QLIP Text Embedding Knows How to Quantize Text-Guided Diffusion Models

Hongjae Lee, Myungjun Son, Dongjea Kang, and Seung-Won Jung

KOREA UNIVERSITY

ICCV OCT 19-23, 2025 HONOLULU HAWAII

Video Github

Why do we need QLIP?

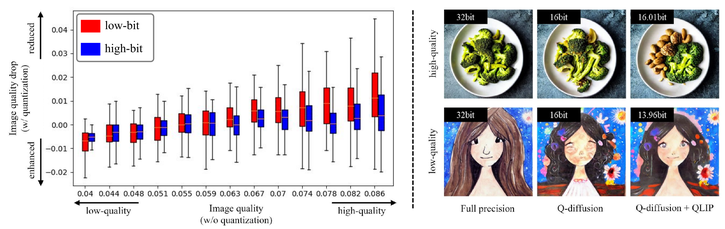
- Diffusion models generate high-quality images but suffer from huge computational cost.
- Existing quantization methods ignore text prompts, which are critical for guiding generation.
- Neglecting text prompts results in inefficient bit allocation across layers and timesteps.

We propose **QLIP**,
the **first text-aware quantization method for diffusion models**.

Motivation

The Impact of Quantization on Generation Quality

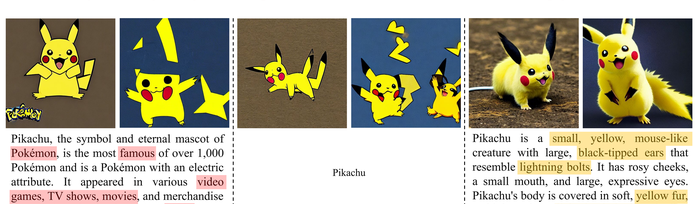
Q1. Should all images use the same quantization bit precision?



- When the full-precision model would produce high-quality images, low-bit quantization causes a severe drop; when quality is low, the drop is minor.
- This shows that the impact of quantization strongly depends on the expected image quality.

The Impact of Prompt Richness on Generation Quality

Q2. Can we predict quality before generation and allocate bits accordingly?



Pikachu, the symbol and eternal mascot of Pokémon, is the most famous of over 1,000 Pokémon and is a Pokémon with an electric attribute. It appeared in various video games, TV shows, movies, and merchandise since its introduction in the late 1990s.

Pikachu

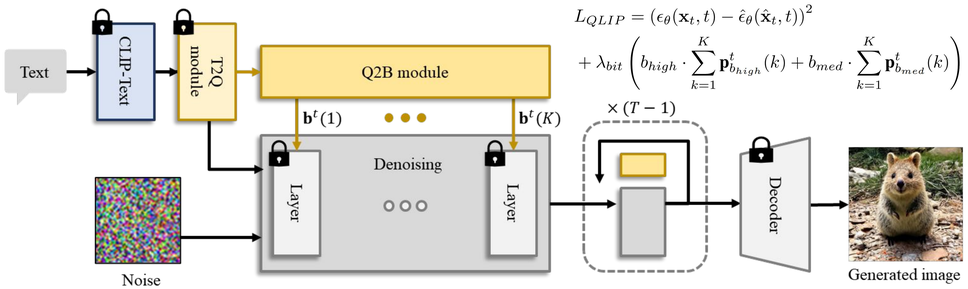
Pikachu is a small, yellow, mouse-like creature with large, black-tipped ears that resemble lightning bolts. It has rosy cheeks, a small mouth, and large, expressive eyes. Pikachu's body is covered in soft, yellow fur, and it has a long, lightning bolt-shaped tail.

- Image generation quality varies with the richness of the text prompt.
- The more richly the prompt describes shapes, textures, and attributes, the higher the quality of the generated image.
- Simply making prompts longer does not guarantee better results

QLIP: Quantization of Language-to-Image diffusion models using text Prompts

Overview of QLIP Framework.

A1. Bit precision should be dynamically allocated across layers and timesteps based on image quality.

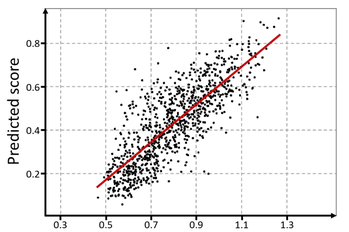


$$L_{QLIP} = (\epsilon_\theta(\mathbf{x}_t, t) - \hat{\epsilon}_\theta(\hat{\mathbf{x}}_t, t))^2 + \lambda_{bit}\left(b_{high}\cdot\sum_{k=1}^{K}\mathbf{p}^t_{b_{high}}(k) + b_{med}\cdot\sum_{k=1}^{K}\mathbf{p}^t_{b_{med}}(k)\right)$$

- The T2Q module predicts generation quality from text embeddings.
- The Q2B module allocates layer- and timestep-wise bit precision based on the predicted quality.
- Together, they enable a text-aware quantized diffusion model that balances efficiency and image quality.

T2Q Module Results

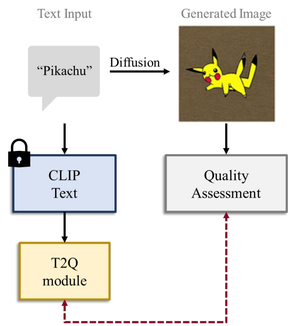
A2. Generation quality can be predicted from the text prompt.



GIQA: Generated image quality assessment. ECCV, 2020

- The T2Q module is trained to predict image quality from CLIP text embeddings.
- Ground-truth quality scores are obtained using GIQA on images generated by a full-precision diffusion model.

T2Q Module Training Process



$$L_{t2q} = \frac{1}{N_{T2Q}}\sum_{i=1}^{N_{T2Q}}\left(\bar{q}^i - \phi\left(\mathbf{z}^i\right)\right)^2$$

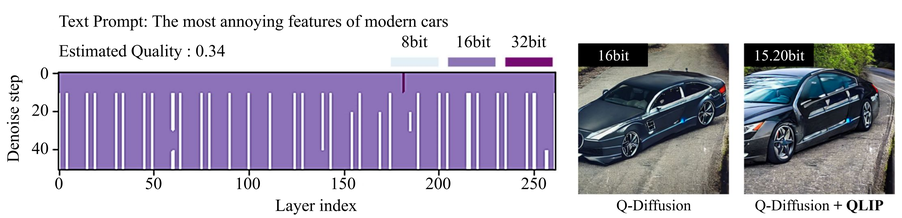
Results

QLIP Bit Allocation by Prompt Richness

Richness controlled captions	FAB
retro camera	13.98
retro camera has a square frame with metal body and leather grip.	14.10
retro camera has a boxy metal body, leather accents, and a large lens.	14.40
retro camera has a metal frame, leather casing, and a large, rounded glass lens with delicate dials.	15.15
retro camera has a classic metal body, leather wrapping, and a large glass lens, complete with dials and knobs for a timeless, vintage feel.	15.19

- The richer and more detailed the prompt, the higher bits QLIP allocates to generate vivid and detailed images.
- For simple prompts, the image is relatively easy to generate, and QLIP accordingly assigns lower bits.

Layer- and Timestep-wise Bit Allocation



- QLIP dynamically assigns bit precision across timesteps and layers, adapting to the predicted generation quality.

Quantitative Results

Method	Bitoptions	FAB↓	BitOPs (T)↓	FID↓	sFID↓	CLIP Score↑
-	W32A32	32.00	10.46	20.00	47.80	0.2983
Q-diffusion	W4A16	16.00	1.03	**24.68**	62.70	**0.2965**
+QLIP	W4A{8,16,32}	**10.58**	**0.82**	24.72	**59.54**	0.2964
PTQD	W4A16	16.00	1.03	**24.66**	62.60	0.2965
+QLIP	W4A{8,16,32}	**10.58**	**0.82**	24.71	**59.25**	**0.2966**

- QLIP reduces computational complexity while preserving image quality.
- Table reports results on a lightweight diffusion model (BK-SDM), QLIP is also effective on larger models (SDXL, FLUX).