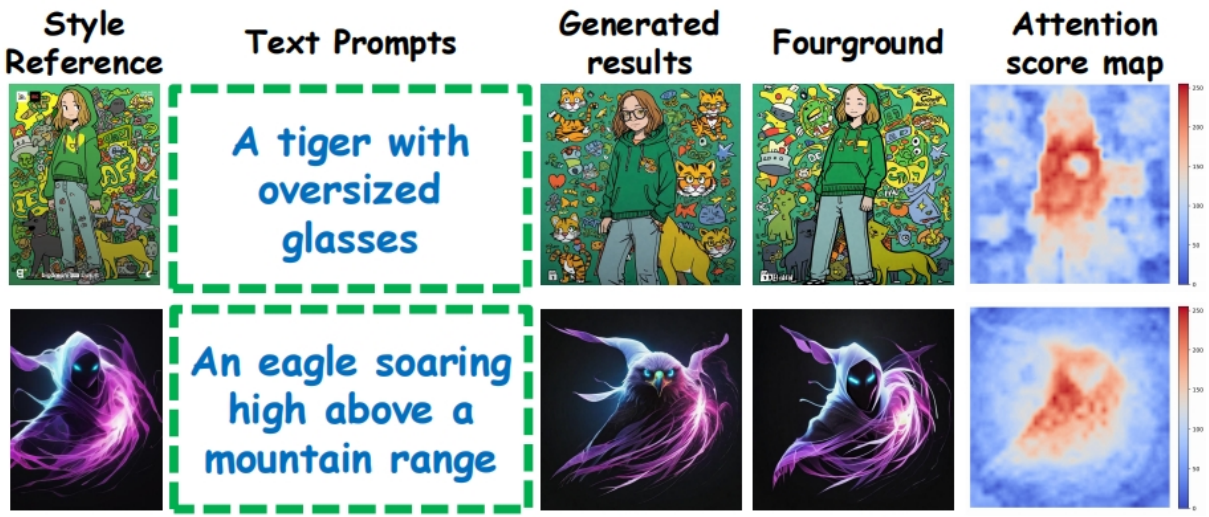


Remark 1. Larger σ_i tend to correspond to the foreground, whereas smaller σ_i are more likely associated with the background.

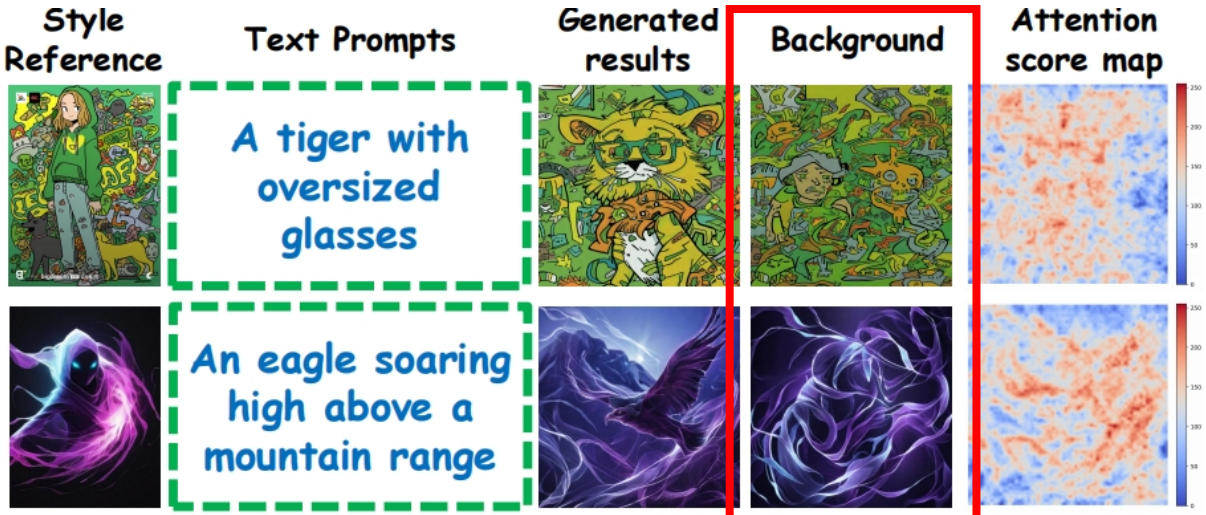
Methodology

$$\overline{W}_k^{im} = U\,\mathrm{diag}\left(\frac{1}{\sigma_1}, \frac{1}{\sigma_2}, \ldots, \frac{1}{\sigma_{d_{im}}}\right) V^T, \quad \sigma_i \neq 0.$$

Remark 2. Inverting the singular values suppresses dominant foreground components while enhancing background-related information.



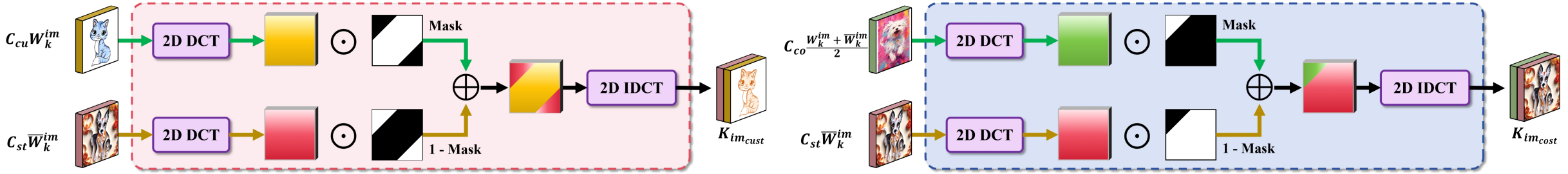
(a) The Generated results of W_k^{im}



(b) The Generated results of W_k^{im}

Methodology

➢ How to disentangle different frequency bands?



(a) Mid-frequency replacing layer

(b) Low-frequency replacing layer

Customization Style Transfer

$$K_{im_{cust}}^{m_{mid}} = \mathrm{IDCT}\Big(\mathrm{DCT}(C_{cu}W_k^{im}) \odot m_{mid} + \mathrm{DCT}(C_{st}\overline{W}_k^{im}) \odot (1 - m_{mid})\Big),$$

$$V_{im_{cust}}^{m_{mid}} = \mathrm{IDCT}\Big(\mathrm{DCT}(C_{cu}W_v^{im}) \odot m_{mid} + \mathrm{DCT}(C_{st}W_v^{im}) \odot (1 - m_{mid})\Big),$$

Color Style Transfer

$$K_{im_{cost}}^{m_{low}} = \mathrm{IDCT}\Big(\mathrm{DCT}\Big(C_{co}\Big(\frac{W_k^{im}+\overline{W}_k^{im}}{2}\Big)\Big) \odot m_{low} + \mathrm{DCT}(C_{st}\overline{W}_k^{im}) \odot (1 - m_{low})\Big),$$

$$V_{im_{cost}}^{m_{low}} = \mathrm{IDCT}\Big(\mathrm{DCT}(C_{co}W_v^{im}) \odot m_{low} + \mathrm{DCT}(C_{st}W_v^{im}) \odot (1 - m_{low})\Big),$$

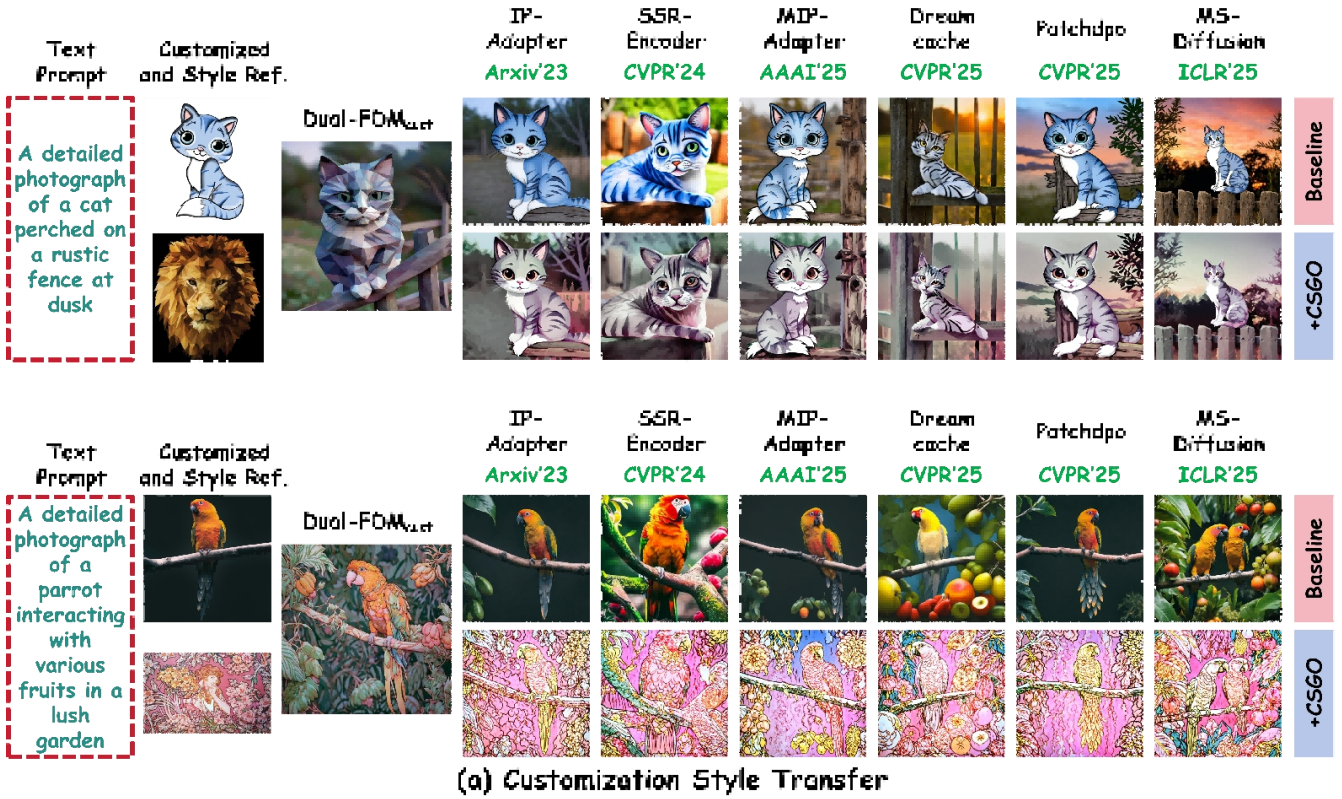
Methodology

- Adaptively Modulating the Early and Late Denoising Stage
  - Early Stage -- Low or Middle Frequency
  - Late Stage -- High Frequency

$$\Delta_{cost} = \frac{H\left(K^{m^e_{low}}_{im_{cost}}\right)}{H\left(K^{m^l_{low}}_{im_{cost}}\right)}, \quad H(x) = -\sum_{i=1}^{d} \mathrm{Softmax}(x)_i \log(\mathrm{Softmax}(x)_i + \epsilon)\,,$$

Experiments

Qualitative Comparisons





Experiments

Quantitative Comparisons

	Baseline	IP-Adapter Arxiv'23	SSR-Encoder CVPR'24	MIP-Adapter AAAI'25	Dreamcache CVPR'25	PatchDPO CVPR'25	MS-Diffusion ICLR'25	DUAL-FDM$_{cust}$
CSGO (ARXIV'24)	CLIP-T ↑	0.2851	0.2829	0.2809	0.2803	0.2793	0.3123	0.3194
	CLIP-TC ↑	0.4802	0.4689	0.4744	0.4518	0.4738	0.4680	0.5031
	CLIP-TS ↑	0.4415	0.4439	0.4420	0.4525	0.4440	0.4577	0.4516
	CLIP-CS ↑	0.6658	0.6587	0.6626	0.6465	0.6638	0.6419	0.6687
	IR$_{\times 10}$ ↑	-5.5733	-7.4241	-6.5255	-6.7782	-7.1456	4.2023	5.3487
	HPS$_{\times 10}$ ↑	2.3401	2.2692	2.2624	2.1794	2.2493	2.5436	2.6007
STYLEID (CVPR'24)	CLIP-T ↑	0.2731	0.2768	0.2848	0.2647	0.2650	0.3012	0.3194
	CLIP-TC ↑	0.4828	0.4773	0.4699	0.4424	0.4744	0.4614	0.5031
	CLIP-TS ↑	0.4349	0.4370	0.4458	0.4488	0.4370	0.4556	0.4516
	CLIP-CS ↑	0.6731	0.6665	0.6584	0.6460	0.6710	0.6423	0.6687
	IR$_{\times 10}$ ↑	-8.6015	-8.7602	-9.5626	-9.9445	-10.4733	2.6485	5.3487
	HPS$_{\times 10}$ ↑	2.0861	2.0979	2.0073	1.9126	1.9978	2.3322	2.6007
STYLESSP (CVPR'25)	CLIP-T ↑	0.2642	0.2656	0.2584	0.2561	0.2591	0.2907	0.3194
	CLIP-TC ↑	0.4664	0.4729	0.4616	0.4355	0.4638	0.4540	0.5031
	CLIP-TS ↑	0.4397	0.4407	0.4409	0.4468	0.4413	0.4583	0.4516
	CLIP-CS ↑	0.6680	0.6592	0.6678	0.6433	0.6700	0.6462	0.6687
	IR$_{\times 10}$ ↑	-10.8440	-10.4537	-12.0747	-11.0862	-12.1339	0.5184	5.3487
	HPS$_{\times 10}$ ↑	2.0087	2.0153	1.9075	1.8562	1.9204	2.2367	2.6007

Experiments

Quantitative Comparisons

		InstantStyle Arxiv'24	CSGO NeurIPS'25	StyleShot Arxiv'24	StyleAligned CVPR'24	DEADiff CVPR'24	StyleCrafter TOG'24	MaskST ICLR'25	StyleStudio CVPR'25	Dual-FDM$_{cost}$
Color	CLIP-TS ↑	0.4587	0.4539	0.4644	0.4832	0.4448	0.4771	0.4641	0.4497	0.5015
	CLIP-S ↑	0.5967	0.6392	0.6782	0.6904	0.6177	0.6304	0.6763	0.5859	0.7178
	DINO-S ↑	0.2146	0.3924	0.4769	0.4437	0.3318	0.3541	0.4764	0.2370	0.4970
Shape	CLIP-TS ↑	0.4507	0.4561	0.4595	0.4797	0.4392	0.4656	0.4597	0.4504	0.4956
	CLIP-S ↑	0.5864	0.6357	0.6602	0.7051	0.5986	0.6274	0.6602	0.6162	0.7271
	DINO-S ↑	0.2087	0.3711	0.4374	0.4658	0.2944	0.3843	0.4368	0.3087	0.4982
Texture	CLIP-TS ↑	0.4524	0.4561	0.4543	0.4819	0.4304	0.4712	0.4534	0.4443	0.4988
	CLIP-S ↑	0.5879	0.6406	0.6553	0.7246	0.5835	0.6348	0.6538	0.5889	0.7358
	DINO-S ↑	0.2131	0.3760	0.4476	0.4844	0.2707	0.3729	0.4465	0.2616	0.4991
Non Spatial	CLIP-TS ↑	0.4424	0.4553	0.4504	0.4839	0.4402	0.4553	0.4504	0.4314	0.4958
	CLIP-S ↑	0.5723	0.6279	0.6387	0.7251	0.5972	0.6064	0.6387	0.5566	0.7261
	DINO-S ↑	0.2336	0.4278	0.4853	0.5171	0.3533	0.3795	0.4854	0.2646	0.5406
2D Spatial	CLIP-TS ↑	0.4568	0.4670	0.4661	0.4912	0.4563	0.4829	0.4668	0.4565	0.5122
	CLIP-S ↑	0.5835	0.6309	0.6592	0.7129	0.6362	0.6421	0.6597	0.6060	0.7314
	DINO-S ↑	0.2262	0.4113	0.4615	0.4998	0.3824	0.3977	0.4647	0.3253	0.5187
3D Spatial	CLIP-TS ↑	0.4573	0.4624	0.4663	0.4875	0.4497	0.4773	0.4663	0.4429	0.5078
	CLIP-S ↑	0.5918	0.6333	0.6680	0.7100	0.6333	0.6357	0.6680	0.5732	0.7334
	DINO-S ↑	0.2166	0.3819	0.4654	0.4760	0.3655	0.3774	0.4696	0.2249	0.5094
Number	CLIP-TS ↑	0.4490	0.4509	0.4597	0.4902	0.4465	0.4783	0.4597	0.4436	0.5029
	CLIP-S ↑	0.5845	0.6182	0.6562	0.7241	0.6226	0.6499	0.6558	0.6011	0.7334
	DINO-S ↑	0.2289	0.3882	0.4779	0.5116	0.3541	0.4346	0.4738	0.3160	0.5469
Complex	CLIP-TS ↑	0.4404	0.4500	0.4556	0.4836	0.4390	0.4636	0.4558	0.4351	0.4954
	CLIP-S ↑	0.5864	0.6445	0.6582	0.7256	0.6021	0.6387	0.6582	0.5791	0.7329
	DINO-S ↑	0.2134	0.3911	0.4533	0.4961	0.2938	0.3956	0.4536	0.2274	0.5051

Conclusion

In this paper, we study personalized generation based on dual references - customization and color and style reference - and propose a paradigm to disentangle dual image references within frequency-aware diffusion models, dubbed Dual-FDM, for two classical personalized image generation tasks in a row: customization style transfer and color style transfer, by disentangling different frequency bands via mask strategy within frequency domain.

Thanks for Listening

Q&A