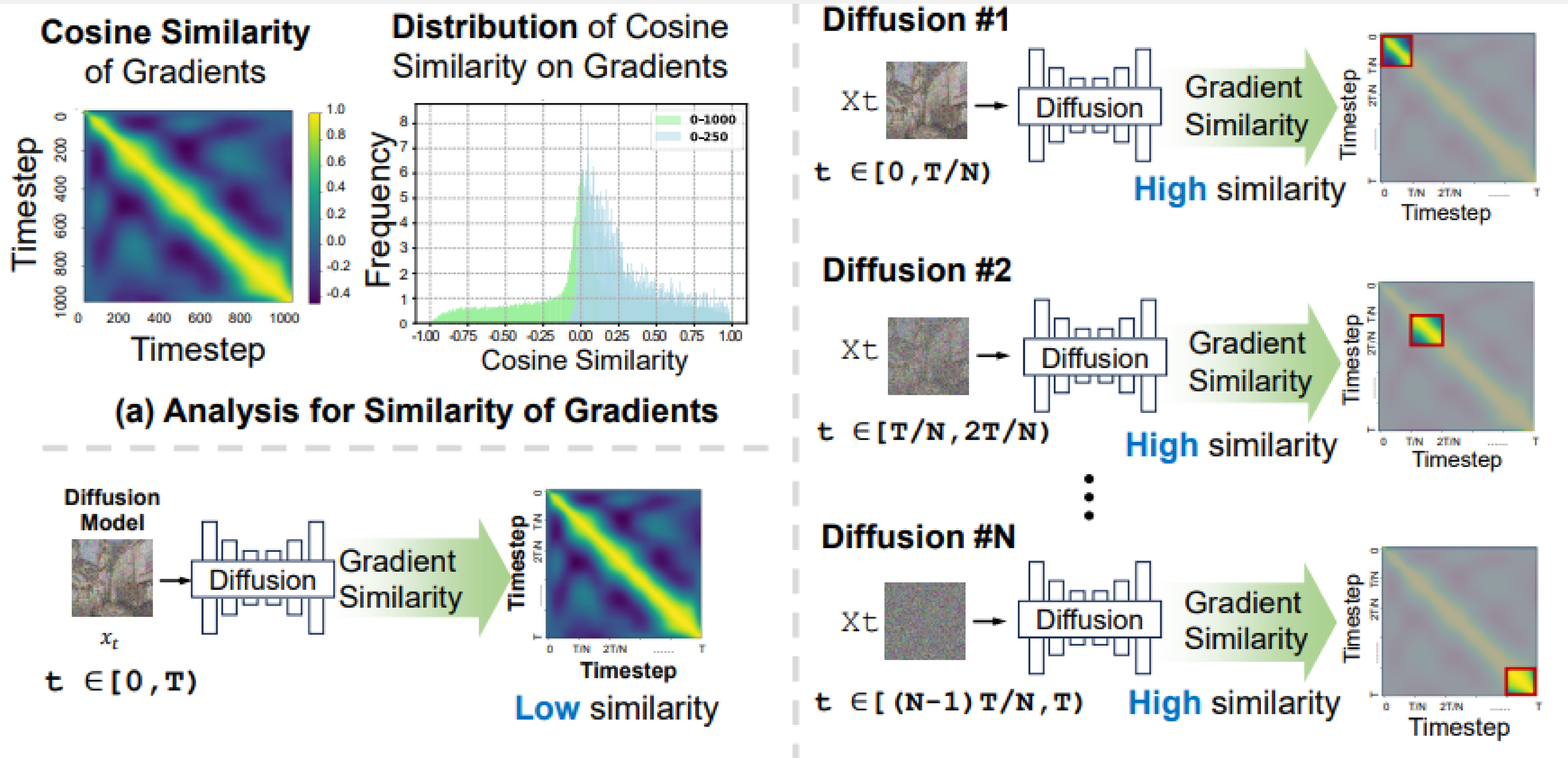
Decouple-Then-Merge: Finetune Diffusion Models as Multi-Task Learning

Qianli Ma, Xuefei Ning, Dongrui Liu, Li Niu, Linfeng Zhang

Shanghai Jiao Tong University, Tsinghua University, Shanghai AI Laboratory

Project Page Wechat

Introduction



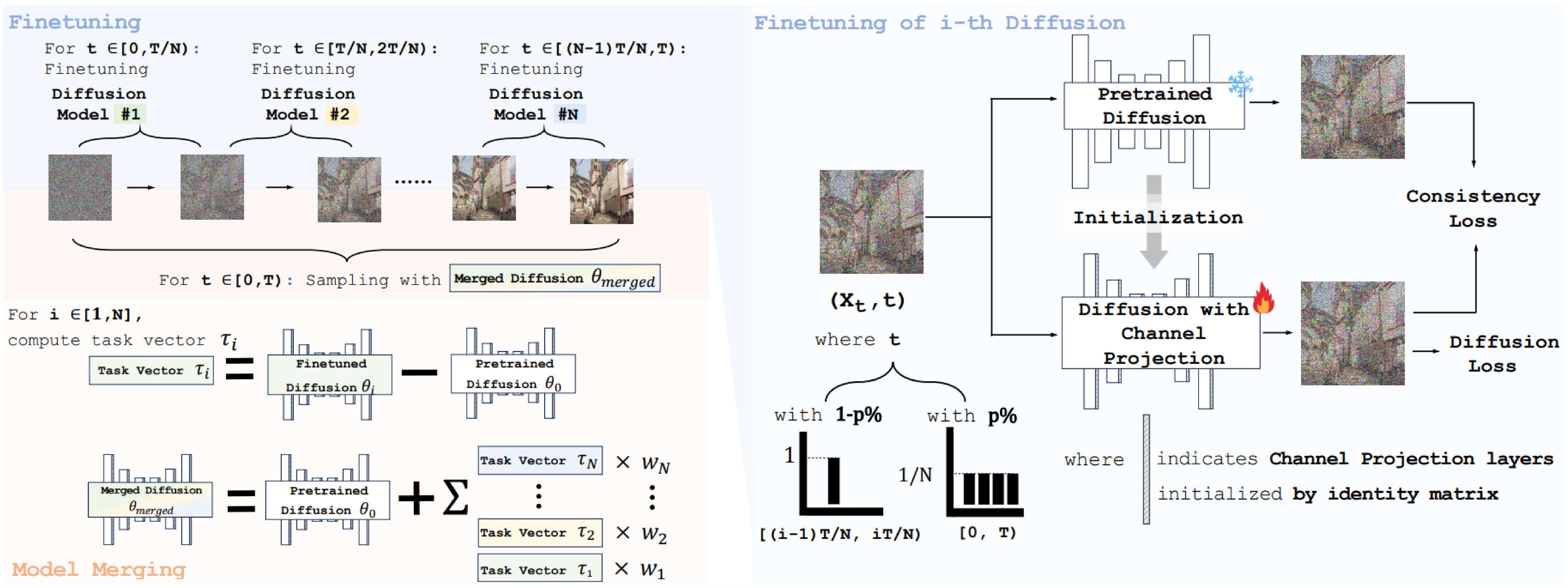
(a) Analysis for Similarity of Gradients

(b)Traditional Diffusion Training Paradigm

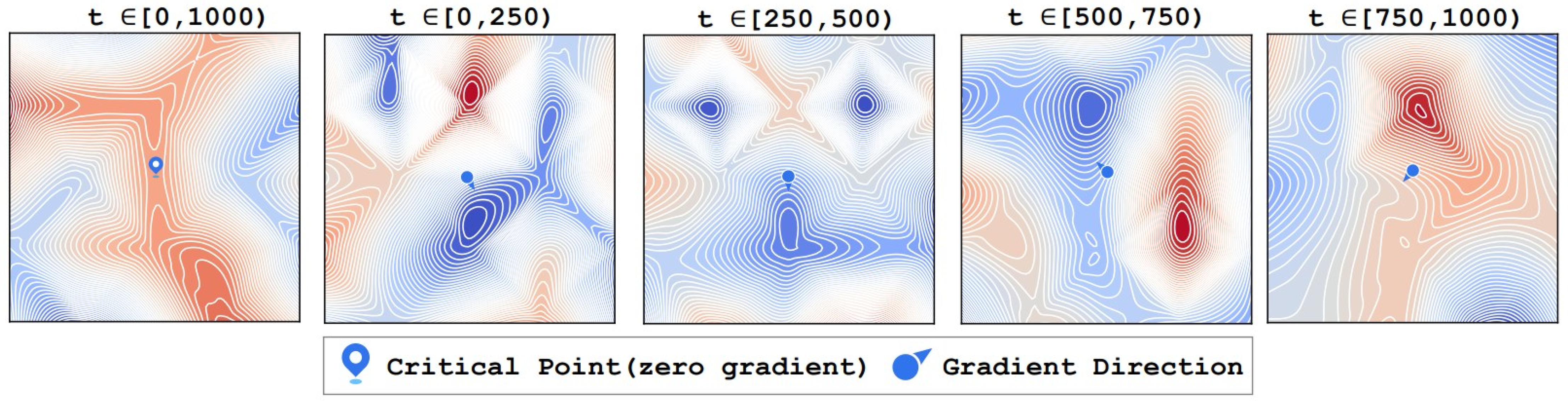
(c) Our Training Paradigm

- Diffusion model learns to denoise across multiple timesteps. However, training a single diffusion model across all timesteps can cause **gradient conflicts**, leading to suboptimal performance.
- To solve this, we propose Decouple-then-Merge (DeMe), a novel finetuning framework that **decouples** training across different timestep ranges to reduce interference, and then **merges** the models in parameter space to retain efficiency. With simple techniques like consistency loss, probabilistic sampling and channel-wise projection, DeMe significantly improves generation quality without extra inference cost.
- We further provide theoretical analysis and conduct extensive experiments to demonstrate its effectiveness

Method: Decouple Training & Models Merging



- In Decoupled training process, **Channel-wise Projection** enables the model to capture the channel-wise difference in different timesteps. **Consistency Loss** preserves the initial knowledge of the diffusion model. **Probabilistic Sampling** enables the finetuned model to mainly learn from its corresponding timestep range, but still possible to preserve the knowledge in the other timestep ranges
- Finally, we **merge** multiple fine-tuned diffusion models into single diffusion model in the parameter space, obtain an enhanced merged model without extra inference cost.



Comparison Experiments

Method	MS-COCO		ImageNet		PartiPrompts	#Iterations
	FID↓	CLIP Score↑	FID↓	CLIP Score↑	CLIP Score↑	
Before-finetuning [41]	13.42	29.88	27.62	27.07	29.78	-
SNR+1 [46]	13.92	29.96	27.56	27.03	29.86	80K
Trun-SNR [46]	13.93	29.95	27.60	27.05	29.85	80K
Min-SNR-γ [13]	13.92	29.93	27.59	27.02	29.87	80K
P2 Weighting [3]	13.23	29.93	26.92	26.44	29.50	80K
ANT-NashMTL [10]	13.39	29.81	27.41	26.99	29.90	80K
ANT-UW [10]	13.17	29.94	26.91	26.78	29.98	80K
DeMe (Before Merge)	12.78 (−0.64)	29.85 (−0.03)	26.36 (−1.26)	26.90 (−0.17)	30.02 (+0.24)	20K×4
DeMe (After Merge)	**13.06** (−0.36)	**30.11** (+0.23)	**27.23** (−0.39)	**27.09** (+0.02)	**29.98** (+0.20)	20K×4

Method	CIFAR10	LSUN-Church	LSUN-Bedroom	#Iterations
Before-finetuning [15]	4.42	10.69	6.46	-
SNR+1 [46]	5.41	10.80	6.41	80K
Trun-SNR [46]	4.49	10.81	6.42	80K
Min-SNR-γ [13]	5.77	10.82	6.41	80K
P2 Weighting [3]	5.63	10.77	6.53	80K
ANT-NashMTL [10]	4.24	10.45	6.43	80K
ANT-UW [10]	4.21	10.43	6.48	80K
DeMe (Before Merge)	3.79 (−0.63)	9.57 (−1.12)	5.87 (−0.59)	20K×4
DeMe (After Merge)	**3.51** (−0.91)	**7.27** (−3.42)	**5.84** (−0.62)	20K×4

N	Probabilistic Sampling	Consistency Loss	Channel-wise Projection	FID ↓
1	✗	✗	✗	4.40
	✗	✗	✔	4.45
8	✔	✗	✗	4.32
	✔	✔	✗	4.27
	✔	✔	✔	3.87

Prompt I: "*A graceful white horse galloping through a field of wildflowers, its mane flowing in the wind as the sun sets behind it.*"

Before Finetuning: Loss of text-image alignment: *as the sun sets behind it*

After Finetuning: Superb text-image alignment, lifelike horse

Prompt II: "*A tropical beach with crystal-clear turquoise water gently lapping against white sandy shores, tall palm trees swaying in the breeze under a clear blue sky.*"

Before Finetuning: Loss of text-image alignment: *white sandy shores*

After Finetuning: Superb text-image alignment, excellent coastal view

Prompt III: "*A dolphin leaping out of the ocean in perfect harmony, water splashing around it as the sun sets on the horizon, casting a golden glow on the sea.*"

Before Finetuning: Loss of text-image alignment: *sun sets on the horizon..*

After Finetuning: Superb text-image alignment, photorealistic illustration

Prompt IV: "*A foggy morning in a quiet countryside, a small wooden cabin surrounded by wildflowers and tall grass, the sun just beginning to rise through the mist.*"

Before Finetuning: Loss of text-image alignment: *a small wooden cabin*

After Finetuning: Superb text-image alignment, serene landscape