Q-Sched: Pushing the Boundaries of Few-Step Diffusion Models with Quantization-Aware Scheduling

Natalia Frumkin, Diana Marculescu

The University of Texas at Austin, Chandra Family Deptartment of Electrical & Computer Engineering

The University of Texas at Austin

Compressing text-to-image diffusion models

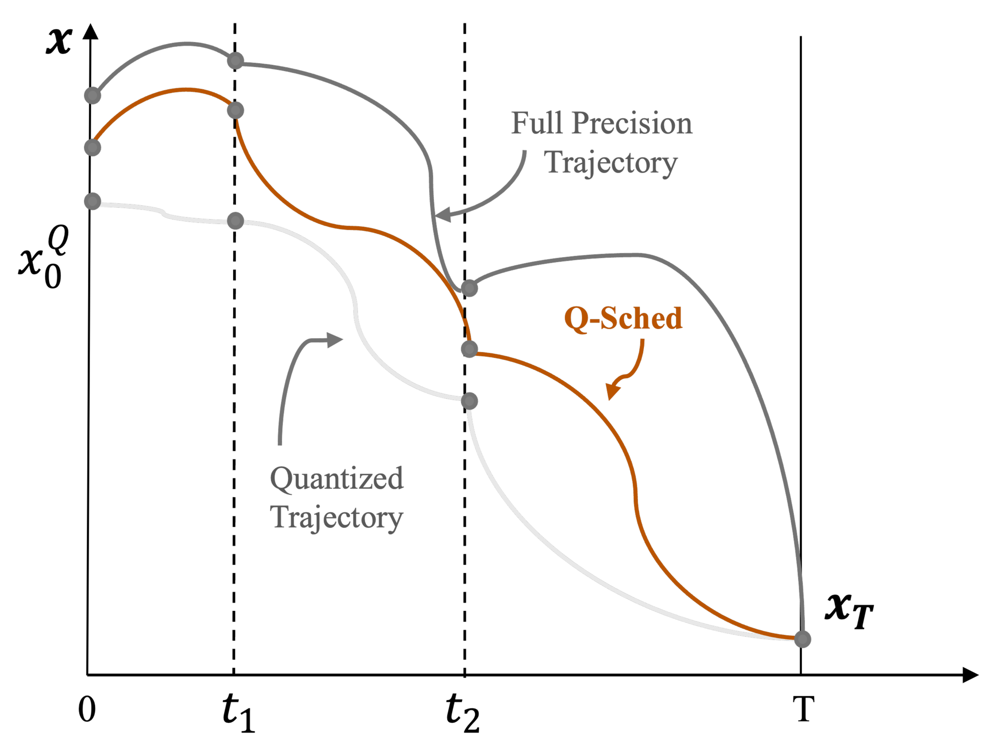
Few-Step Diffusion Distillation

- Reduces multi-step diffusion (e.g. 50–100 steps) to a few-step regime (e.g. 1–4 steps) for fast generation.
- Achieves this via distillation from a teacher model into a lightweight student sampler.
- Now a dominant paradigm for real-time and on-device text-to-image generation.

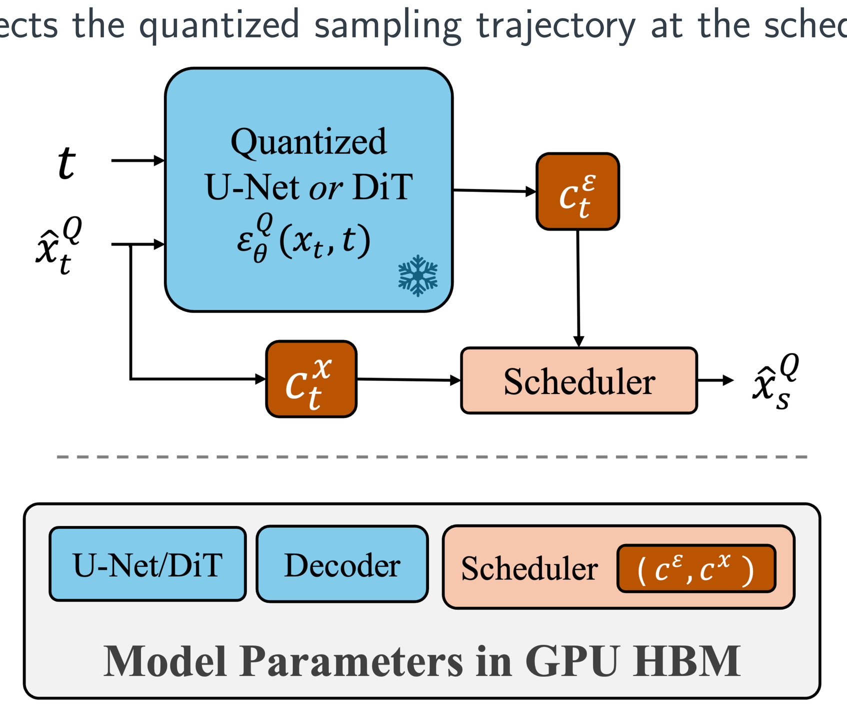
Post-Training Quantization

- Further compresses models by reducing weight/activation precision (e.g. W4A8, W8A8) at test time.
- Commonly combined with few-step distillation to maximize efficiency and deployability.
- However, introduces significant quality degradation due to accumulated quantization error.

Correcting the quantized sampling trajectory



Q-Sched corrects the quantized sampling trajectory at the scheduler level.



Per-timestep coefficients (c_t^ϵ, c_t^x) live in the scheduler; model weights stay frozen.

Q-Sched: A scheduler-level adaptation

- Introduces lightweight, per-timestep coefficients that correct quantization-induced trajectory drift.
- Operates directly at the sampler level with zero additional model overhead or latency.
- Enables quantized models to match or exceed full-precision quality in few-step regimes.

Q-Sched coefficient in the TCD scheduler:

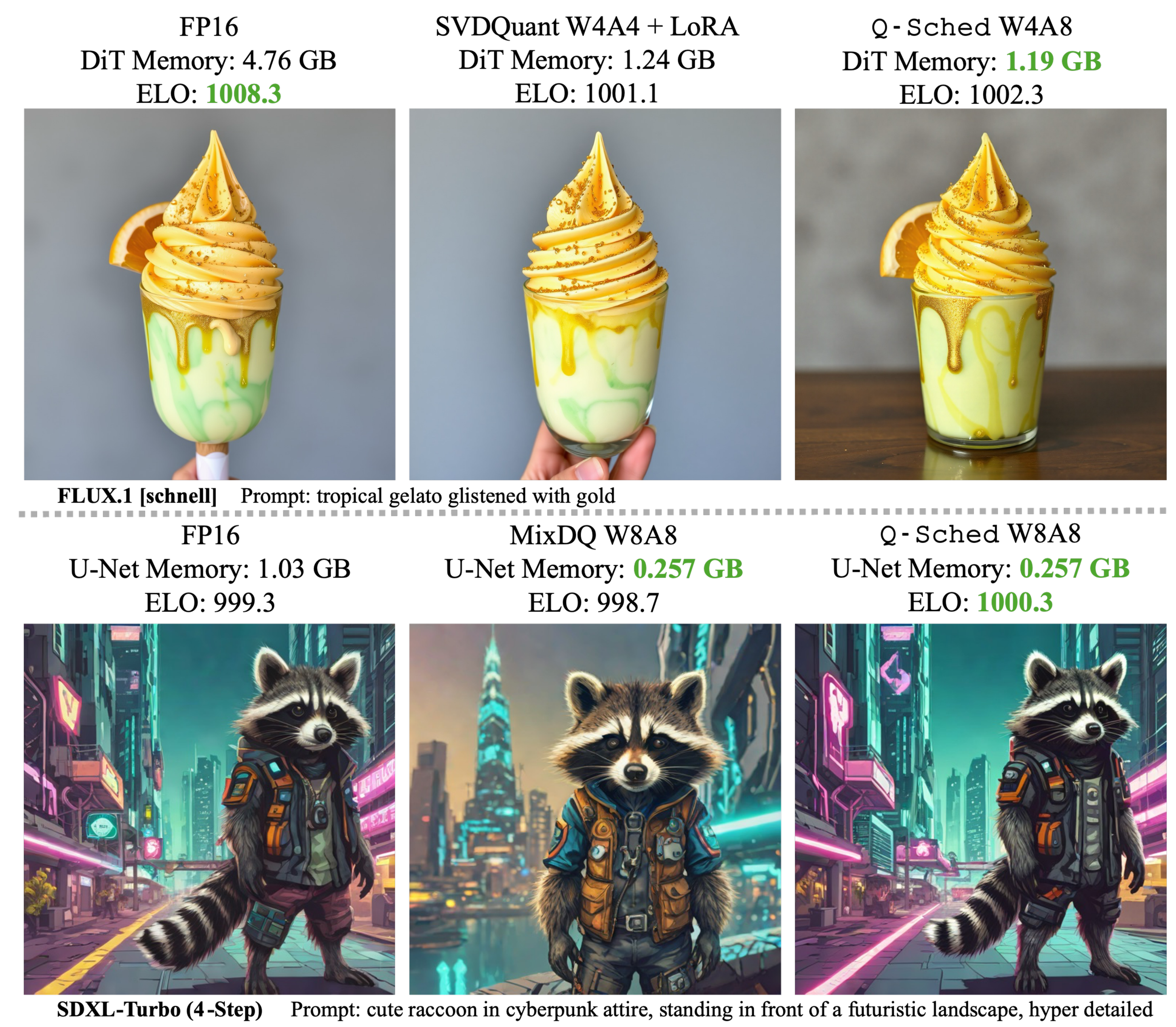
$$\mathbf{x}_s = \frac{\alpha_s}{\alpha_{s'}}\left(\frac{c_t^x \mathbf{x}_t - \sigma_t c_t^\epsilon \mathcal{E}_\theta^Q(x_t, t)}{\alpha_t} + \sigma_{s'} c_t^\epsilon \mathcal{E}_\theta^Q(x_t, t)\right) + \sqrt{1 - \frac{\alpha_s^2}{\alpha_{s'}^2}}\,\mathbf{z}$$

The Joint Alignment Quality (JAQ) loss

$$\mathrm{JAQ}(x) = \mathrm{TC}(x) + k \cdot \mathrm{IQ}(x)$$

- **Text Consistency (TC):** encourages alignment between generated images and input prompts.
- **Image Quality (IQ):** promotes perceptual fidelity and realism using reference-free quality metrics.

Qualitative results



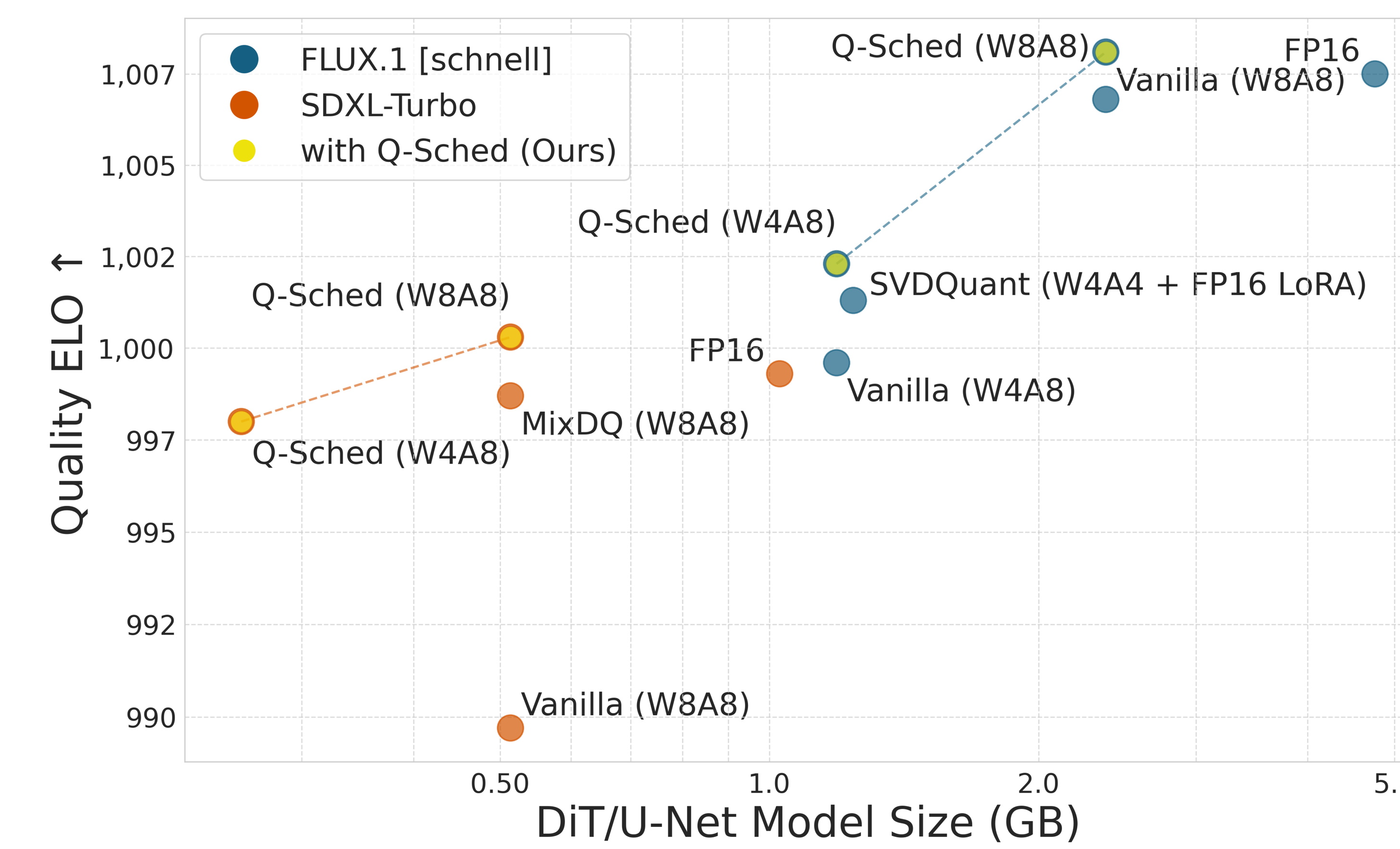
Q-Sched achieves competitive image quality on large-scale few-step diffusion models.

Quantitative results

Q-Sched outperforms MixDQ [1] and SVDQuant [2]

	SDXL-Turbo (4-Step)				FLUX.1 [schnell]			
	FP16	W8A8	MixDQ	**Q-Sched**	FP16	W4A8	SVDQ	**Q-Sched**
CLIP Score ↑	**25.62**	**25.62**	25.38	25.36	**25.61**	25.17	25.52	25.27
CLIP IQA ↑	0.725	0.727	0.727	**0.731**	**0.716**	0.712	0.714	0.707
HPV2 ↑	0.276	0.276	0.275	**0.278**	0.275	0.274	**0.275**	0.272
AQ-MAP ↑	0.693	0.694	0.693	**0.696**	**0.700**	**0.700**	0.697	**0.700**
Pick Score ↑	18.48	18.49	18.48	**18.51**	18.43	18.42	18.40	**18.46**
MANIQA ↑	0.508	0.513	0.502	**0.511**	**0.528**	0.500	0.514	0.506
JAQ (ours) ↑	1.663	1.665	1.659	**1.669**	**1.676**	1.675	1.669	1.673

Q-Sched achieves pareto-optimality on large-scale human evaluation.



Quality ELO vs. model size from large-scale human evaluation (Rapidata).

In Summary

Q-Sched is a scheduler-level PTQ approach that adapts the diffusion sampler while keeping quantized weights fixed.

References

[1] Zhang, T., et al. "MixDQ: Mixed-Precision Quantization for Diffusion Models." *ECCV* (2024).

[2] Li, M., et al. "SVDQuant: Absorbing Outliers via Low-Rank Components for 4-bit Diffusion Models." *arXiv:2411.05007* (2024).

Q-Sched on GitHub

This work was supported in part by NSF CCF Grant No. 2107085, iMAGiNE — the Intelligent Machine Engineering Consortium at UT Austin, and UT Cockrell School of Engineering Doctoral Fellowships. Human evaluation studies were conducted using Rapidata.