A Experimental Details

A.1 Details of Diffusion Models

In order to conduct a comprehensive analysis of the early-stage robustness in diffusion models, we employ five different diffusion models to generate various datasets, including CIFAR10 [24], FFHQ [27], CelebA-HQ [26], LSUN-bedrooms, and LSUN-churches [25]. To meausre the Fréchet Inception Distance (FID), we utilize the torch-fidelity library [1], following the methodology established in previous works [6, 8]. Subsequently, we apply the proposed RAQ method to perform high-resolution image generation tasks, encompassing both unconditional image generation for LSUN-bedrooms/LSUN-churches and conditional image generation using Stable Diffusion [8]. The details of the diffusion models utilized in these experiments are thoroughly presented in Table 2. Note that the calculation of σ_t in Eq. 2 is performed using η in Table 2 as specified in the following equation [6]:

$$\sigma_t = \eta \cdot \sqrt{(1-\alpha_{t-1})/(1-\alpha_t)}\sqrt{1-\alpha_t/\alpha_{t-1}} \tag{5}$$

Table 2: Implementation specifications of the diffusion models

	CIFAR-10	FFHQ/ CelebA-HQ	LSUN-Bedrooms	LSUN-Churches	Unconditional Generation
Image Size	32×32	256×256	256×256	256×256	512×512
Architecture	DDIM [2]	LDM-4 [3]	LDM-4 [3]	LDM-8 [3]	Stable Diffusion v1.4 [4]
Sampler	DDIM [6]	DDIM	DDIM	DDIM	PLMS [28]
Step Count	100	200	200	200	50
η	0	0/1	1	0	0

A.2 Details of Entropy Analysis

In Section 3.1, we calculate the entropy of the latent variables x_t for each diffusion step. To facilitate this calculation, we transform the values of x_t into histogram bins. Specifically, we map x_t to a histogram bin using the following equation:

$$h(x_t) = clamp(\lfloor \frac{x_t}{256} \rceil,\ -3,\ +3) \tag{6}$$

This equation ensures that the values of the histogram bins are constrained within the range of -3 to $+3$, allowing us to effectively create a histogram with 256 bins. Once we have the histogram representation of x_t, we can calculate the entropy using the equation[5]:

$$H(X) = \sum_x -p(X)\log p(X) \tag{7}$$

A.3 Intermediate Image Prediction during Reverse Diffusion Process

In Fig. 1(a) and Fig. 2, we showcase the intermediate image prediction results during the reverse diffusion process, aiming to illustrate the characteristics of diffusion models. To visualize the prediction results, we utilize the following x_0 prediction of each diffusion step as stated in [6]:

$$x_0 = \frac{x_t - \sqrt{1-\alpha_t}\epsilon_\theta(x_t, t)}{\sqrt{\alpha_t}} \tag{8}$$

[1] https://github.com/toshas/torch-fidelity

[2] https://github.com/ermongroup/ddim

[3] https://github.com/CompVis/latent-diffusion

[4] https://github.com/CompVis/stable-diffusion

[5] Claude Elwood Shannon, "A Mathematical Theory of Communication", Bell system technical journal, 1948

A.4 Details of Activation Quantization

During the activation quantization process, we observed that the skip connections of ResBlocks, the first convolutional layer responsible for transforming the latent variable into the input of the denoising network, and the last convolutional layer responsible for transforming the output of the denoising network (Fig. 1(b)) had a significant impact on the quality of the final image. However, these components constitute a negligible fraction of the overall computation. Therefore, in order to balance computational efficiency and image quality, the proposed RAQ adjust the activation bits of the diffusion models to the desired bit precision while keeping the activation bits of these three components fixed at 8 bits.

B Additional Results

B.1 Activation Quantization across Diffusion Steps

Fig. 9 complements the image generation results presented in Fig. 6. Fig. 9 showcases image generation with 4-bit activation quantization at different diffusion steps, alongside the image generation with floating-point activations. The results of the activation quantization demonstrate a consistent trend with the noise injection test (4). When 4-bit activation quantization is applied to the early stages, the resulting images closely resemble those generated using floating-point activations, showcasing high quality with minor shape variations. However, applying 4-bit activation quantization to the entire diffusion process leads to a significant compromise in the generated image quality. This is primarily due to the degradation in image quality caused by the quantization applied to the later diffusion steps.

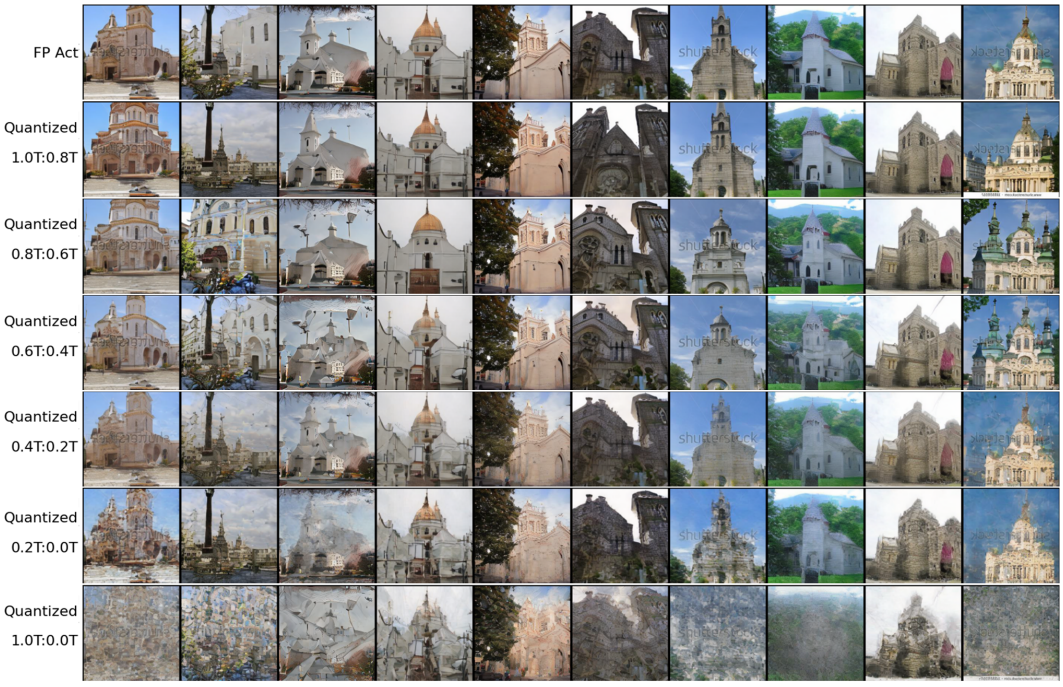
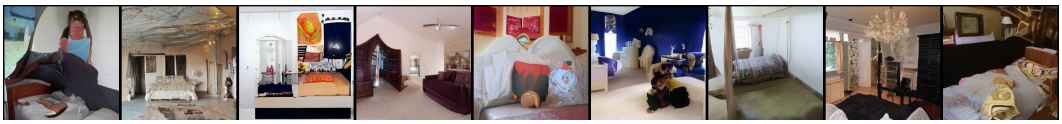


Figure 8: Examples of 256×256 LSUN-Churches generation with FP32 activation or activation quantization across different diffusion steps. Example images with activation quantization are generated by applying 4-bit activation quantization in the target diffusion steps.

B.2 Explanation on FID Improvement with Proposed RAQ on LSUN-Bedrooms

In Table 1 of Section 5.1, we observe that the LSUN-Bedrooms images generated with the proposed RAQ exhibit slightly better FID scores compared to Q-diffusion with W4A32 and W4A8. To investigate the reason behind this FID improvement, we conduct a detailed comparison of the images generated using full-precision activations and 4-bit activations in the early stage of the diffusion



(a) FP32 Activation

(b) 4-bit Activation for $[1.0T{:}0.8T]$

Figure 9: Examples of 256×256 LSUN-Bedrooms generation with different activation precision.

process. We find that the models with full-precision activations sometimes generate images with complex structures that are not easily recognizable as bedrooms. However, when 4-bit activation quantization is applied to the early stage, it simplifies the complex structures and results in images that more closely resemble bedrooms. This observation suggests that the step-wise activation quantization strategy employed in the proposed RAQ method helps refine the generated images, leading to improved quality and better alignment with the target LSUN-Bedrooms dataset.

B.3 Activation Quantization of Stable Diffusion

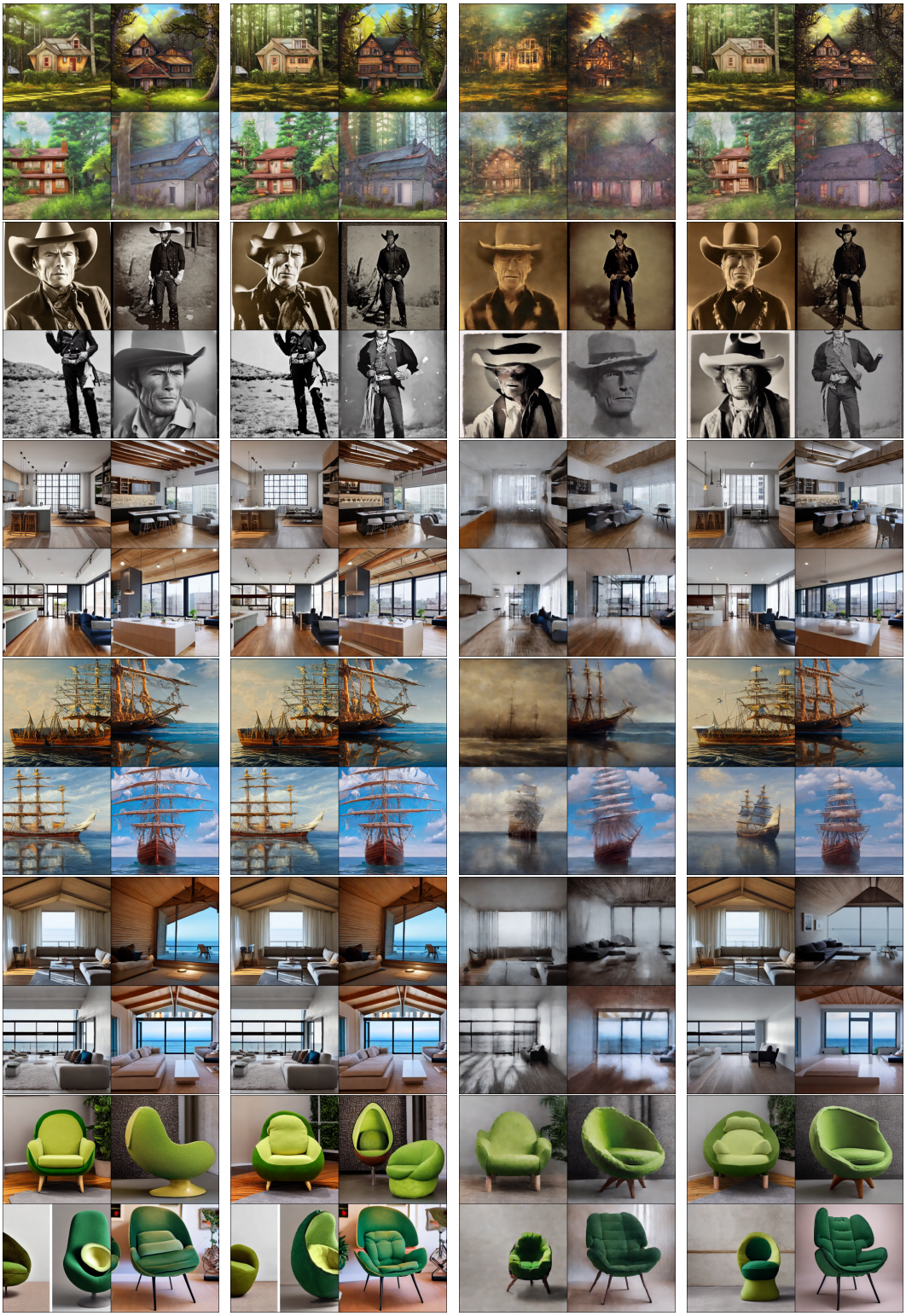
For the conditional image generation with Stable Diffusion, we utilize the prompt examples that are publicly available online [6] [7]. The prompts used in Section 5.2 are as follows:

1. a puppy wearing a hat
2. cluttered house in the woods in anime oil painting style*
3. Old photo of Clint Eastwood dressed as cowboy,1800s, centered, by professional photographer, wide-angle lens, background saloon*
4. interior design, open plan, kitchen and living room, modular furniture with cotton textiles, wooden floor, high ceiling, large steel windows viewing a city Artstation and Antonio Jacobsen and Edward Moran, (long shot), clear blue sky, intricated details, 4k*
5. a tree on the hill, bright scene, highly detailed, realistic photo
6. a highly detailed, majestic royal tall ship on a calm sea,realistic painting, by Charles Gregory*
7. medium shot side profile portrait photo of the Takeshi Kaneshiro warrior chief, tribal panther make up, blue on red, looking away, serious eyes, 50mm portrait, photography, hard rim lighting photography –ar 2:3 –beta –upbeta
8. a picture of dimly lit living room, minimalist furniture, vaulted ceiling, huge room, floor to ceiling window with an ocean view, nighttim*
9. an armchair in the shape of an avocado, an armchair imitating an avocado*

In this section, we additionally present non-cherry-picked samples generated using Stable Diffusion with and without activation quantization. We use the prompts that are highlighted with asterisk (*). For each prompt, we generate four images to demonstrate the variety and quality of the generated results (Fig. 10 in the next page).

[6] https://stablediffusion.fr/prompts

[7] https://mpost.io/best-100-stable-diffusion-prompts-the-most-beautiful-ai-text-to-image-prompts/



(a) Full Precision (b) W4A8 (Q-Diffusion) (c) W4A6 (d) W4A6/8 (Proposed)

Figure 10: Text-guided 512×512 image generation results with Stable Diffusion.