PCA Lab Work Report

[CVPR 2026] WaDi: Weight Direction-aware Distillation for One-step Image Synthesis

Lei Wang Yang Cheng Senmao Li Ge Wu Yaxing Wang ✉ Jian Yang ✉

Nankai University

Task Background

Wide Applications of Diffusion Models

Text-to-image [1]

Text-to-video [2]

[1] Esser P, Kulal S, Blattmann A, et al. Scaling rectified flow transformers for high-resolution image synthesis[C]//Forty-first international conference on machine learning. 2024.
[2] Kong W, Tian Q, Zhang Z, et al. Hunyuanvideo: A systematic framework for large video generative models[J]. arXiv preprint arXiv:2412.03603, 2024.

Task Background

Wide Applications of Diffusion Models

Image-to-video [3]

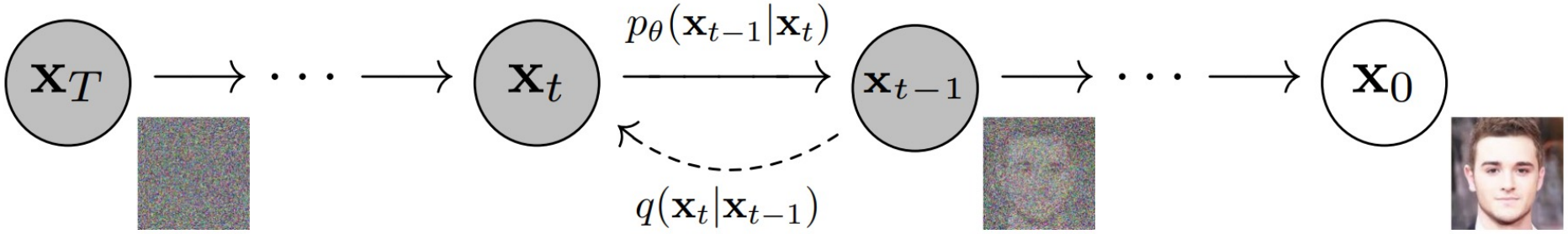
Use an image as the first frame and combine it with a prompt to generate a video.

[3] Yang Z, Teng J, Zheng W, et al. Cogvideox: Text-to-video diffusion models with an expert transformer[J]. arXiv preprint arXiv:2408.06072, 2024.

Task Background

Limitation of Diffusion Models: Speed Bottleneck

迭代去噪



DDPM [4]： 1000 step **30+s** DDIM [5]： 50step **2-3s**

Problem: Slow Generation Speed

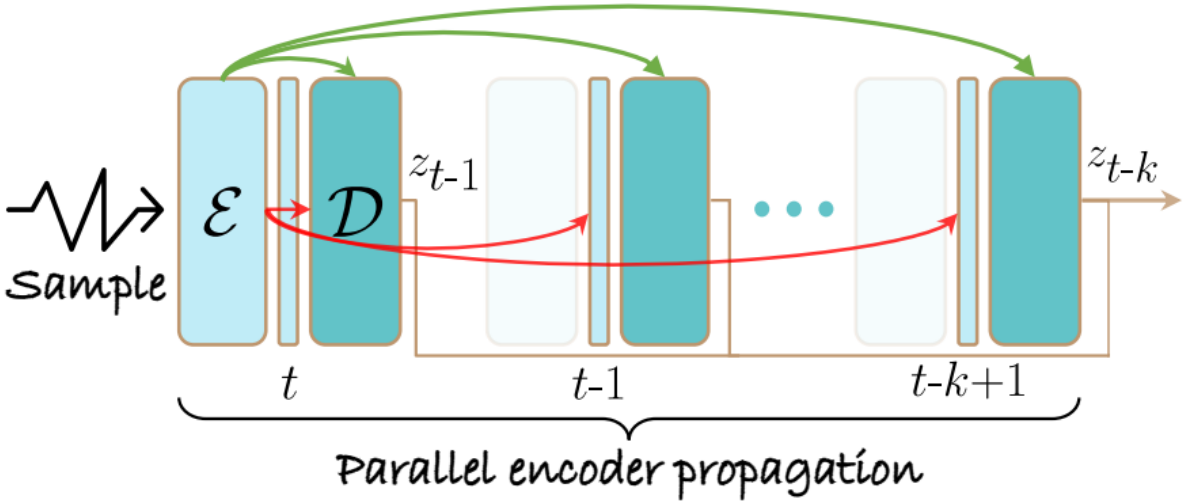
[4] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models[J]. Advances in neural information processing systems, 2020, 33: 6840-6851.

[5] Song J, Meng C, Ermon S. Denoising diffusion implicit models[J]. arXiv preprint arXiv:2010.02502, 2020.

Related Work

Diffusion Model Acceleration

Cache-based Methods

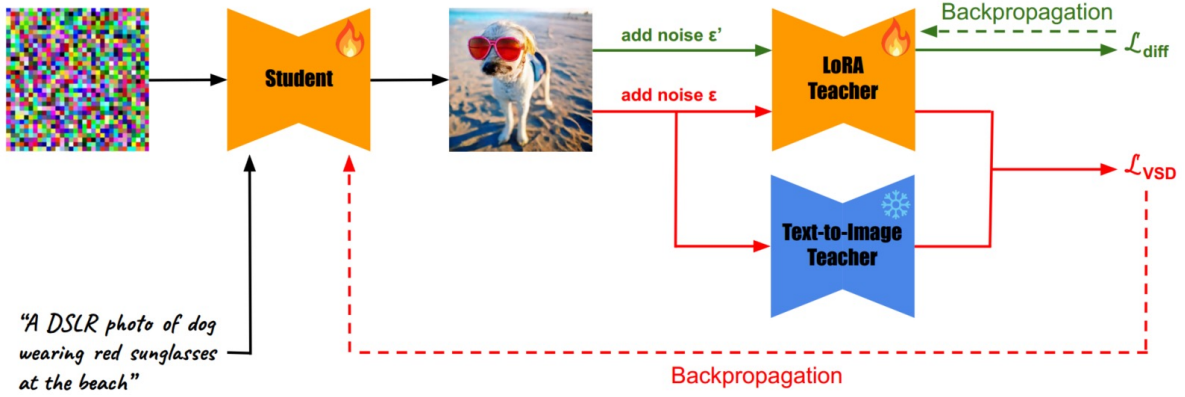


Faster diffusion [6]

Advantage: No Training Required

Disadvantage: Limited Acceleration Performance

Distillation-based Methods



Swiftbrush [7]

Advantage: Significant Acceleration Performance

Disadvantage: High Computational Cost

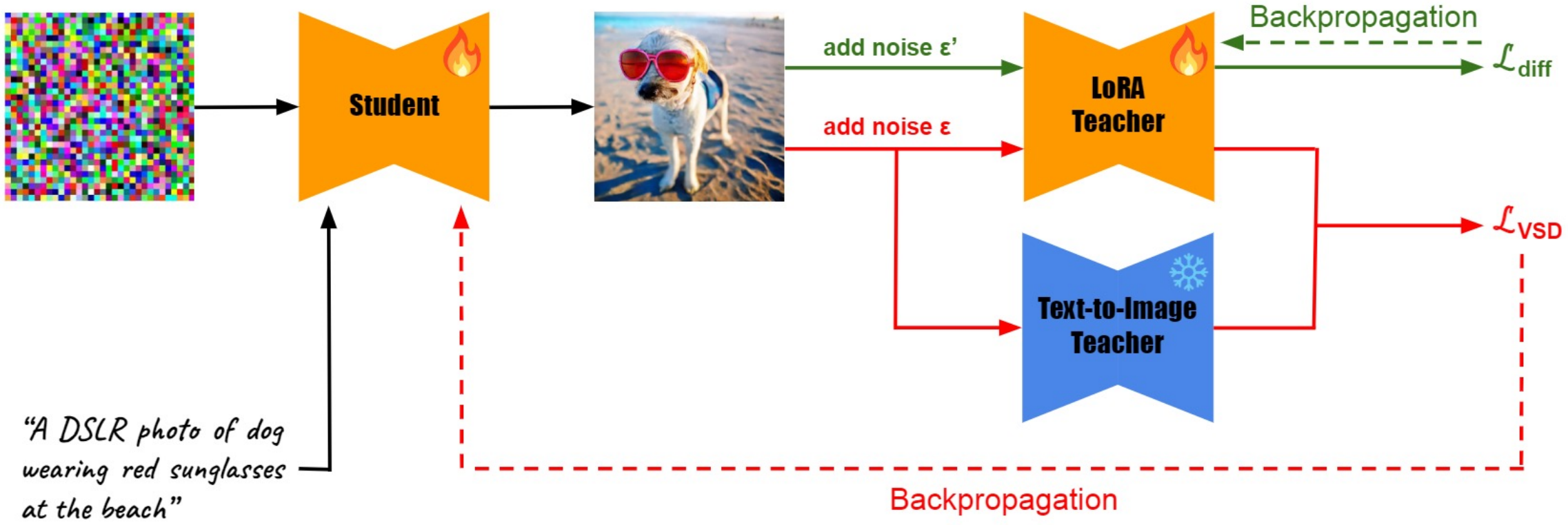
[6] Li S, Hu T, Shahbaz Khan F, et al. Faster diffusion: Rethinking the role of unet encoder in diffusion models[J]. arXiv e-prints, 2023: arXiv: 2312.09608.

[7] Nguyen T H, Tran A. Swiftbrush: One-step text-to-image diffusion model with variational score distillation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 7807-7816.

Related Work

Methods Based on Variational Score Distillation (VSD)

Image-Free Distillation

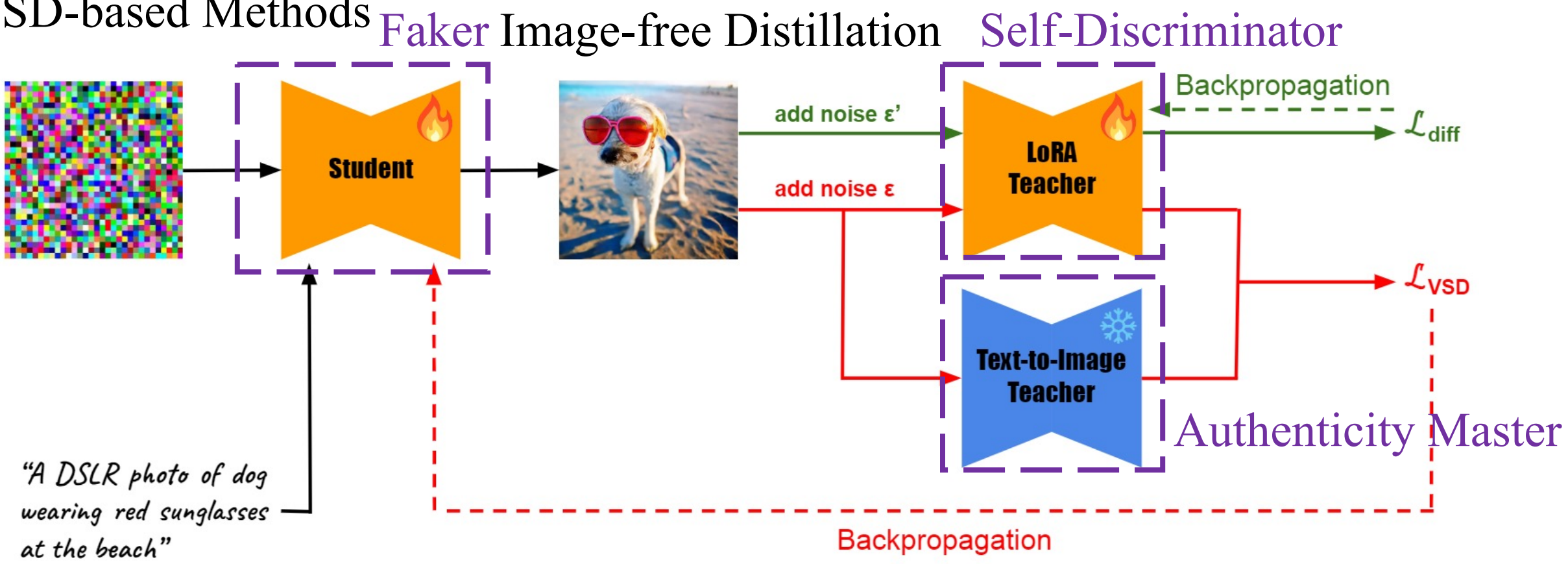


$$\nabla_\theta \mathcal{L}_{VSD} = \mathbb{E}_{t,\mathbf{y},\mathbf{z}} \left[w(t)(\hat{\epsilon}_\phi(\mathbf{x}_t, t, \mathbf{y}) - \hat{\epsilon}_\psi(\mathbf{x}_t, t, \mathbf{y})) \frac{\partial f_\theta(\mathbf{z}, \mathbf{y})}{\partial \theta} \right],$$

[7] Nguyen T H, Tran A. Swiftbrush: One-step text-to-image diffusion model with variational score distillation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 7807-7816.

Related Work

VSD-based Methods



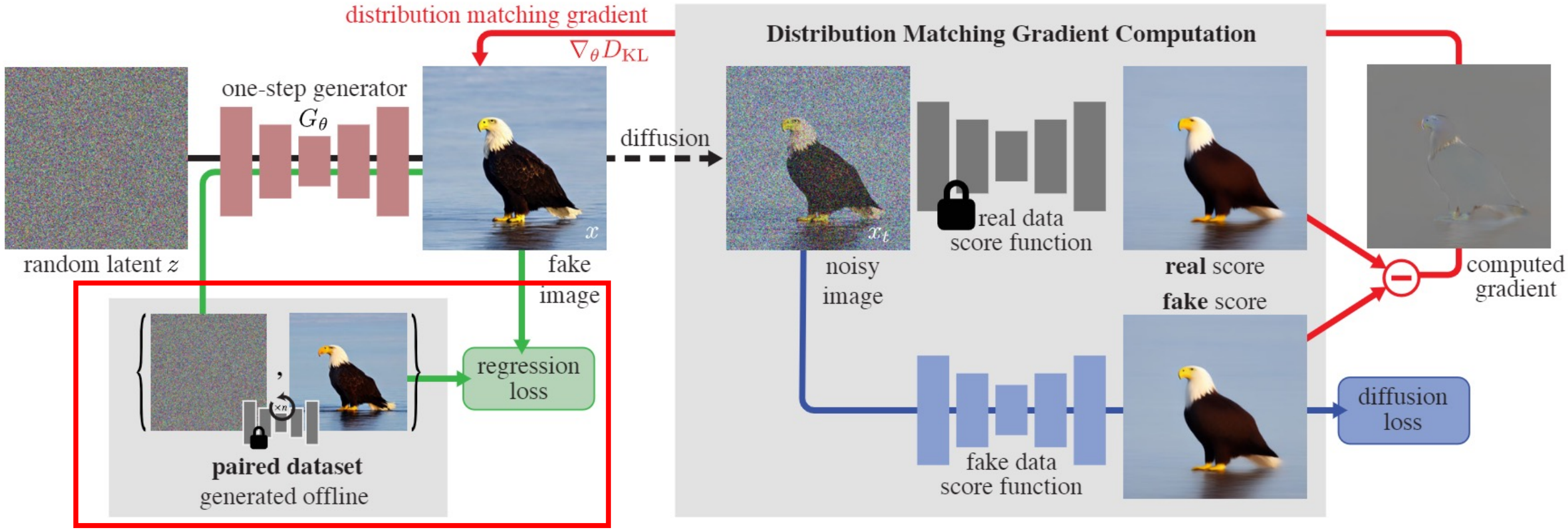
$$\nabla_\theta \mathcal{L}_{VSD} = \mathbb{E}_{t,\mathbf{y},\mathbf{z}} \left[w(t)(\hat{\epsilon}_\phi(\mathbf{x}_t, t, \mathbf{y}) - \hat{\epsilon}_\psi(\mathbf{x}_t, t, \mathbf{y})) \frac{\partial f_\theta(\mathbf{z}, \mathbf{y})}{\partial \theta} \right],$$

[7] Nguyen T H, Tran A. Swiftbrush: One-step text-to-image diffusion model with variational score distillation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 7807-7816.

Related Work

VSD-based Methods

Image-free？ × Teacher LoRA→Fake model（FT）



DMD [8]

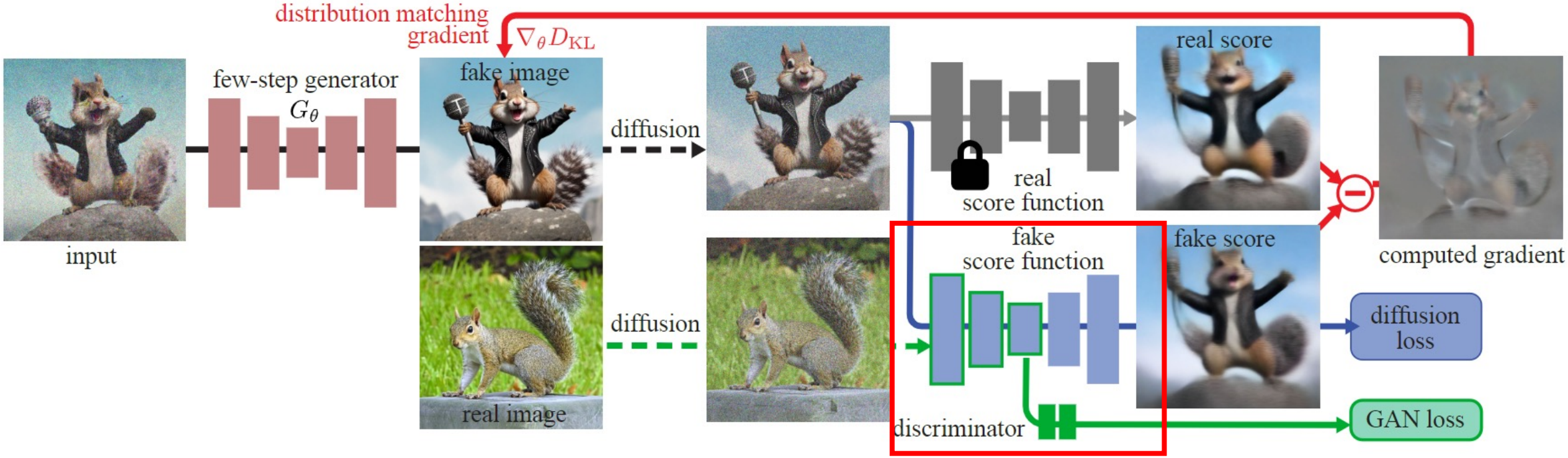
[8] Yin T, Gharbi M, Zhang R, et al. One-step diffusion with distribution matching distillation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024: 6613-6623.

Related Work

VSD-based Methods

Image-free？ ×

Introduced GAN loss to stabilize training

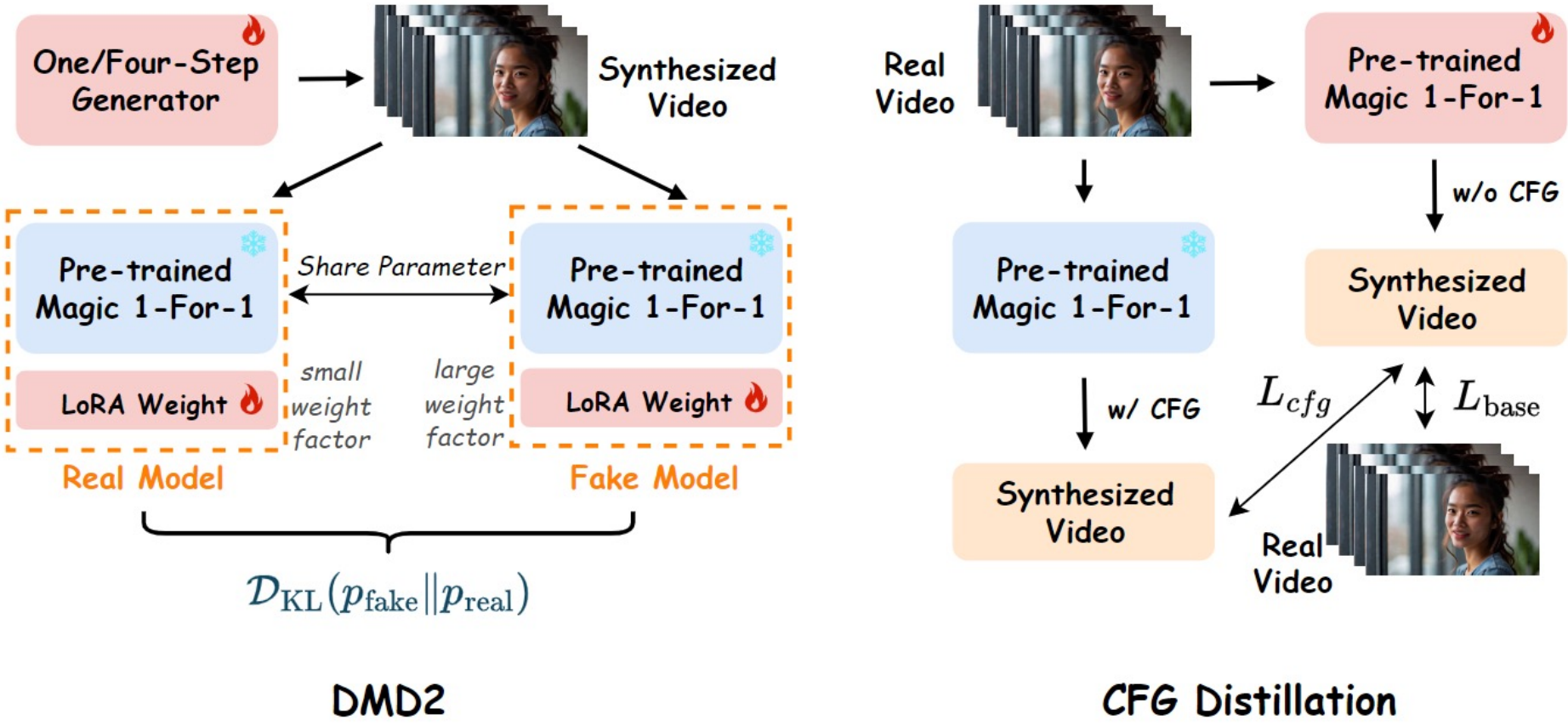


DMD2 [9]

[9] Yin T, Gharbi M, Zhang R, et al. One-step diffusion with distribution matching distillation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024: 6613-6623.

Related Work

VSD-based Methods

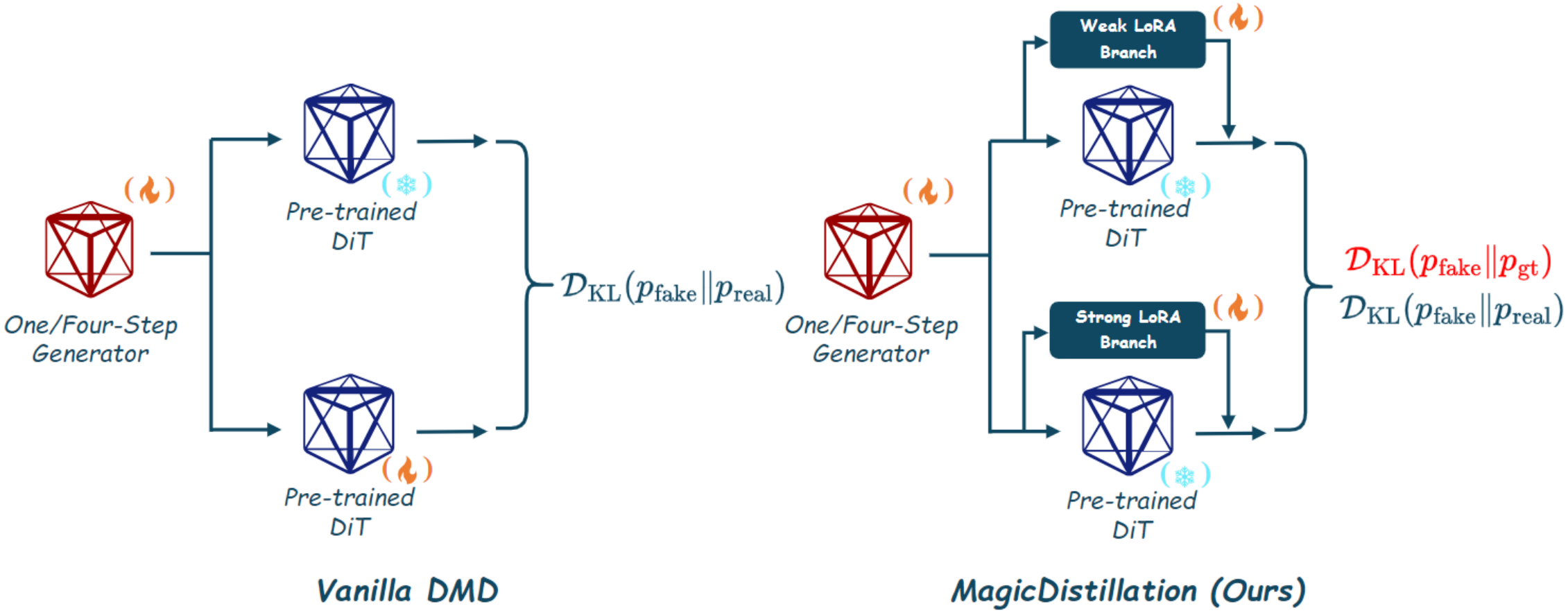


Magic141 [10]

[10] Yi H, Shao S, Ye T, et al. Magic 1-For-1: Generating One Minute Video Clips within One Minute[J]. arXiv preprint arXiv:2502.07701, 2025.

Related Work

VSD-based Methods

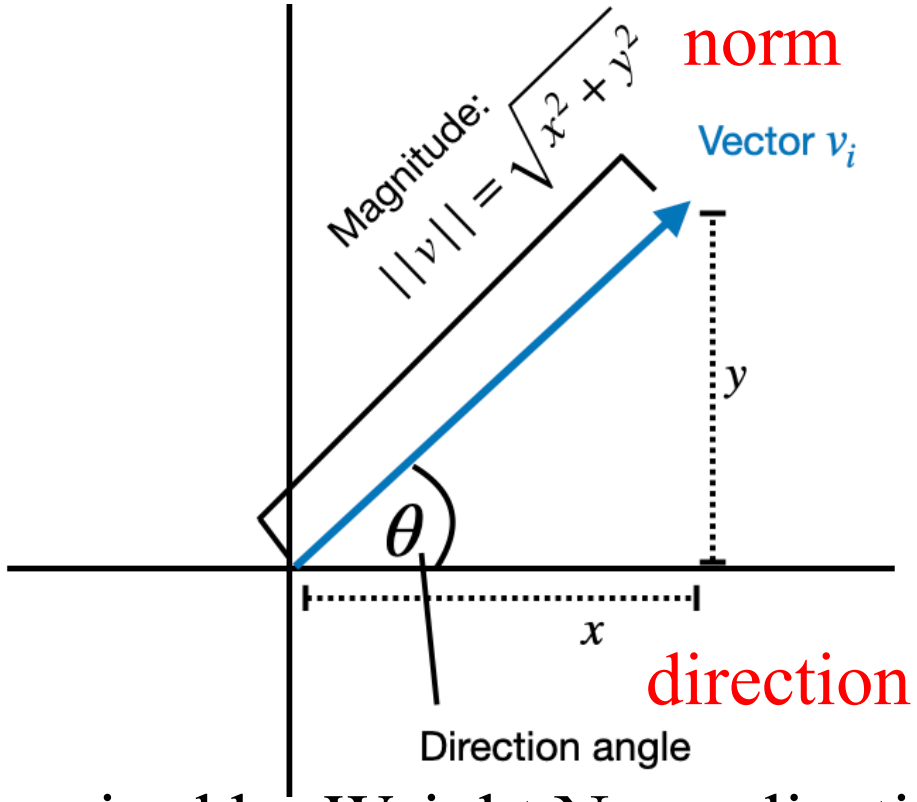


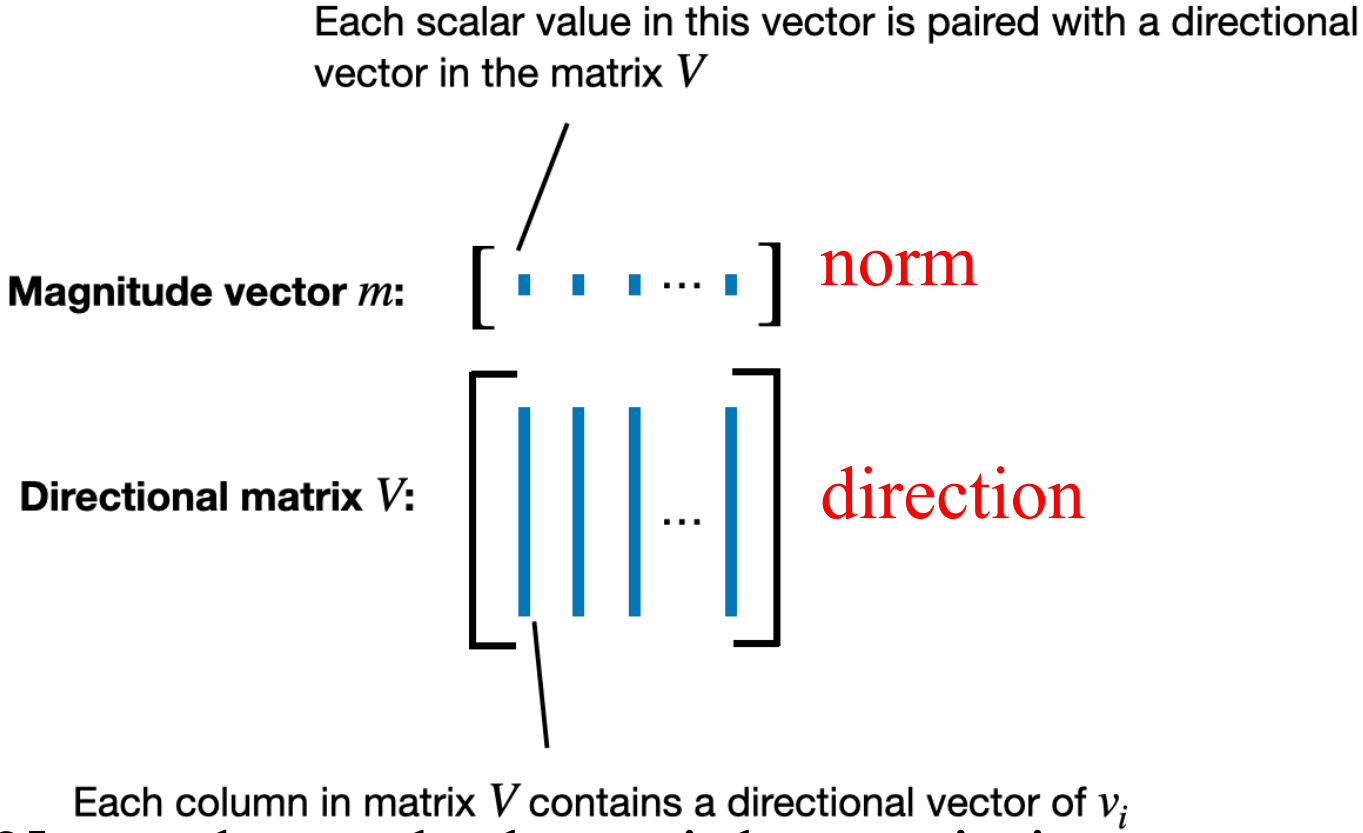
MagicDistillation [11]

[11] Shao S, Yi H, Guo H, et al. MagicDistillation: Weak-to-Strong Video Distillation for Large-Scale Portrait Few-Step Synthesis[J]. arXiv preprint arXiv:2503.13319, 2025.

Research Motivation

What exactly happens during distillation that enables the U-Net/DiT to generate in a single step?



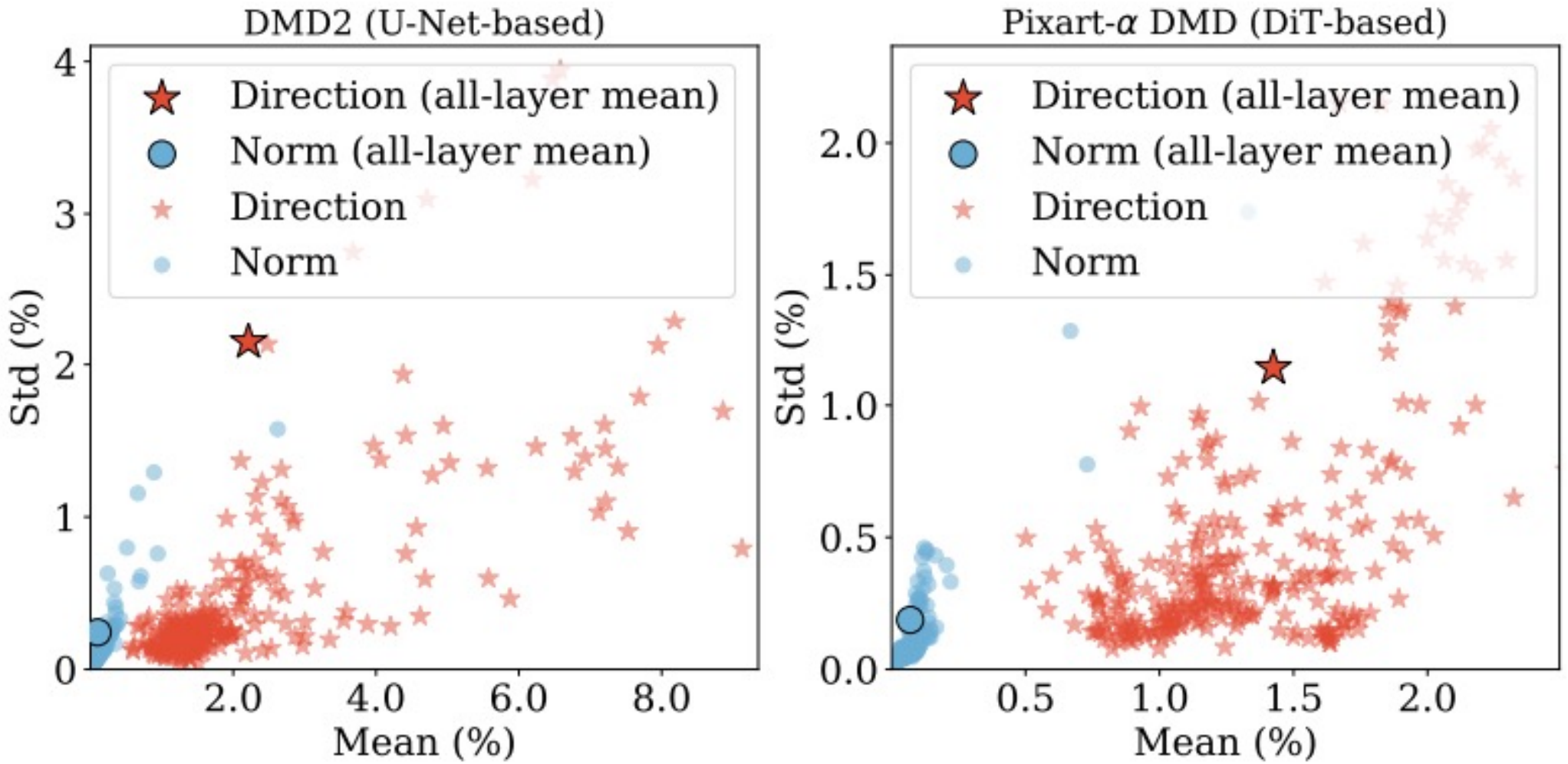


Inspired by Weight Normalization [12], we decouple the weight matrix into norm and direction.

[12] Salimans T, Kingma D P. Weight normalization: A simple reparameterization to accelerate training of deep neural networks[J]. Advances in neural information processing systems, 2016, 29.

Research Motivation

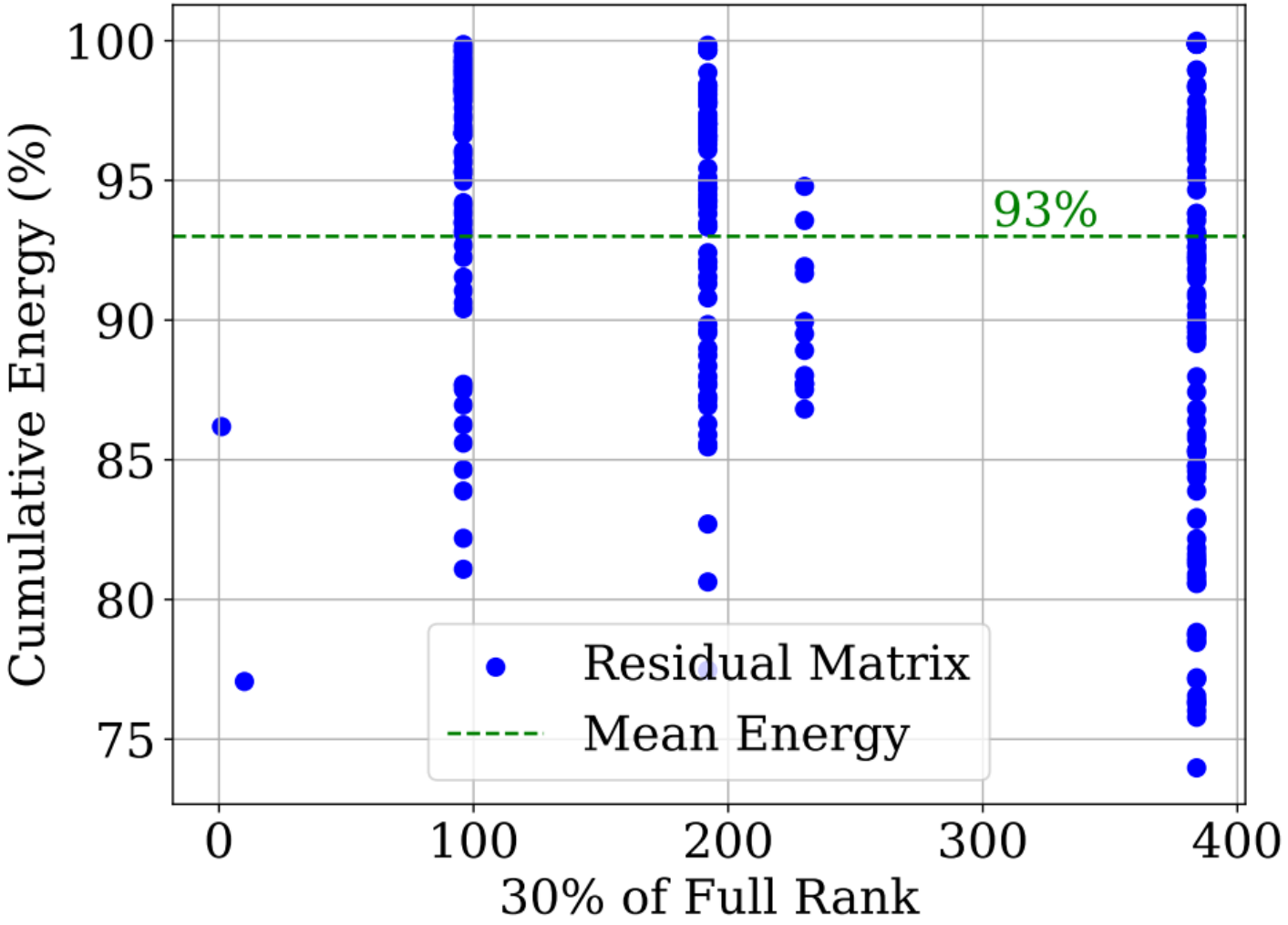
Empirical Observation



The norm changes slightly, while the direction changes significantly.

Research Motivation

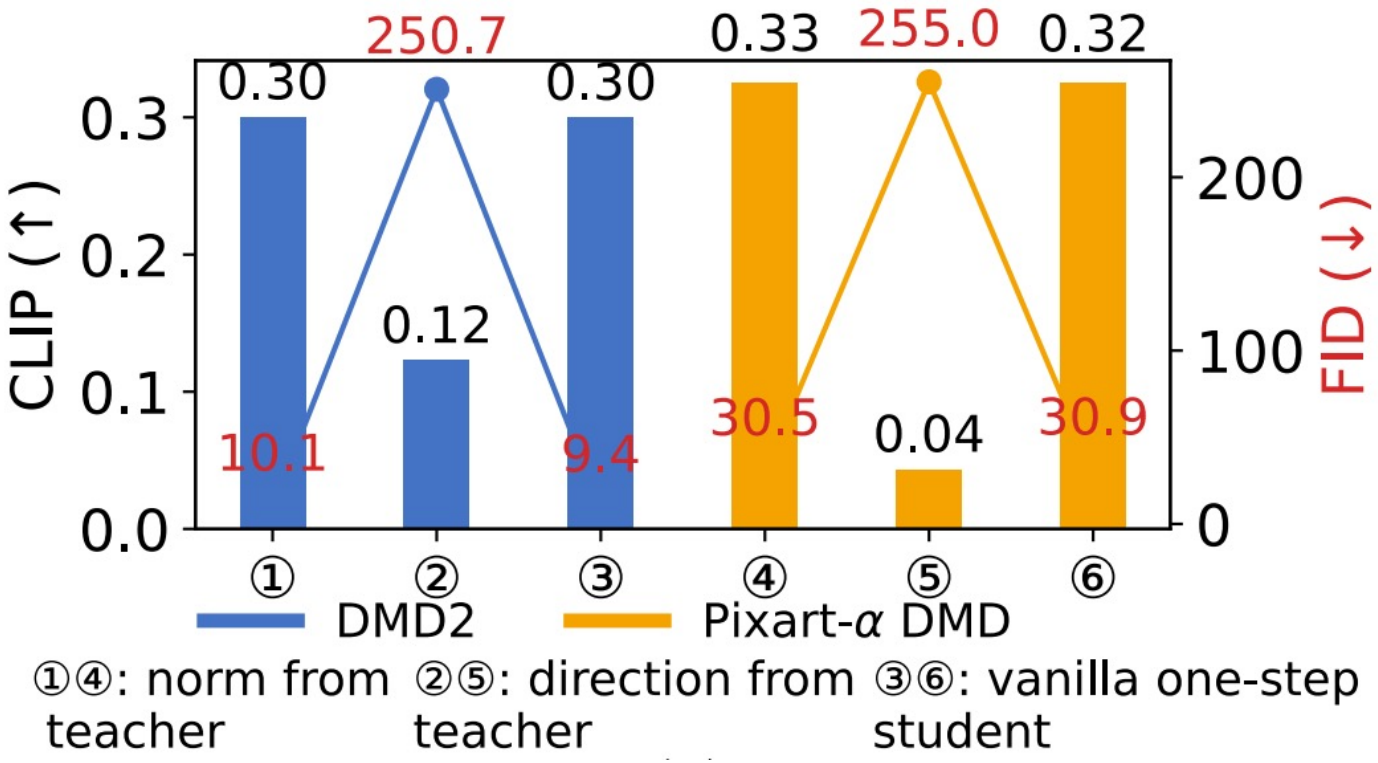
Low-Rank Property of Direction Changes

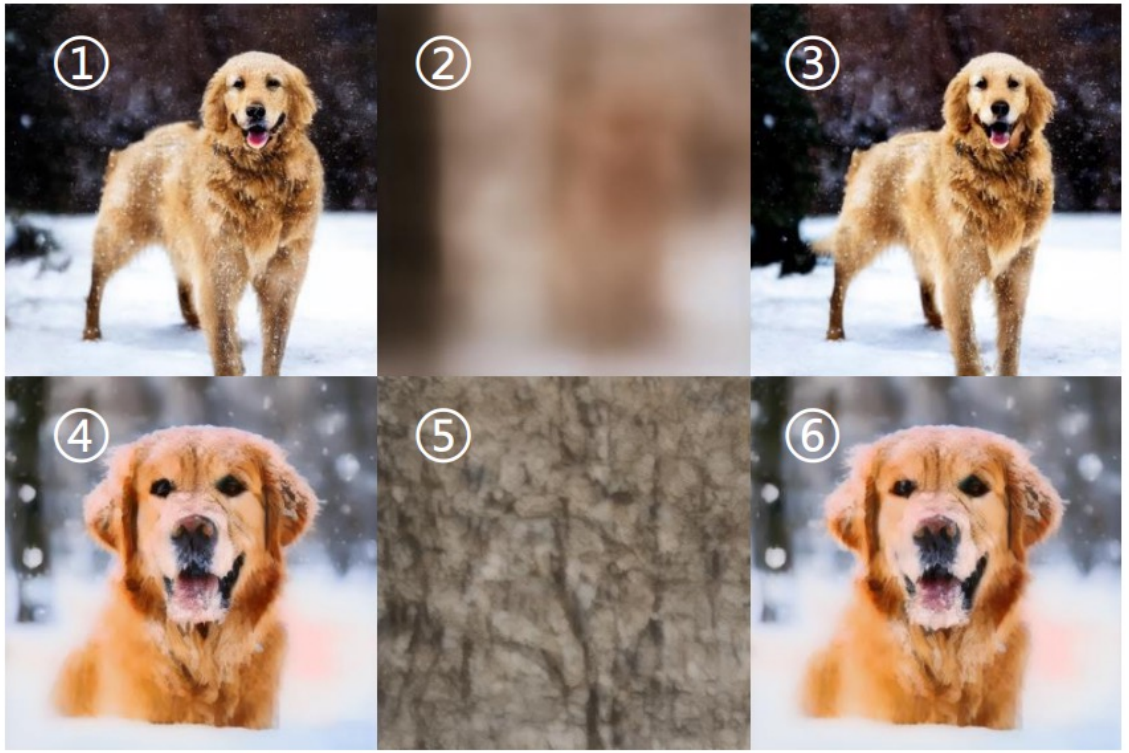


The residual of the direction matrix directly reveals its low-rank property.

Research Motivation

Direct Verification

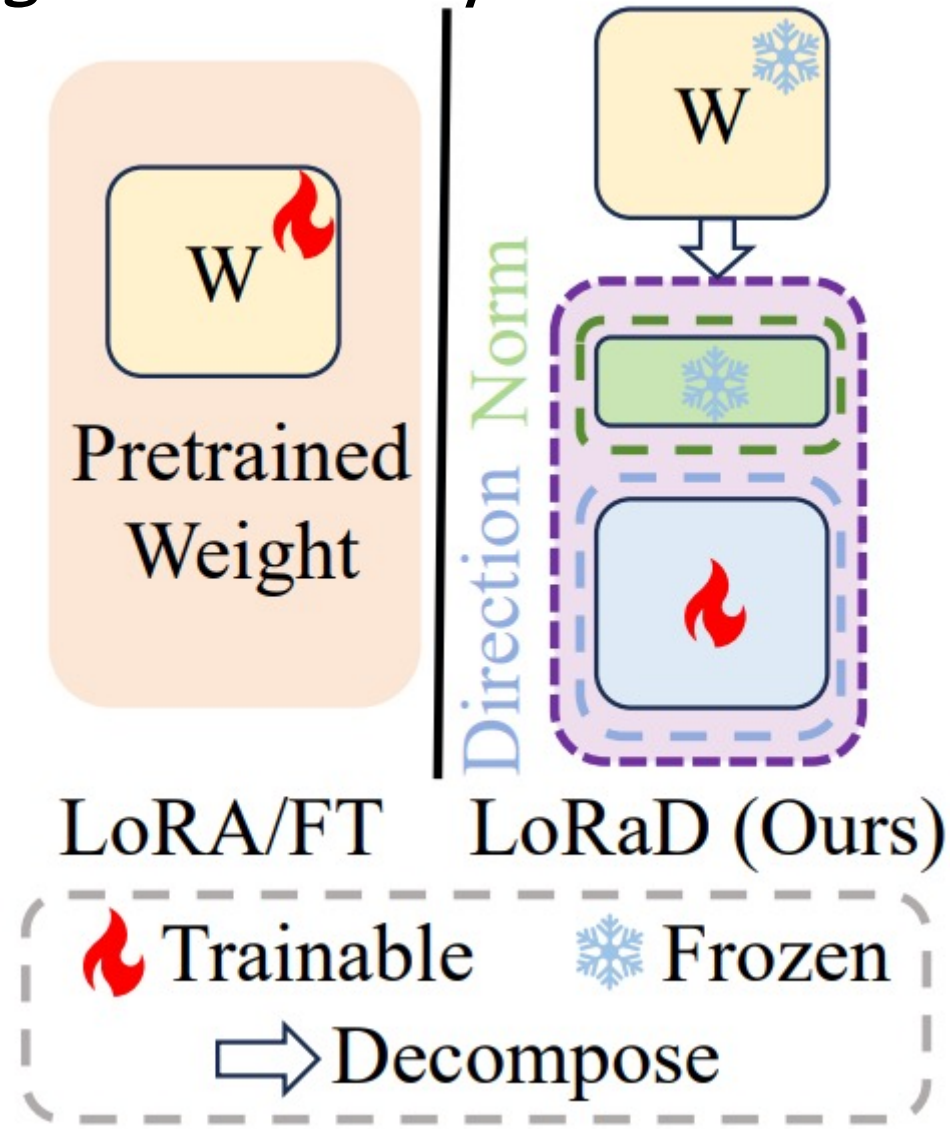




- Replacing only the norm → The model is barely affected.
- Replacing only the direction → The model performance drops significantly or changes noticeably.

Research Motivation

High-Level Comparison with LoRA

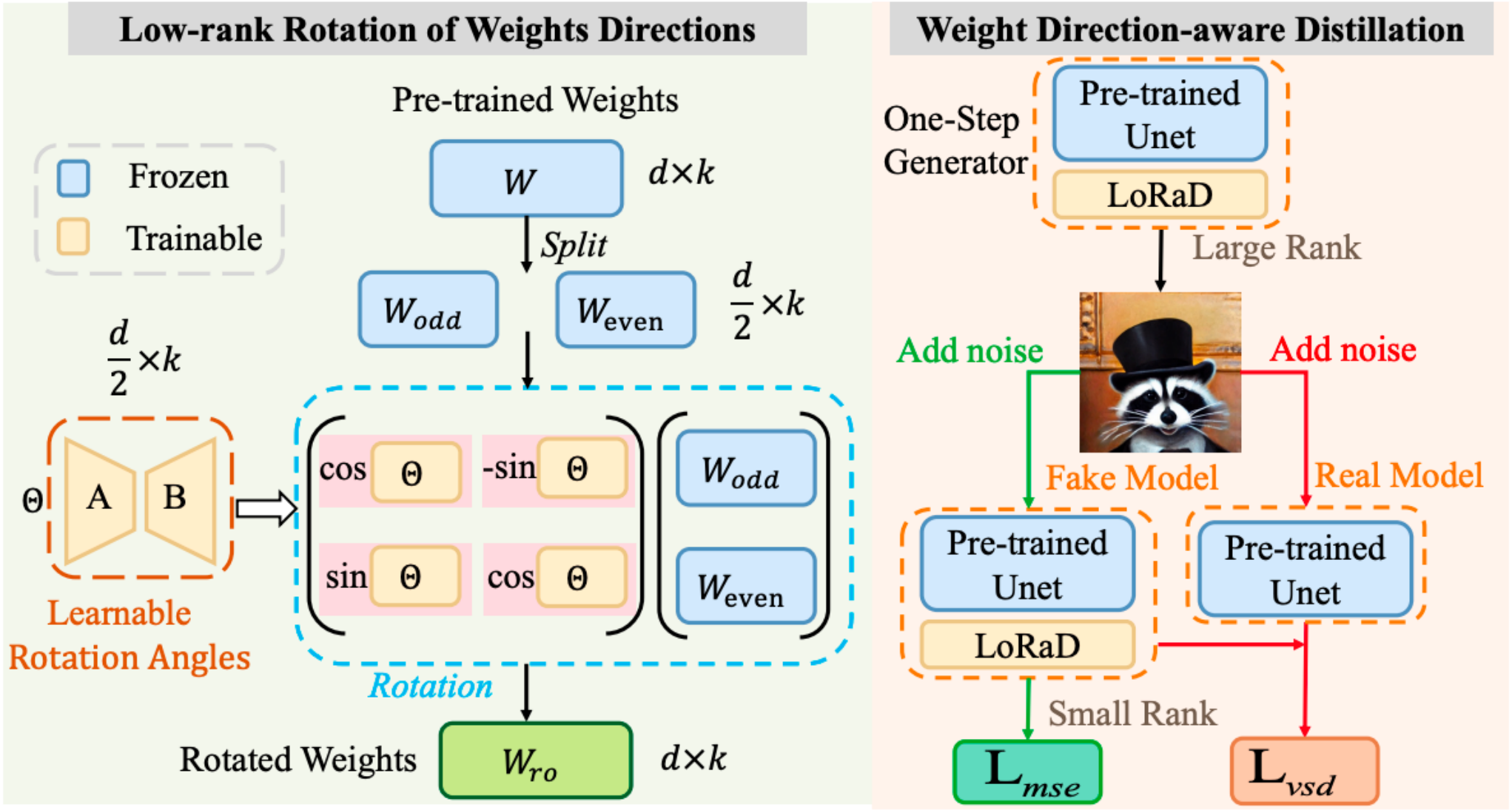


Problems of LoRA and FT:

- Difficult Optimization
- Slow Convergence
- Unstable Convergence
- Overfitting

Research Method

Pipeline



Experimental Results

Quantitative Results of Zero-shot T2I Generation

Table 1. Quantitative comparison of WaDi and other methods on zero-shot COCO 2014 results. * indicates our reproduced results, and ≀ indicates results using the official pre-trained models. '-' denotes unknown. Best and second-best scores are in **bold** and underline, respectively. "Image-free" refers to training without supervision from real images.

Method	#Params	NFEs	Type	Trainable params	FID ↓	CLIP ↑	Precision ↑	Recall ↑	Image-free?	Training Data
Stable Diffusion 1.5-based backbone										
SD 1.5 ($cfg = 3.0$)	860M	25	U-Net	860M	8.78	0.30	0.59	0.53	✗	5B
LCM-LoRA≀	860M	1	LoRA	67.50M	77.73	0.24	0.22	0.15	✗	12M
InstaFlow	860M	1	U-Net	860M	13.10	0.28	0.53	0.45	✗	3.2M
UFOGen	860M	1	U-Net	860M	12.78	-	-	-	✗	12M
DMD	860M	1	U-Net	860M	11.49	**0.32**	-	-	✗	3M
DMD2*	860M	1	U-Net	860M	12.96	0.30	0.60	0.47	✓	1.4M
SiD-LSG*	860M	1	U-Net	860M	14.27	0.30	0.56	**0.48**	✓	1.4M
PCM	860M	1	U-Net	860M	17.91	0.29	-	-	✗	3M
Hyper-SD≀	860M	1	LoRA	67.25M	22.90	0.31	**0.62**	0.25	✗	-
YOSO≀	860M	1	LoRA	67.25M	23.68	0.29	0.56	0.36	✗	4M
WaDi	860M	1	LoRaD	83.80M	**10.79**	0.31	**0.62**	**0.48**	✓	1.4M
Stable Diffusion 2.1-based backbone										
SD 2.1 ($cfg = 3.0$)	865M	1	U-Net	865M	9.60	0.32	0.59	0.50	✗	5B
SD-Turbo≀	865M	1	U-Net	865M	16.14	**0.33**	**0.65**	0.35	✗	-
Swiftbrush	865M	1	U-Net	865M	16.67	0.29	0.47	0.46	✓	1.4M
Swiftbrushv2*	865M	1	U-Net+LoRA	884.14M	15.98	**0.33**	0.58	0.47	✓	1.4M
SiD-LSG*	865M	1	U-Net	865M	15.17	0.30	0.56	0.46	✓	1.4M
TiUE≀	865M	1	U-Net	865M	13.49	0.31	0.59	**0.48**	✓	1.4M
WaDi	865M	1	LoRaD	94.43M	**12.34**	0.31	0.60	**0.48**	✓	1.4M
PixArt-α-based backbone										
PixArt-α ($cfg = 4.5$)≀	610.86M	20	DiT	610.86M	8.75	0.32	0.75	0.45	✗	25M
Swiftbrush*	610.86M	1	DiT	610.86M	29.89	0.28	0.50	0.26	✓	1.4M
PG-SB*	610.86M	1	DiT	610.86M	25.58	0.28	0.53	0.27	✓	1.4M
WaDi	610.86M	1	LoRaD	81.22M	**18.99**	**0.30**	**0.64**	**0.29**	✓	1.4M

Experimental Results

Quantitative Results of Zero-shot T2I Generation

Table A6. Quantitative comparison of WaDi and other methods on zero-shot COCO 2017 results. * indicates our reproduced results, and ≀ indicates results using the official pre-trained models. '-' denotes unknown. Best and second-best scores are in **bold** and underline, respectively.

Method	#Params	NFEs	Type	Trainable params	FID ↓	CLIP ↑	Precision ↑	Recall ↑	Image-free?	Training Data
Stable Diffusion 1.5-based backbone										
SD 1.5 ($cfg = 3.0$) [28]	860M	25	U-Net	860M	19.80	0.31	0.64	0.60	✗	5B
LCM-LoRA [21]≀	860M	1	LoRA	67.50M	89.65	0.24	0.22	0.24	✗	12M
InstaFlow [19]	860M	1	U-Net	860M	23.49	**0.31**	0.53	0.46	✗	3.2M
UFOGen [38]	860M	1	U-Net	860M	22.50	**0.31**	-	-	✗	12M
DMD2 [39]*	860M	1	U-Net	860M	23.30	0.30	0.60	0.49	✓	1.4M
SiD-LSG [43]*	860M	1	U-Net	860M	24.22	0.30	0.60	0.52	✓	1.4M
Hyper-SD [27]≀	860M	1	LoRA	67.25M	32.49	**0.31**	0.52	0.33	✗	-
YOSO [22]≀	860M	1	LoRA	67.25M	33.54	0.29	0.50	0.44	✗	4M
WaDi	860M	1	LoRaD	83.80M	**20.86**	**0.31**	**0.63**	**0.54**	✓	1.4M
Stable Diffusion 2.1-based backbone										
SD 2.1 ($cfg = 3.0$) [28]	865M	25	U-Net	865M	19.66	0.32	0.66	0.57	✗	5B
SD-Turbo [31]≀	865M	1	U-Net	865M	26.36	**0.34**	**0.69**	0.47	✗	-
Swiftbrush [25]	865M	1	U-Net	865M	26.87	0.32	0.61	0.44	✓	1.4M
Swiftbrushv2 [7]*	865M	1	U-Net+LoRA	884.14M	25.96	0.33	0.65	0.45	✓	3.3M
SiD-LSG [43]*	865M	1	U-Net	865M	25.02	0.30	0.62	0.51	✓	1.4M
WaDi	865M	1	LoRaD	94.43M	**22.62**	0.31	0.65	**0.53**	✓	1.4M
PixArt-α-based backbone										
PixArt-α ($cfg = 4.5$) [6]≀	0.6B	20	DiT	0.6B	20.85	0.27	0.65	0.59	✗	25M
Swiftbrush [25]*	0.6B	1	DiT	0.6B	41.07	0.28	0.53	0.35	✓	1.4M
PG-SB [26]*	0.6B	1	DiT	0.6B	35.84	0.28	0.57	0.36	✓	1.4M
WaDi	0.6B	1	LoRaD	81.22M	**28.91**	**0.30**	**0.62**	**0.37**	✓	1.4M

Experimental Results

Efficiency Analysis of Zero-shot T2I

Table 3: Comparison of inference and training times of our method vs. other methods on the zero-shot benchmark of COCO 2014. * indicates our reproduced results, and ≀ indicates results using the official pre-trained models. '-' denotes unknown. Best and second-best scores are in **bold** and underline, respectively.

Method	NFEs	Type	Trainable params	FID ↓	CLIP ↑	Image-free?	Inference	A100 Days
Stable Diffusion 1.5-based backbone								
SD 1.5 ($cfg = 3.0$) [4]	25	U-Net	860M	8.78	0.30	✗	1.11s	4783
LCM-LoRA [21]≀	1	LoRA	67.50	77.73	0.24	✗	0.11s	1.3
InstaFlow [24]	1	U-Net	860M	13.10	0.28	✗	0.11s	183.2
UFOGen [25]	1	U-Net	860M	12.78	-	✗	-	-
DMD [26]	1	U-Net	860M	11.49	**0.32**	✗	0.11s	108
DMD2 [18]*	1	U-Net	860M	12.96	0.30	✓	0.11s	5.1
SiD-LSG [30]*	1	U-Net	860M	14.27	0.30	✓	0.11s	6.4
PCM [22]	1	U-Net	860M	17.91	0.29	✗	-	unk
Hyper-SD [32]≀	1	LoRA	67.25M	22.90	0.31	✗	0.11s	33.3
YOSO [31]≀	1	LoRA	67.25M	23.68	0.29	✗	0.11s	20
DKD	1	LoRaD	83.80M	**10.79**	0.31	✓	0.11s	2.1
Stable Diffusion 2.1-based backbone								
SD 2.1 ($cfg = 3.0$) [4]	25	U-Net	865M	9.60	0.32	✗	1.04s	8332
SD-Turbo [16]≀	1	U-Net	865M	16.14	**0.33**	✗	0.11s	-
Swiftbrush [28]	1	U-Net	865M	16.67	0.29	✓	0.11s	4.1
Swiftbrushv2 [29]*	1	U-Net+LoRA	884.14M	15.98	**0.33**	✓	0.11s	24.1
SiD-LSG [30]*	1	U-Net	865M	15.17	0.30	✓	0.11s	6.4
DKD	1	LoRaD	94.43M	**12.34**	0.31	✓	0.11s	2.1
PixArt-α-based backbone 256×256								
PixArt-α ($cfg = 4.5$) [49]≀	20	DiT	0.6B	8.75	0.32	✗	0.59s	753
Swiftbrush [28]*	1	DiT	0.6B	29.89	0.28	✓	0.05s	2.6
PG-SB [76]*	1	DiT	0.6B	25.58	0.28	✓	0.05s	2.6
DKD	1	LoRaD	81.22M	**18.99**	**0.30**	✓	0.05s	1.6

Experimental Results

Visualization Results of WaDi (SD2.1)

U-Net-based

Experimental Results

Visualization Results of WaDi (SD1.5)

U-Net-based

Experimental Results

Visualization Results of WaDi (PixArt-α)

DiT-based

Experimental Results

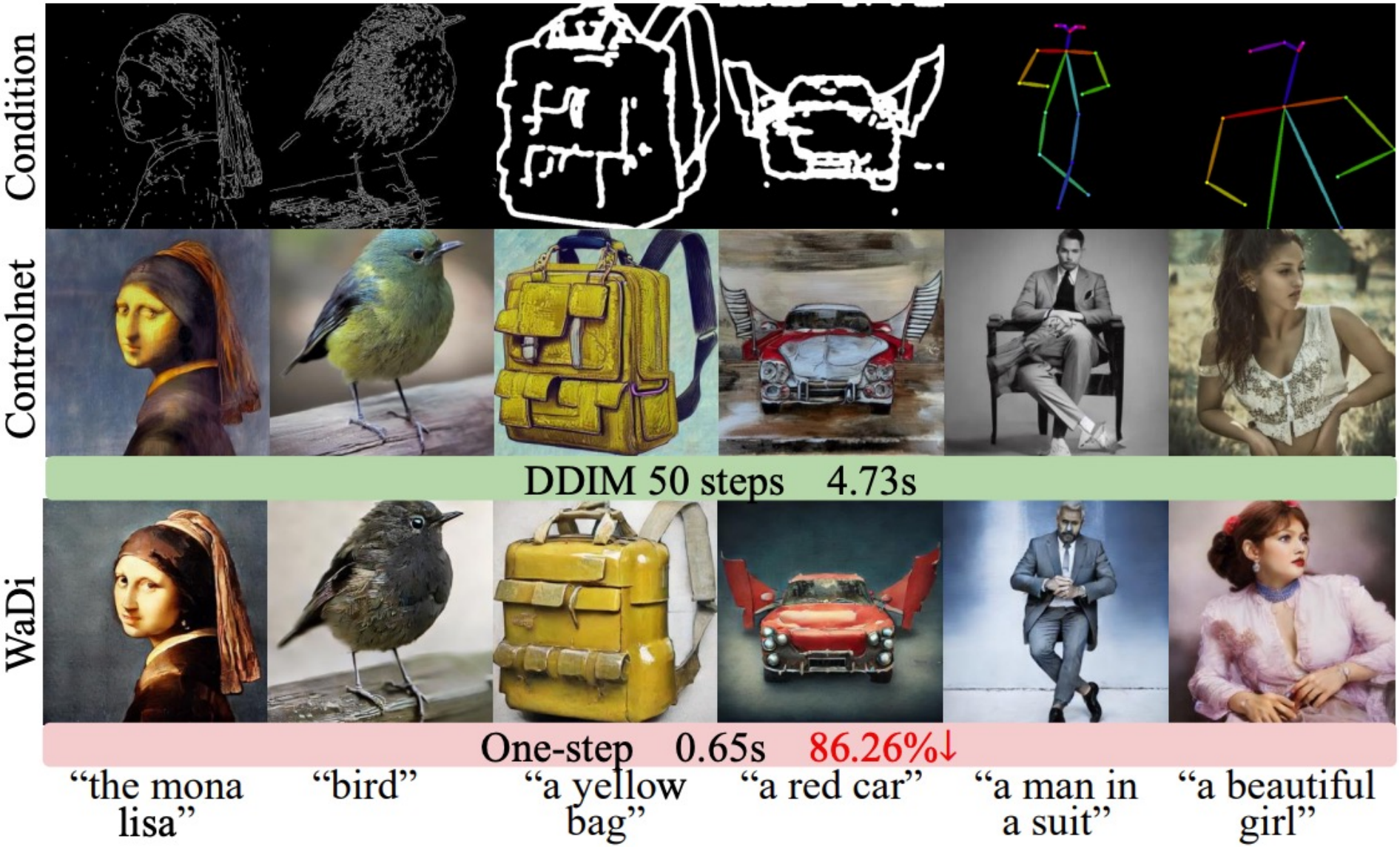
Qualitative Results of Zero-shot T2I Generation

SD 1.5 based

SD 2.1 based

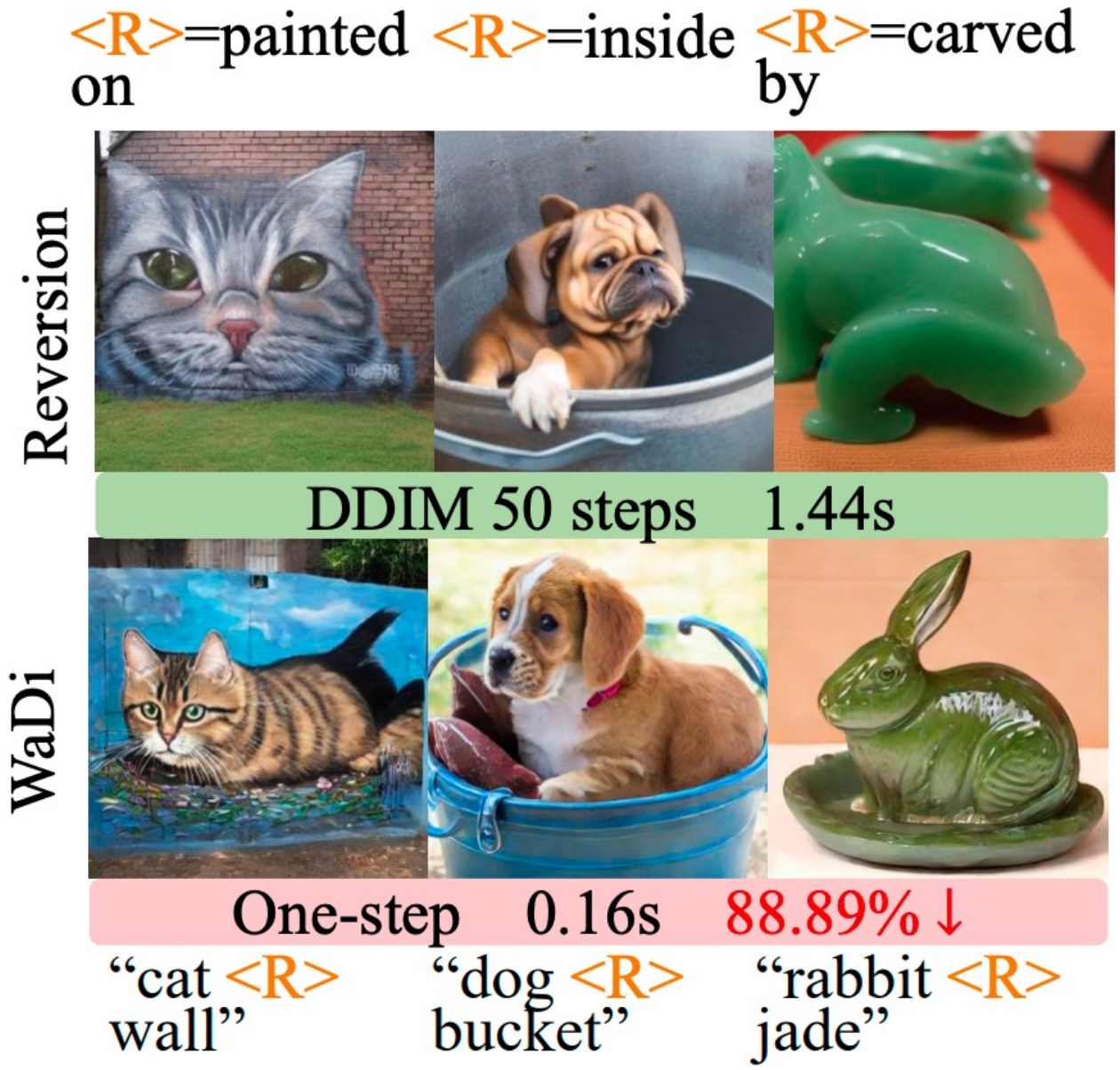
Experimental Results

Controllable Generation



Experimental Results

Relation Inversion



Experimental Results

High-Resolution Synthesis

Experimental Results

Image Customization

Figure 7. Quality results by Dreambooth with or without LoRaD.

Experimental Results

User Study

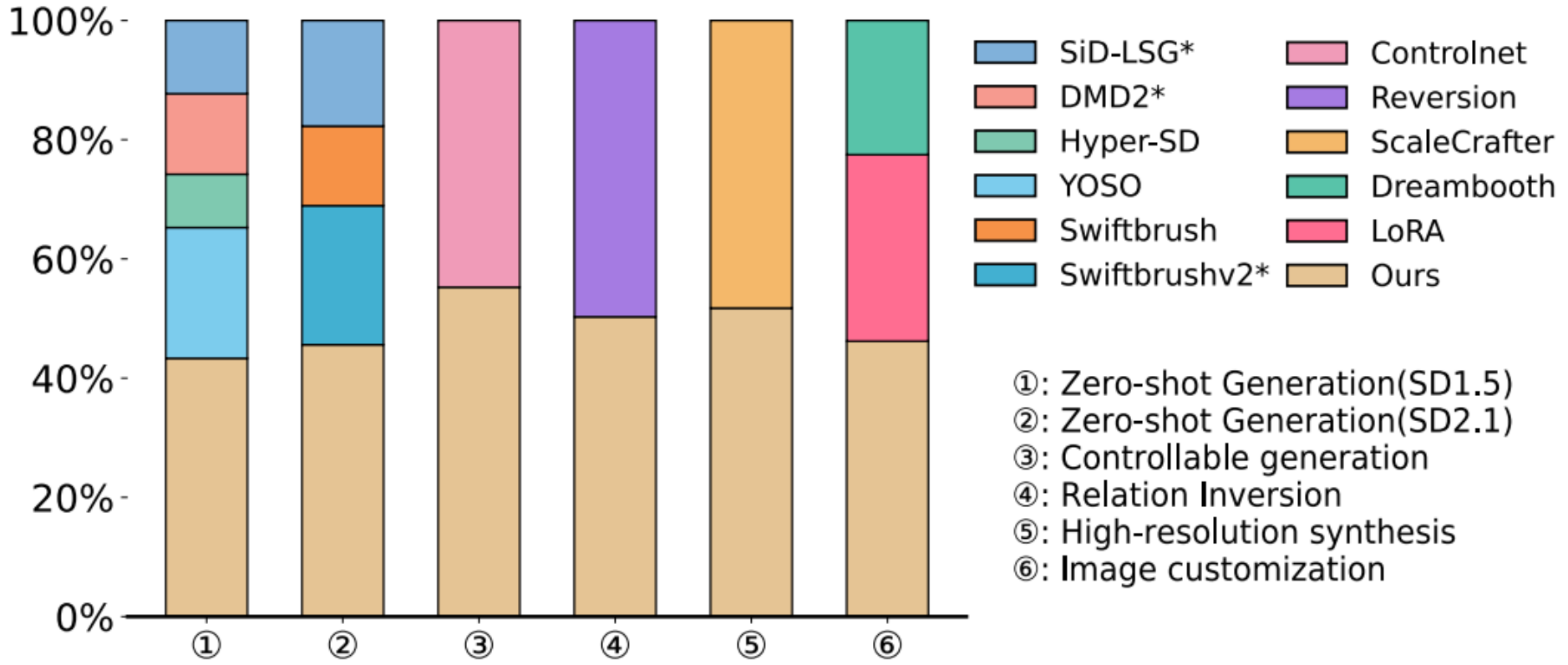


Figure 10: User study results compared to other methods.

DKD Outperforms Existing Baseline Methods

Experimental Results

Ablation Study

Table R4. Orthogonal maps.

Method (**COCO 17**)	FID	CLIP	A100 Days
orthogonal (matrix_exp)	33.43	0.27	12.1
orthogonal (cayley)	26.72	0.29	10.4
orthogonal (house-holder)	121.15	0.24	104.4 (2.1)
WaDi	**20.86**	**0.31**	**2.1**

Compared with the simplest baseline, orthogonal matrix constraints, LoRaD achieves better convergence and performance.

Experimental Results

Ablation Study

Table 2. Ablation study on the impact of adapter type in WaDi (SD 1.5, **VSD loss**) on the COCO 2017 dataset."NM" and "DM" denote the norm mean and direction mean for all layers, respectively.

Type	#Params	FID	CLIP	NM	DM
LoRA	120.9M	25.27	0.29	0.06	0.83
DoRA	121.2M	26.56	0.30	0.03	0.55
DoRA (frozen norm)	120.9M	24.52	0.30	-	0.92
FT (DMD2)	860.0M	23.30	0.30	0.10	2.21
LoRaD	**83.8M**	**20.86**	**0.31**	-	2.89

LoRaD achieves the best FID and CLIP scores.

Experimental Results

Ablation Study

Table A1. Comparison of different methods in terms of memory, number of trainable parameters, FID, CLIP score, and latency.

Type	Memory (M)	#Train Param.	FID	CLIP	Latency
LoRA	1259	120.9M	25.27	0.29	0.11s
LoRaD	2021	**83.8M**	**20.86**	**0.31**	0.11s
FT (DMD2)	17397	860M	23.30	0.30	0.11s

LoRaD's memory efficiency lies between LoRA and FT.

Experimental Results

Ablation Study

Table 2: Ablation study on the impact of the rank on WaDi (SD 1.5) on COCO 2014 dataset.

Setting	Rank				FID	CLIP
	Student	#Params	Fake model	#Params		
A	64	20.95M	32	9.38M	13.64	0.30
B	128	41.90M	32	9.38M	13.16	0.29
C	256	83.80M	32	9.38M	**10.79**	**0.31**
D	512	167.59M	32	9.38M	12.75	0.30
E	256	83.80M	16	4.69M	17.53	0.29
F	256	83.80M	64	18.76M	16.98	**0.31**

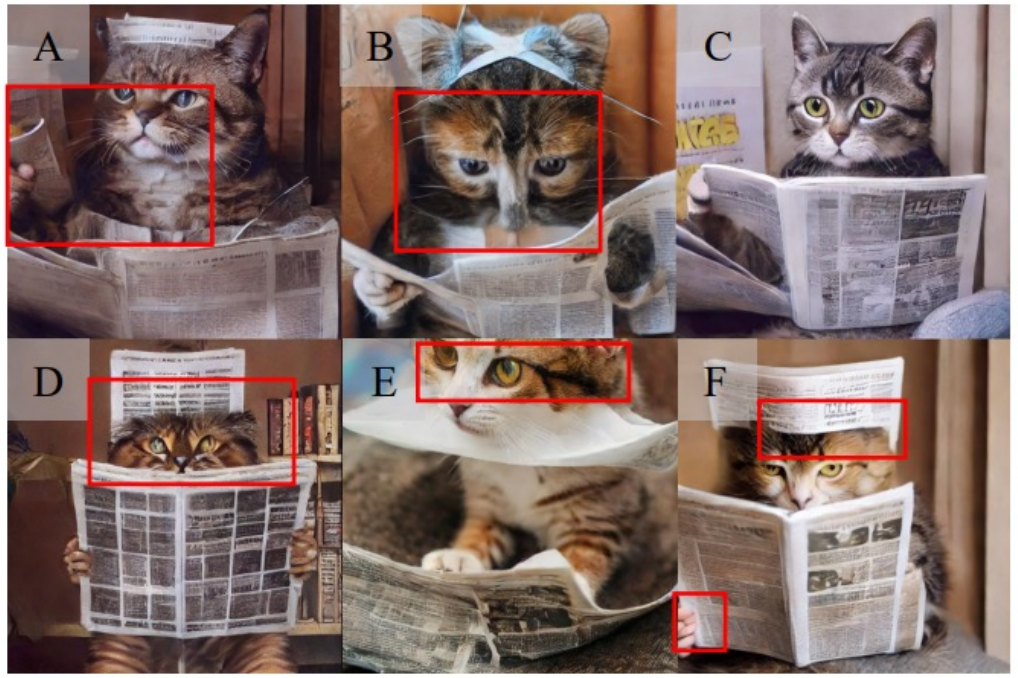
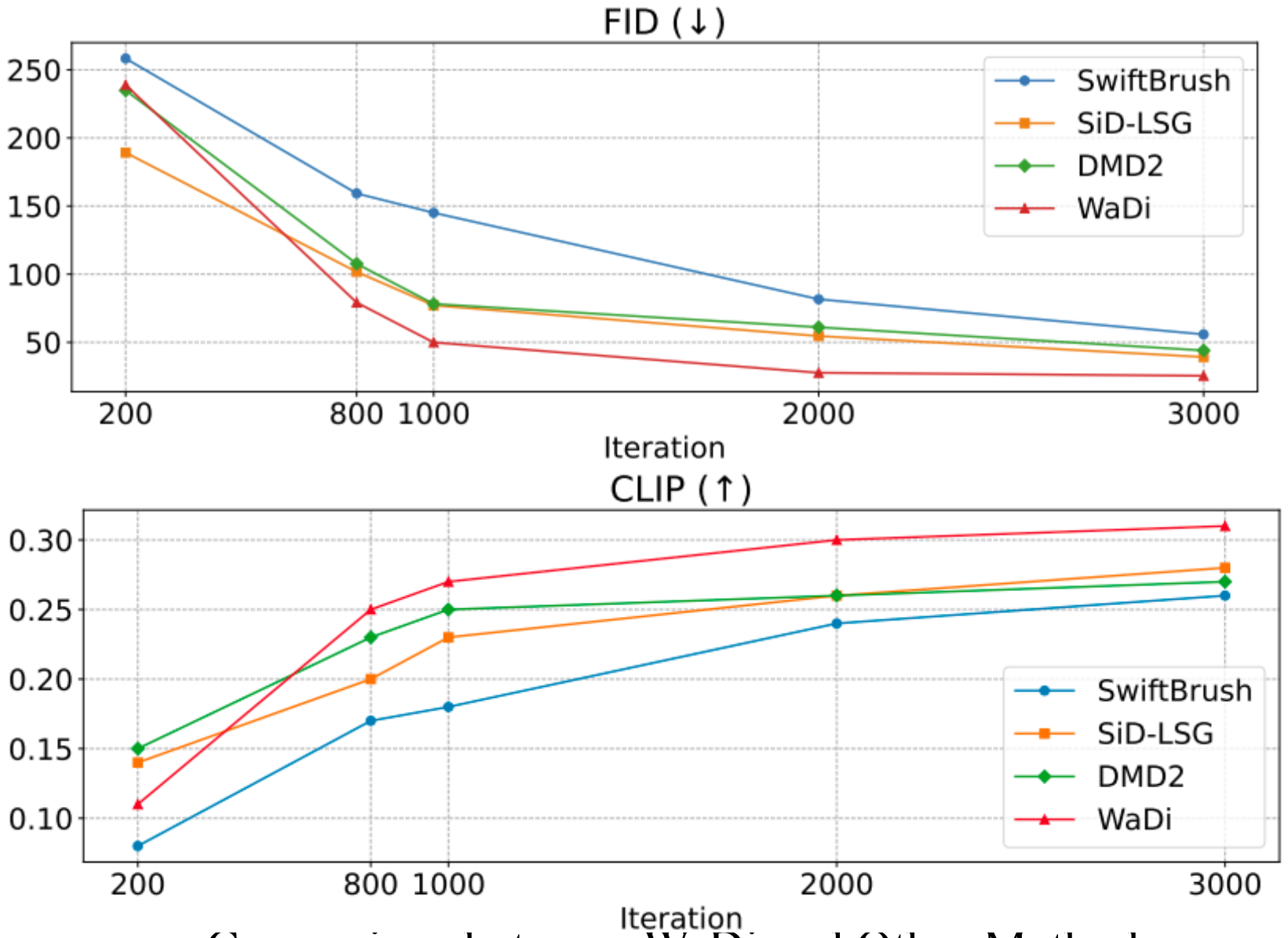


Figure 9: One-step image generation with various settings.

Configuration C achieves a trade-off between performance and parameter efficiency.

Experimental Results

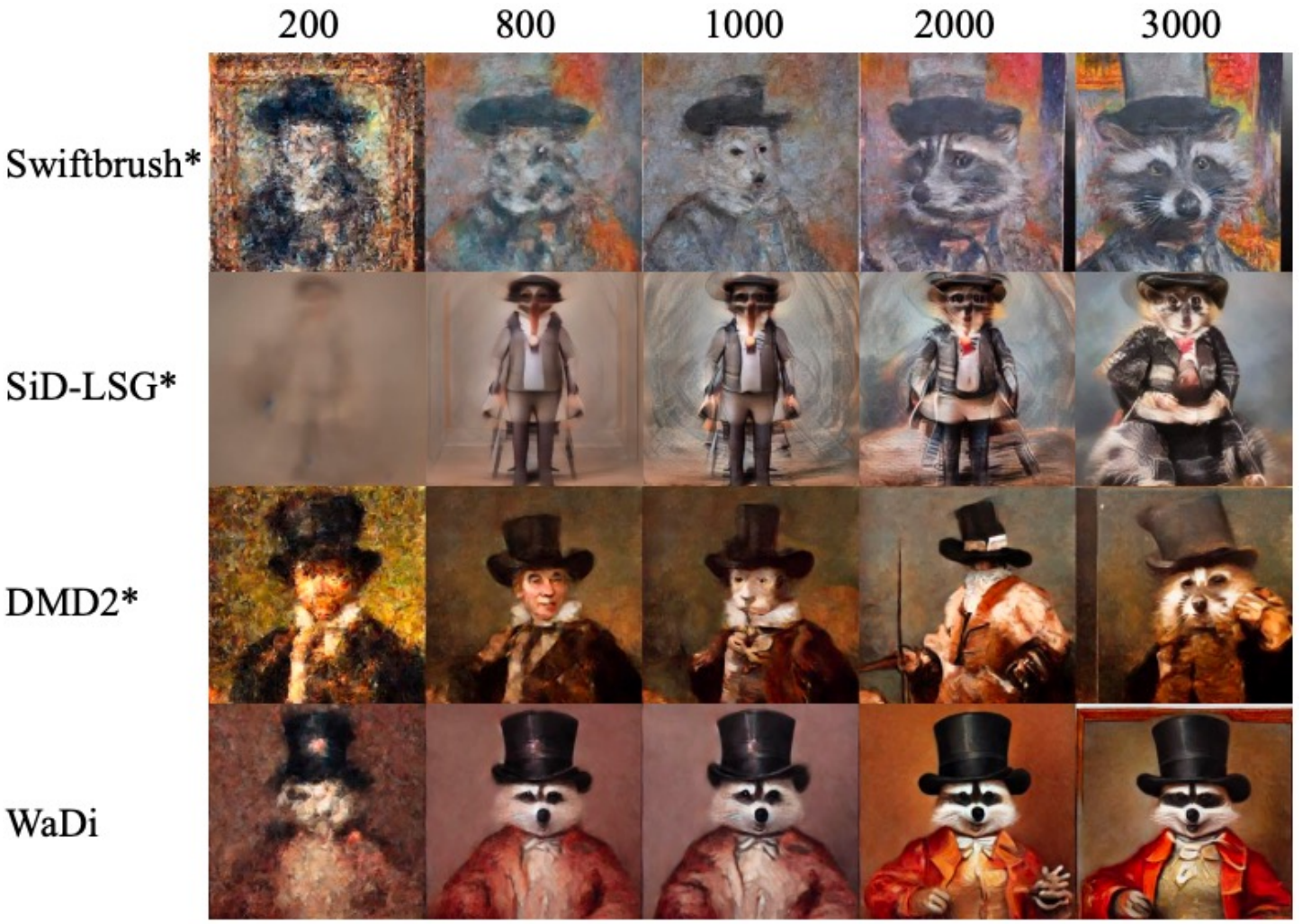
Convergence Analysis



Convergence Comparison between WaDi and Other Methods.

Experimental Results

Convergence Analysis



"A racoon wearing formal clothes, wearing a tophat. Oil painting in the style of Rembrandt"

Visual Comparison of WaDi's Convergence Behavior.

PCA Lab Work Report

Thank you for your attention! I sincerely welcome your comments and suggestions.