

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

## Supplementary Material

In this supplementary document, we present additional results and analyses, including the following:

- Results on a large-scale diffusion model (Section A).
- Evaluation on mixed-precision quantization method (Section B).
- Ablation studies on bit selection criteria (Section C).
- Ablation studies on bit precision options (Section D).
- Batch inference scenario analysis (Section E).
- Memory requirement analysis (Section F).
- Details on the structure and implementation of the T2Q module (Section G).
- Implementation details for ablation studies (Section H).
- Quantization sensitivity analysis for each layer (Section I).
- Additional visual comparison results.(Section J).

### A. Results on a Large-Scale Diffusion Model

We conducted experiments using SDXL [6] with the Euler scheduler as the base model to evaluate the effectiveness of QLIP on a large-scale diffusion model. As shown in Table S1, QLIP demonstrated significant improvements in computational efficiency while maintaining competitive image quality.

For the COCO2017 dataset, QLIP effectively reduced the computational complexity of the baseline diffusion quantization methods. Specifically, Q-diffusion+QLIP (W4A{8,16,32}) demonstrated a significant reduction in FAB while maintaining comparable FID and sFID scores to Q-diffusion (W4A16). This result indicates that QLIP optimizes the bit precision selection effectively, reducing computational overhead without compromising image quality. The results on the Conceptual Captions dataset exhibited similar trends to those observed with COCO2017. These results suggest that QLIP generalizes well to large-scale diffusion models.

### B. Evaluation on Mixed-Precision Method

We compared with PCR [7], a recent mixed-precision quantization method, using FLUX [1] as the base model. For image quality, we reported ImageReward [8] and PickScore [4], using 500 prompts from COCO2017. As shown in Table S2, applying our QLIP to PCR further improved performance by achieving higher quality scores while also reducing overall bit usage, demonstrating its effectiveness even on the recent diffusion model FLUX.

COCO2017				
Method	FAB $\downarrow$	FID $\downarrow$	sFID $\downarrow$	CLIP Score $\uparrow$
SDXL [6]	32.00	23.75	65.85	0.3180
Q-diffusion [5]	16.00	28.46	67.54	0.3178
+QLIP	12.68	28.16	66.31	0.3177

Conceptual Captions				
Method	FAB $\downarrow$	FID $\downarrow$	sFID $\downarrow$	CLIP Score $\uparrow$
SDXL [6]	32.00	19.25	47.72	0.3085
Q-diffusion [5]	16.00	22.23	49.57	0.3075
+QLIP	11.34	21.95	48.61	0.3074

Table S1. Quantitative comparisons at a resolution of  $768\times 768$  using SDXL [6]. For the bit precision options, W4A16 was used for Q-diffusion and W4A{8,16,32} were used for QLIP.

Method	FAB $\downarrow$	Image Reward $\uparrow$	Pick Score $\uparrow$
FLUX [1]	16.00	1.1013	23.07
PCR [7]	9.60	0.9986	22.97
+QLIP	7.92	1.0214	23.01

Table S2. Quantitative comparisons with the mixed-precision quantization method PCR, using FLUX as the baseline model. Evaluation was conducted on 500 prompts from the COCO2017 dataset at a resolution of  $1024\times 1024$ . For the bit precision options, W4A{8,16} was used for PCR, and W4A{6,8,16} were used for QLIP.

Bit Selection Strategy	FAB $\downarrow$	FID $\downarrow$
Image Complexity	14.86	31.09
Prompt Length	12.79	31.49
Image Quality (QLIP)	12.14	30.01

Table S3. Comparison of bit selection criteria.

### C. Ablation Studies on Bit Selection Criteria

We conducted experiments to explore alternative criteria for determining bit precision, replacing the predicted image quality used in our T2Q module. Specifically, we investigated two alternative metrics: image complexity and prompt length, as shown in Table S3.

Image complexity has been effectively used in other tasks, such as super-resolution, as a criterion for dynamic

Bit-Options	FAB $\downarrow$	FID $\downarrow$
{8,16}	10.51	24.78
{8,16,32}	10.58	24.72
{6,8,16,32}	9.24	25.22

Table S4. Ablation study on bit precision options for QLIP.

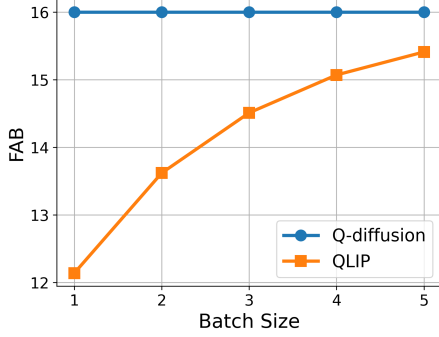


Figure S1. Bit allocation over different batch sizes.

Layer No.	Operator	Kernel ( $C_{in} \times C_{out}$ )
1	Linear / ReLU	$C_{clip} \times C_{clip}$
2	Linear / ReLU	$C_{clip} \times 512$
3	Linear / ReLU	$512 \times 1$
4	Sigmoid	-

Table S5. The structure of the T2Q module.

quantization [3]. To test its applicability in diffusion models, we replaced the T2Q and Q2B modules with T2C and C2B modules that utilize image complexity, measured as average image gradient magnitude. However, this configuration resulted in worse performance (FAB 14.86, FID 31.09) compared to the original QLIP design (FAB 12.14, FID 30.01), suggesting that image complexity alone is not a reliable criterion for bit selection in diffusion models.

We also evaluated a prompt length-based bit allocation strategy, where the number of tokens in the input prompt was used to decide bit precision. This variant also underperformed (FAB 12.79, FID 31.49) relative to our original approach. Unlike prompt length, which only reflects the input length, the T2Q module captures richer semantic representations from text, leading to more accurate bit assignment and better image quality.

## D. Ablation Studies on Bit-Options

The impact of different bit-options on FAB and FID is presented in Table S4. The results demonstrate that while the {6,8,16,32} configuration achieves the lowest FAB, it

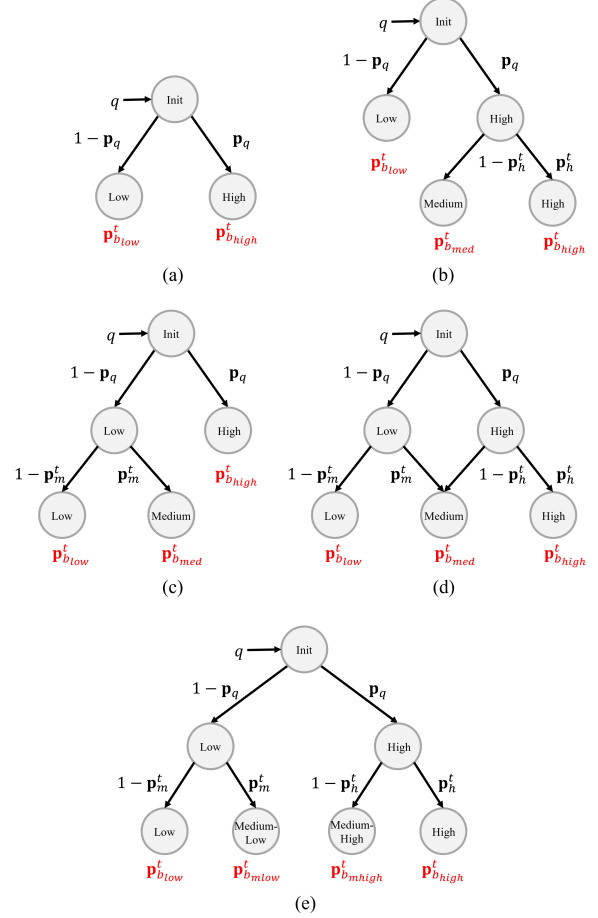


Figure S2. Implementation details of ablation study variants for the Q2B module. (a) Bit-Options correspond to the case with 2 bit candidates using  $p_q$ . (b), (c), and (d) correspond to the case with 3 bit candidates using  $p_q + p_h$ ,  $p_q + p_m$ , and  $p_q + p_m + p_h$ , respectively. (e) Bit-Options correspond to the case with 4 bit candidates.

also leads to an increase in FID. Conversely, the {8,16,32} configuration results in the lowest FID while maintaining a relatively low FAB, making it a better choice for minimizing image degradation and effectively reducing computational overhead. Based on these findings, we select {8,16,32} as the default bit-option.

## E. Batch Inference Scenario

Our practical implementation strategy for batch processing is to assign the maximum required bit precision for each layer across the batch. As shown in Fig. S1, the effectiveness of QLIP decreases as the batch size increases, which is a known limitation of input-adaptive quantization methods. However, text-to-image generation is typically performed on individual prompts or small batches. For exam-

ple, the recent GPT Image 1 API [2] provided by OpenAI officially supports small batch sizes per API call, and processing many images requires issuing multiple parallel API calls, rather than increasing the batch size within a single request. Therefore, this limitation has less significant practical impact.

## F. Memory Requirements

QLIP requires additional memory to store the model, including 1.5MB for the T2Q module and 12.1KB for the Q2B module, and additional 2.2KB for scale and zero-point values. However, this additional memory overhead is negligible compared to the total memory of BK-SDM-Tiny-2M 4bit (154.2MB). While our method does not reduce memory usage, it improves both energy efficiency and inference time by reducing the computational cost.

## G. Structure of the T2Q module

As shown in Table S5, the T2Q module consists of a simple 3-layer MLP with a ReLU activation function. The first linear layer is the projection layer of the CLIP text encoder, which is frozen during training. In the final layer, a sigmoid function is used to limit the output range of the quality  $q$ .

## H. Implementation Details on Ablation Studies

The implementation of the Q2B modules used in ablation studies is shown in Figure S2. When only  $\mathbf{p}_q$  is used,  $\mathbf{p}_{med}^t$  is not included, limiting the bit-options to two candidates. When  $\mathbf{p}_l^t$  (or  $\mathbf{p}_h^t$ ) is not used,  $\mathbf{p}_{low}^t$  (or  $\mathbf{p}_{high}^t$ ) is determined using  $\mathbf{p}_q$ , adjusting the available three-bit candidates accordingly. Incorporating additional bit-options, such as  $\mathbf{p}_{mlow}^t$  and  $\mathbf{p}_{mhigh}^t$ , can further increase the number of bit candidates.

## I. Analysis on Quantization Sensitivity

Figure S3 shows the proportion of bits used in each layer of the quantized denoising model. Stable Diffusion utilizes cross-attention blocks to incorporate text prompts into the image latent space, in addition to basic residual blocks. Notably, our proposed QLIP aims to manage the overall bit usage, particularly by assigning low bits more frequently on the layers in cross-attention blocks. Specifically, it is evident that bit reduction occurs mainly within three layers: (1) “proj\_in” which leverages the denoised image latent passed into the cross-attention block, (2) “at2.to.v” which projects text prompt to denoised image latent, (3) “proj\_out” which passes out the result of cross-attention to the next denoising blocks. These results are attributed to many cases with a weak correlation between the generated image and the text prompt, promoting the T2Q and Q2B modules to determine low bits for efficient quantization.

## J. Additional Qualitative Results

Figure S4 presents additional examples illustrating how changes in FAB are influenced by the richness and specificity of the text descriptions, along with the generated images. As shown, QLIP assigned higher bits when the text provides more specific and detailed descriptions, demonstrating that it leverages textual information to adapt bit precision, enabling effective synthesis of vivid details and textures.

Figures S5 and S6 show additional examples of the bit selection results and generated images using Q-diffusion or PTQD as baseline quantization methods. Figure S7 provides examples of images generated by QLIP using Q-diffusion or PTQD as baseline quantization methods, along with the full-precision model. The text prompts for generating images were sourced from the captions of COCO2017 and Conceptual Captions datasets.

## References

- [1] FLUX: [huggingface.co/black-forest-labs/flux.1-dev](https://huggingface.co/black-forest-labs/flux.1-dev). 1
- [2] GPT Image 1 API: <https://platform.openai.com/docs/models/gpt-image-1>. 3
- [3] Cheeun Hong, Sungyong Baik, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. CADyQ: Content-aware dynamic quantization for image super-resolution. In *Proceedings of the European Conference on Computer Vision*, pages 367–383, 2022. 2
- [4] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-Pic: An open dataset of user preferences for text-to-image generation. In *Proceedings of the Advances in Neural Information Processing Systems Workshop*, 2023. 1
- [5] Xiuyu Li, Yijiang Liu, Long Lian, Huanrui Yang, Zhen Dong, Daniel Kang, Shanghang Zhang, and Kurt Keutzer. Q-diffusion: Quantizing diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17535–17545, 2023. 1
- [6] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *Proceedings of the International Conference on Learning Representations*, 2024. 1
- [7] Siao Tang, Xin Wang, Hong Chen, Chaoyu Guan, Zewen Wu, Yansong Tang, and Wenwu Zhu. Post-training quantization with progressive calibration and activation relaxing for text-to-image diffusion models. In *Proceedings of the European Conference on Computer Vision*, pages 404–420, 2024. 1
- [8] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. ImageReward: Learning and evaluating human preferences for text-to-image generation. In *Proceedings of the Advances in Neural Information Processing Systems Workshop*, 2023. 1

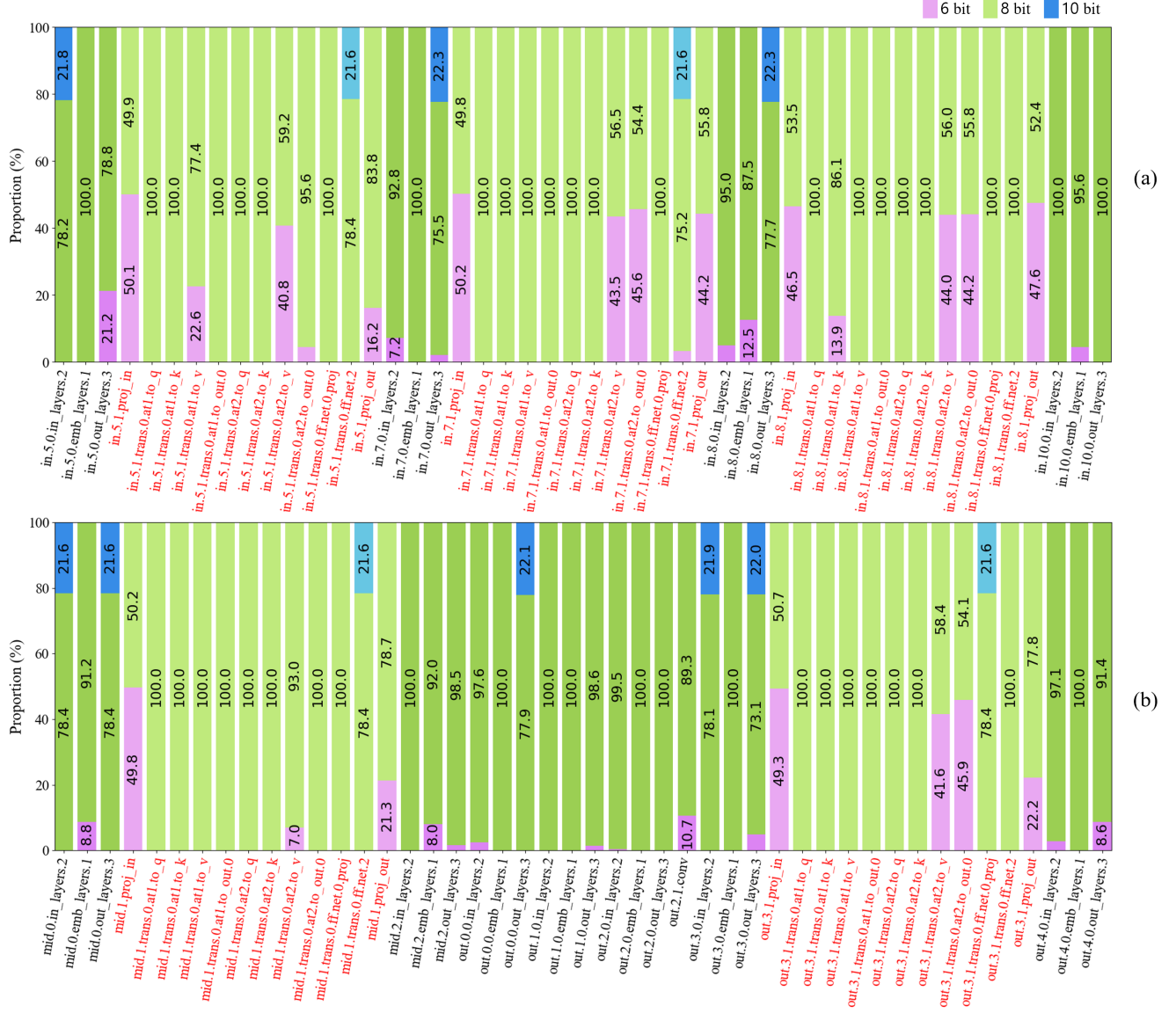


Figure S3. The proportion of bits assigned in each layer of the Stable Diffusion model quantized by W4A{6,8,10} using Q-diffusion w/ QLIP while generating 10k images using the COCO2017 validation dataset. On the x-axis, cross-attention blocks and residual blocks are indicated in red and black, respectively. (a) and (b) show the statistics of several layers in the input, middle and output blocks.









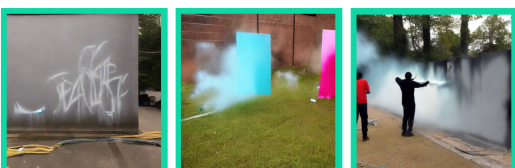
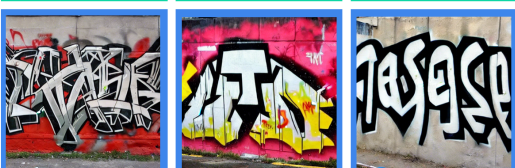
Images			Richness controlled captions	FAB
			volcano	14.30
			volcano erupted with a loud roar.	14.40
			volcano erupted with a loud roar, spewing lava.	14.48
			volcano erupted with a loud roar, spewing lava and ash.	14.51
			volcano erupted with a loud roar, spewing lava and ash into the sky.	14.97
			rainbow	14.54
			rainbow, a stunning natural phenomenon.	14.81
			rainbow, a stunning natural phenomenon, forms a bright, multicolored arc in the sky.	14.99
			rainbow, a stunning natural phenomenon, forms a bright, multicolored arc in the sky, displaying colors.	15.00
			rainbow, a stunning natural phenomenon, forms a bright, multicolored arc in the sky, displaying colors in the order of red, orange, yellow, green, blue, indigo, and violet.	15.12
			snowflake	14.04
			snowflake is a tiny intricate crystal of ice.	14.06
			snowflake is a tiny intricate crystal of ice with patterns.	14.10
			snowflake is a tiny intricate crystal of ice often shaped like a star with patterns.	14.28
			snowflake is a tiny intricate crystal of ice often shaped like a star with symmetrical patterns.	14.78
			spray painting	13.64
			spray painting creates bold, colorful images on walls.	13.83
			spray painting adds layers of bright, swirling graffiti on urban walls.	13.85
			spray painting can be made more descriptive by adding the following words: graffiti or street art	14.28
			spray painting can be made more descriptive by adding the words graffiti or street art to emphasize its appearance.	14.24

Figure S4. Examples of variations in FAB by QLIP for the texts with different levels of richness and detail, along with the generated images.

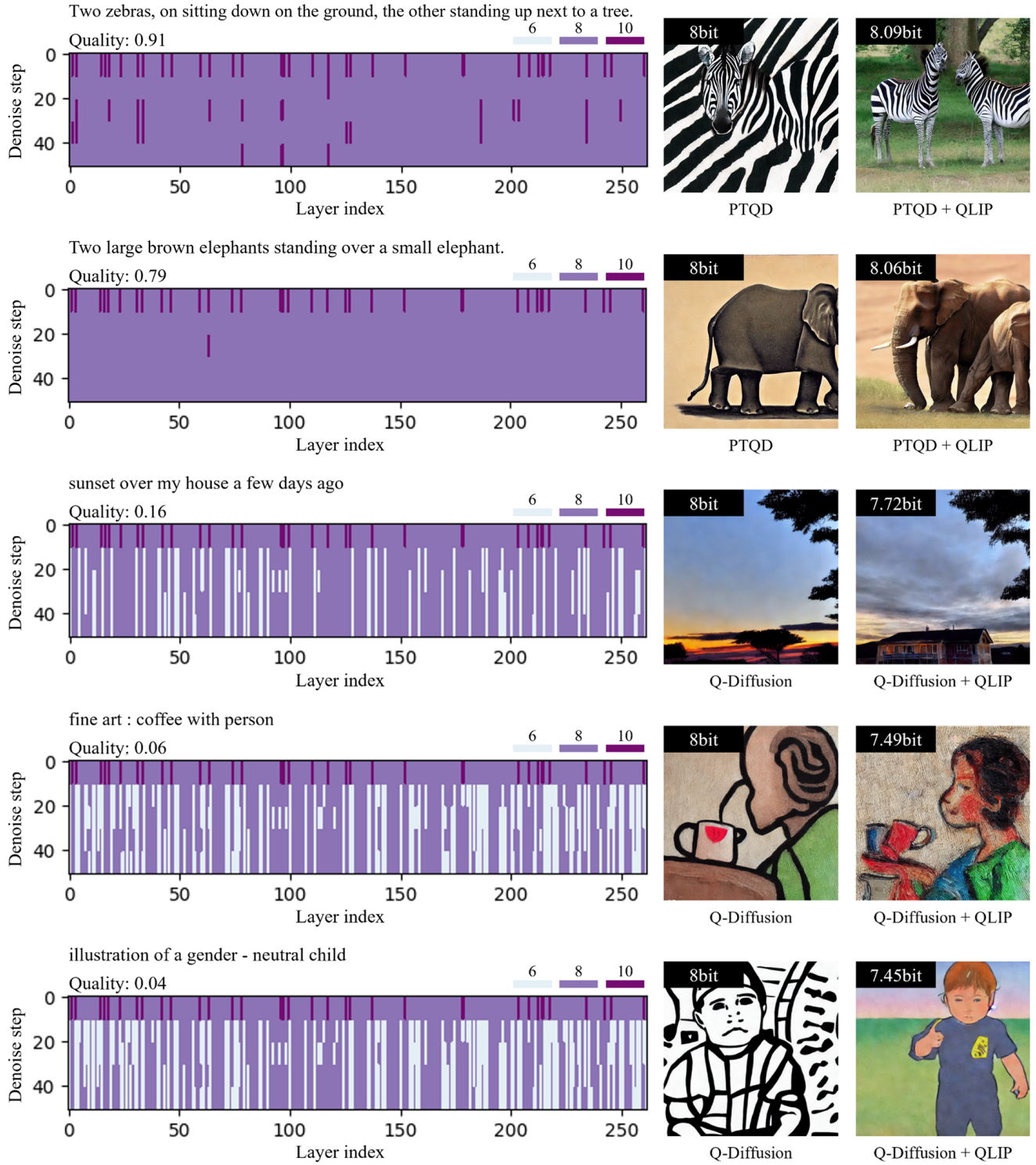


Figure S5. Examples of the bit selection results and generated images using Q-diffusion or PTQD as baseline quantization methods. QLIP is applied with the bit precisions of  $W4A\{6,8,10\}$ .

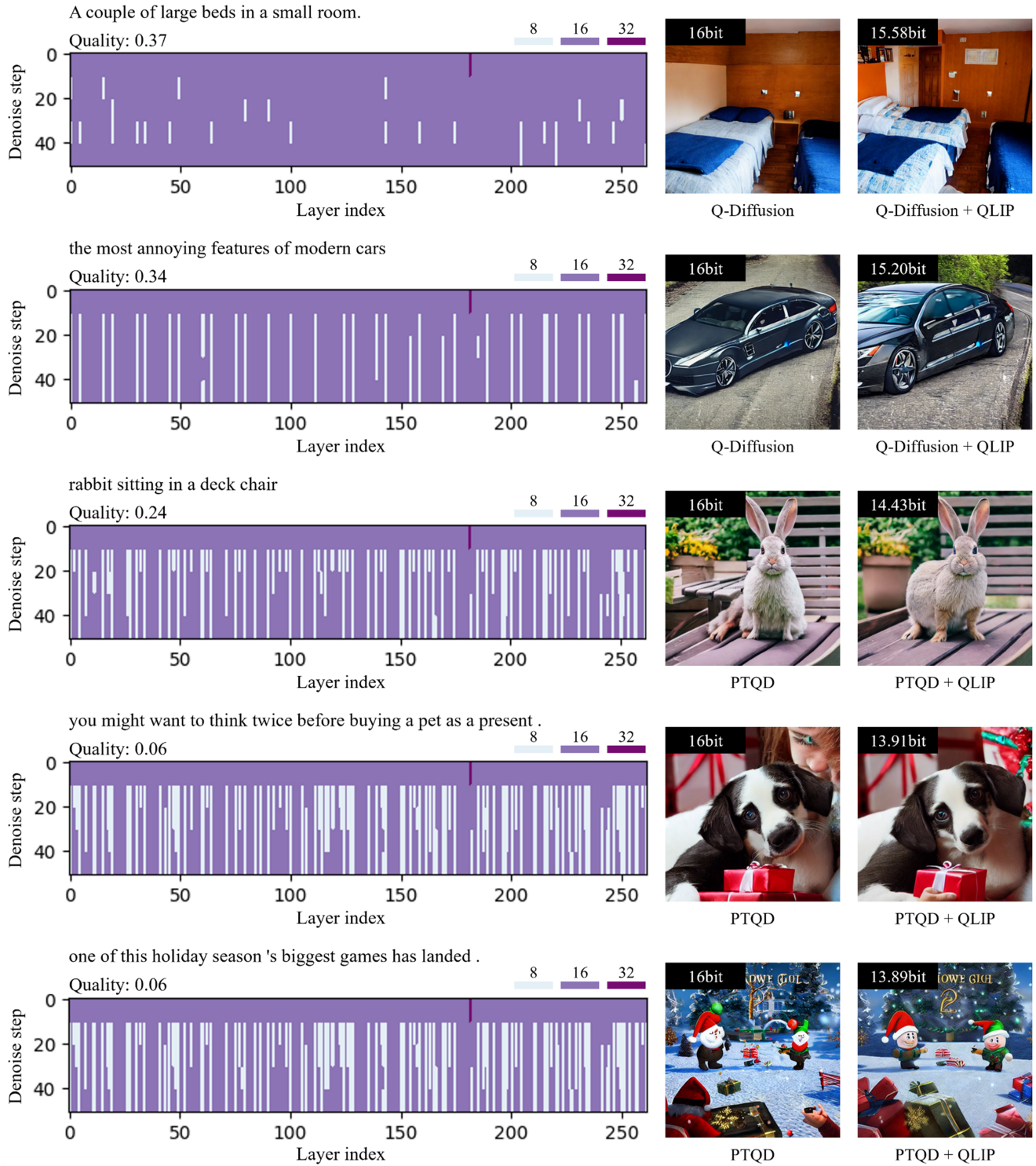


Figure S6. Examples of the bit selection results and generated images using Q-diffusion or PTQD as baseline quantization methods. QLIP is applied with the bit precisions of W4A{8,16,32}.



A large long train on a steel track.



A very tall church with a clock below a tower.



rainbow after a light shower .



a ferris wheel and wires in a cloudy sky



movement of water in the stones



Full precision

Q-diffusion

Q-diffusion + QLIP

PTQD

PTQD + QLIP

Figure S7. Examples of generated images using QLIP with Q-diffusion or PTQD as baseline quantization method.