TFMQ-DM: Temporal Feature Maintenance Quantization for Diffusion Models

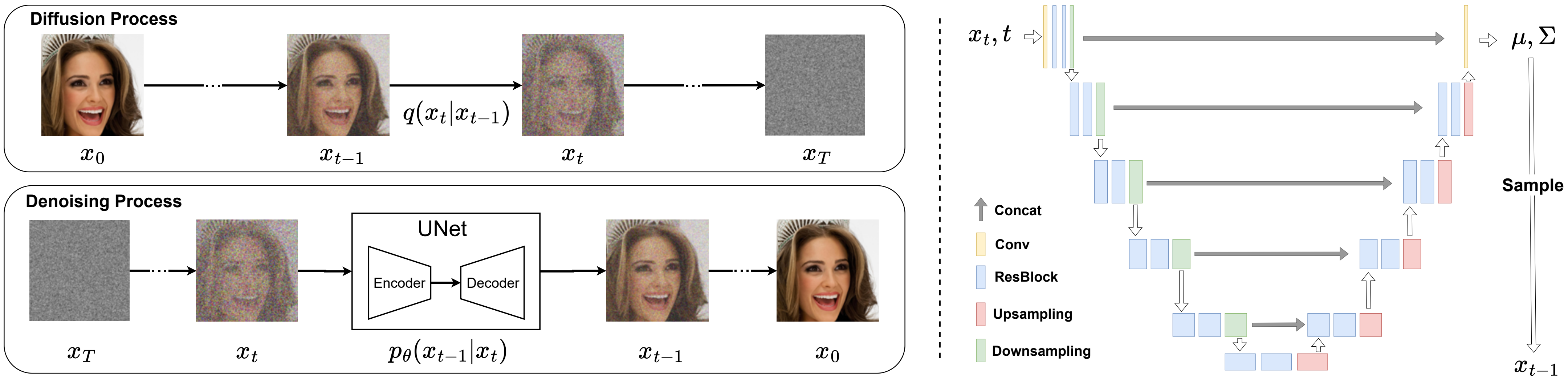
Yushi Huang*, Ruihao Gong*, Jing Liu, Tianlong Chen,
Xianglong Liu

CVPR2024 Highlight

Diffusion Models

The impact of quantized t/temporal feature

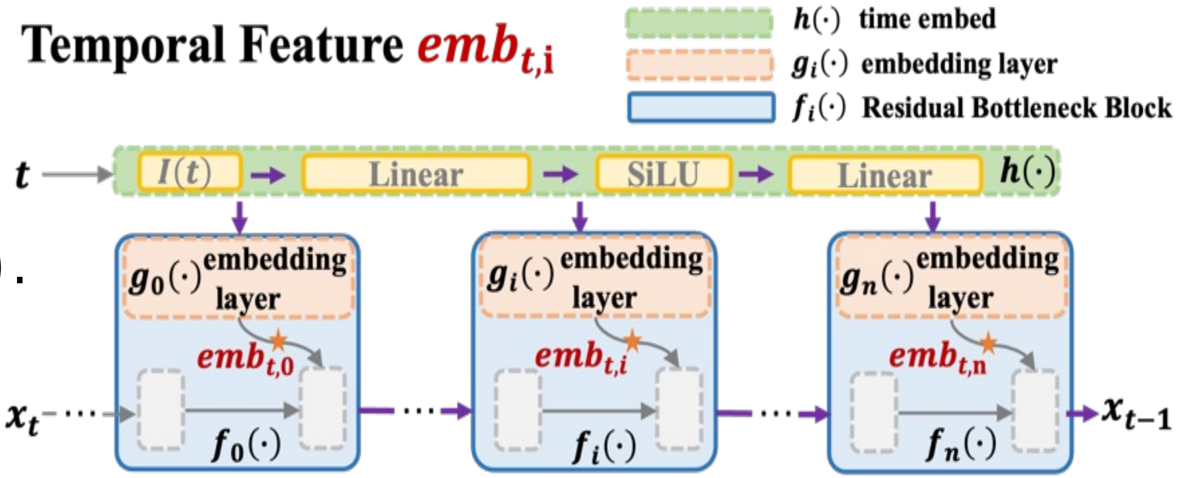
- Significantly time-steps-**varying activation range**
  - **a. How to choose calibration data?** Design for sampling at different t.
  - **b. How to determine quantization parameters?** Introduce more parameters related to t.
- **Accumulative quantization errors** with time-steps
  - **c. How to make some remedies?** Compensate forward pass errors with statistical methods or time-aware mixed precision quantization.



Left: Diffusion/Denoising process. Right: Structure of Unet.

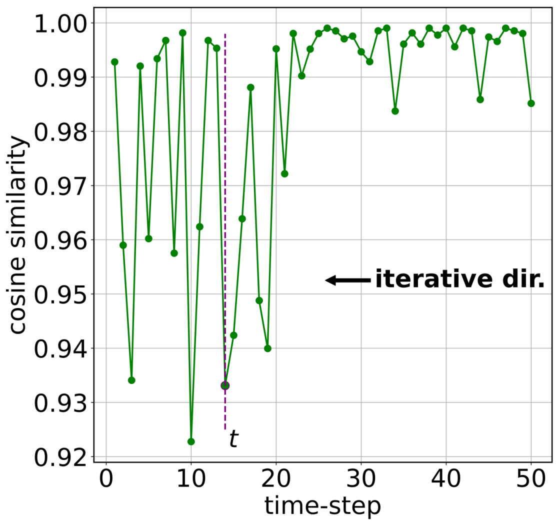
Background & New Problem

- **Temporal Feature** in Diffusion Models

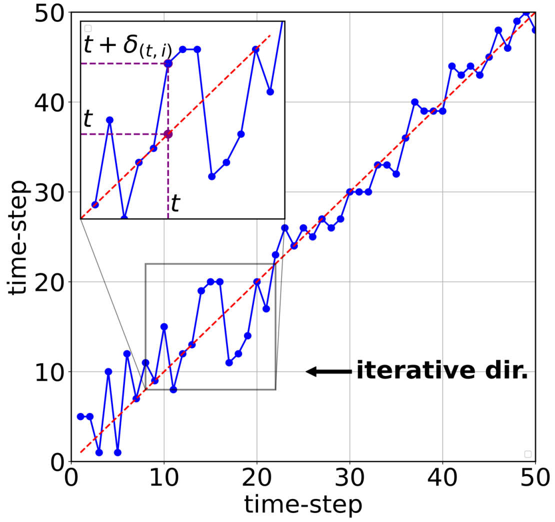


Temporal Feature in diffusion models.

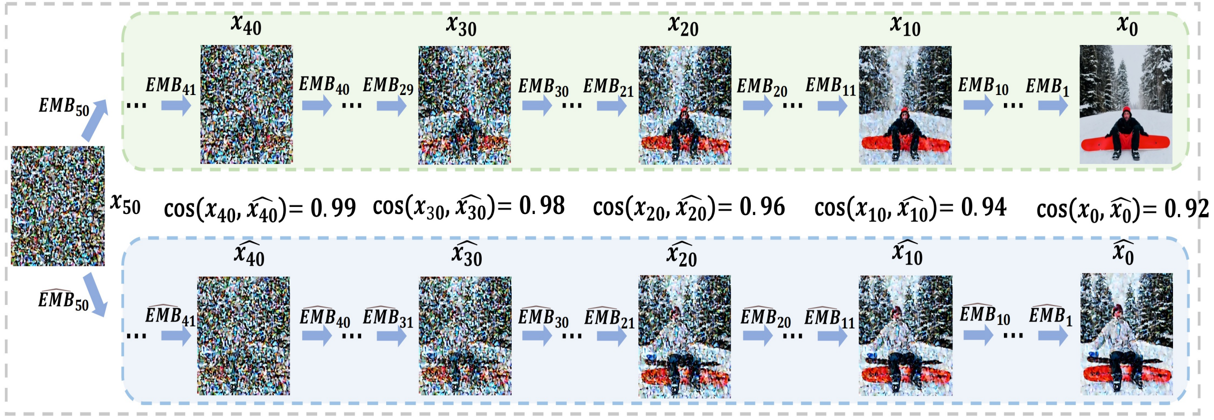
- Temporal Feature **Disturbance**
  - Temporal feature **error:** $cos\left(\boldsymbol{emb}_{t,i}\, ,\widehat{\boldsymbol{emb}}_{t,i}\right)$.
  - Temporal information **mismatch:** $t \nleftarrow\ \widehat{\boldsymbol{emb}}_{t,i}$.
  - Trajectory **deviation:** $\boldsymbol{x}_t \nRightarrow\ \boldsymbol{x}_{t-1}$.



(a) Temporal feature error.



(b) Temporal information mismatch.



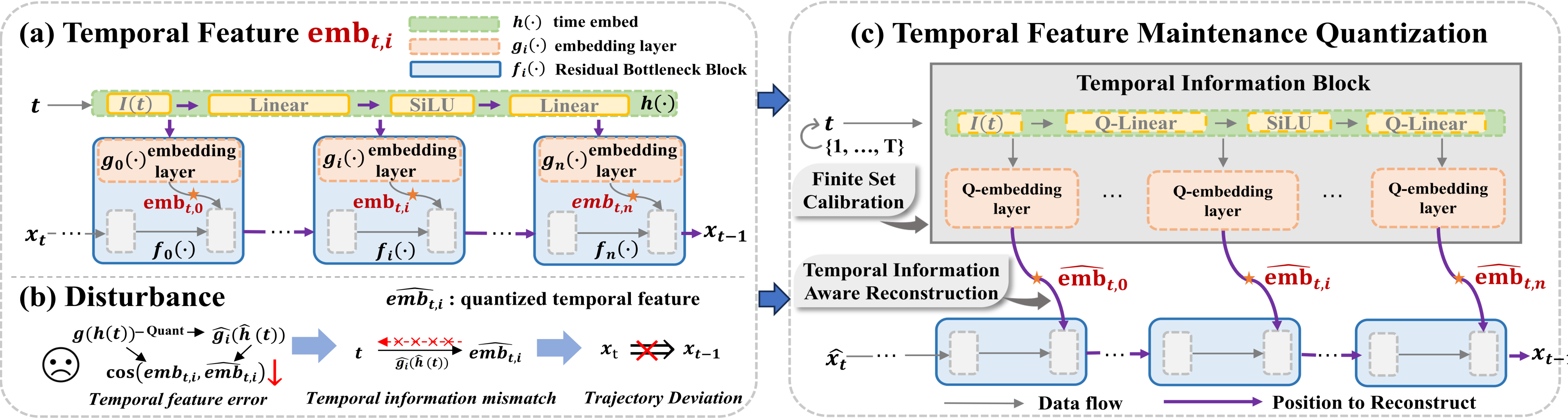
(c) Trajectory deviation. We only quantize embedding layers and time embed in diffusion models.

Inducement analyses & Temporal Feature Maintenance

- Two key inducements:
  - **Inappropriate reconstruction target:** wrong objective, overfit.
  - **Unaware of finite activations:** prev. methods for distribution.
- Solution:
  - **Temporal Information Block:** $\{g_i(h(\cdot))\}_{i=0,\ldots,n}$.
  - **Temporal Information Aware Reconstruction:** $\mathcal{L}_{TIAR} = \sum_{i=0}^{n} \left\| g_i(h(t)) - \hat{g}_i\left(\hat{h}(t)\right)\right\|_F^2$.
  - **Finite Set Calibration:** $\hat{x} = \Phi\left(\left\lfloor \frac{x}{s_t} \right\rceil + z_t, 0, 2^b - 1\right)$.

Methods	Bits (W/A)	FID↓	sFID↓
Full Prec.	32/32	2.98	7.09
Prev	8/8	7.51	12.54
Freeze	8/8	**5.76** (-1.75)	**8.42** (-4.12)
Prev	4/8	9.36	22.73
Freeze	4/8	**7.08** (-2.28)	**16.82** (-5.91)

Freeze means we freeze t-related parts, when quantizing ResBlock and utilize naïve strategy for that component.



Performance & efficiency

- We conduct experiments for various datasets across unconditional/class-conditional/text-guided image generation. **All of experiments** exhibit our method's **SOTA** accuracy.

Methods	Bits (W/A)	LSUN-Bedrooms 256 × 256		LSUN-Churches 256 × 256		CelebA-HQ 256 × 256		FFHQ 256 × 256	
		FID↓	sFID↓	FID↓	sFID↓	FID↓	sFID↓	FID↓	sFID↓
Full Prec.	32/32	2.98	7.09	4.12	10.89	8.74	10.16	9.36	8.67
PTQ4DM*	4/32	4.83	7.94	4.92	13.94	13.67	14.72	11.74	12.18
Q-Diffusion†	4/32	4.20	7.66	4.55	11.90	11.09	12.00	11.60	10.30
PTQD*	4/32	4.42	7.88	4.67	13.68	11.06	12.21	12.01	11.12
TFMQ-DM (Ours)	4/32	**3.60** (-0.60)	**7.61** (-0.05)	**4.07** (-0.48)	**11.41** (-0.49)	**8.74** (-2.32)	**10.18** (-1.82)	**9.89** (-1.71)	**9.06** (-1.24)
PTQ4DM*	8/8	4.75	9.59	4.80	13.48	14.42	15.06	10.73	11.65
Q-Diffusion†	8/8	4.51	8.17	4.41	12.23	12.85	14.16	10.87	10.01
PTQD	8/8	3.75	9.89	4.89*	14.89*	12.76*	13.54*	10.69*	10.97*
TFMQ-DM (Ours)	8/8	**3.14** (-0.61)	**7.26** (-0.91)	**4.01** (-0.40)	**10.98** (-1.25)	**8.71** (-4.05)	**10.20** (-3.34)	**9.46** (-1.23)	**8.73** (-1.28)
PTQ4DM	4/8	20.72	54.30	4.97*	14.87*	17.08*	17.48*	11.83*	12.91*
Q-Diffusion†	4/8	6.40	17.93	4.66	13.94	15.55	16.86	11.45	11.15
PTQD	4/8	5.94	15.16	5.10*	13.23*	15.47*	17.38*	11.42*	11.43*
TFMQ-DM (Ours)	4/8	**3.68** (-2.26)	**7.65** (-7.51)	**4.14** (-0.52)	**11.46** (-1.77)	**8.76** (-6.71)	**10.26** (-6.60)	**9.97** (-1.45)	**9.14** (-2.01)

Unconditional image generation with LDM-4/8. Resolution of images: 256×256

Methods	Bits (W/A)	MS-COCO		
		FID↓	sFID↓	CLIP↑
Full Prec.	32/32	13.15	19.31	0.3146
Q-Diffusion†	4/32	13.58	19.50	0.3143
TFMQ-DM (Ours)	4/32	**13.21** (-0.37)	**19.03** (-0.47)	**0.3144** (+0.0001)
Q-Diffusion†	8/8	13.31	20.54	0.3134
TFMQ-DM (Ours)	8/8	**13.09** (-0.22)	**19.91** (-0.63)	**0.3134** (+0.0000)
Q-Diffusion†	4/8	14.49	20.43	0.3121
TFMQ-DM (Ours)	4/8	**13.36** (-1.13)	**20.14** (-0.29)	**0.3128** (+0.0007)

Text-guided image generation with Stable Diffusion on MS-COCO prompts. Resolution of images: 512×512

Methods	Bits (W/A)	ImageNet 256 × 256		
		IS↑	FID↓	sFID↓
Full Prec.	32/32	235.64	10.91	7.67
PTQ4DM	4/32	-	-	-
Q-Diffusion*	4/32	213.56	11.87	8.76
PTQD†	4/32	201.78	11.65	9.06
TFMQ-DM (Ours)	4/32	**223.81** (+10.25)	**10.50** (-1.15)	**7.98** (-0.78)
PTQ4DM	8/8	161.75	12.59	-
Q-Diffusion*	8/8	187.65	12.80	9.87
PTQD	8/8	153.92	11.94	8.03
TFMQ-DM (Ours)	8/8	**198.86** (+11.21)	**10.79** (-1.15)	**7.65** (-0.38)
PTQ4DM	4/8	-	-	-
Q-Diffusion*	4/8	212.51	10.68	14.85
PTQD	4/8	214.73	10.40	12.63
TFMQ-DM (Ours)	4/8	**221.82** (+7.09)	**10.29** (-0.11)	**7.35** (-5.28)

Class-conditional image generation with LDM-4 on ImageNet. Resolution of images: 256×256

Performance & efficiency

- We present some visualization results.

Random samples from w4a8 quantized and full-precision Stable Diffusion. Resolution of images: 512×512.

- We also deploy our quantized models on CPU.

Methods	Bits (W/A)	UNet Size (Mb)	Latency (s)	Speedup
Full Prec.	32/32	3278.81	81.01	-
OpenVINO	8/8	821.15	33.93	2.39×
TFMQ-DM	8/8	821.77	34.07	2.38×

Efficiency of quantized Stable Diffusion on Intel CPU.

Unconditional image generation with LDM-8 on CelebA-HQ. Resolution of images: 256×256.

Unconditional image generation with LDM-4 on LSUN-Bedrooms. Resolution of images: 256×256.

Summary

- We discover that existing quantization methods suffer from **temporal feature disturbance** affecting the quality of generated images.
- We reveal that the disturbance comes from two aspects: **inappropriate reconstruction target** and **unaware of finite activations**. Both inducements ignore the special characteristics of time information-related modules.
- An advanced framework **(TFMQ-DM)** is proposed, consisting of **TIAR** for weight quantization and **FSC** for activation quantization. Both are based on a **Temporal Information Block** specially devised for diffusion models.
- Extensive experiments on various datasets show that our novel framework achieves a new **SOTA** result in PTQ of diffusion models, especially under 4-bit weight quantization, and significantly accelerates quantization time.