

A Appendix

A.1 Proof for Scale-Aware AdaRound

AdaRound (Nagel et al., 2020) and BRECQ (Li et al., 2021) use gradient descent to update the rounding mask. According to Eqs. 4, 5 and 6, ignoring the regularizer $f_{\text{reg}}(\mathbf{V})$, we have following theorem which provides theoretical evidence that the origin AdaRound is imbalance across different scales s .

Theorem 1. *Let s be the quantization scale corresponding to the rounding mask V . Then for gradient descent, given as $\mathbf{V}_{n+1} = \mathbf{V}_n - \alpha \nabla F(\mathbf{V}_n)$, the subtraction $\nabla F(\mathbf{V}_n)$ is dependent on the scalar s .*

Proof. Refers to Eqs. 4, 5 and gives:

$$\nabla F(\mathbf{V}_n) = \nabla_{\widetilde{W}} \|W\mathbf{x} - \widetilde{W}\mathbf{x}\|_F^2 \cdot s \cdot \frac{\partial h(\mathbf{V}_n)}{\partial \mathbf{V}_n} \quad (9)$$

Refers to Eq. 6 and gives:

$$\frac{\partial h(\mathbf{V})}{\partial \mathbf{V}} = (\zeta - \gamma) \cdot \sigma(\mathbf{V}) \cdot (1 - \sigma(\mathbf{V})) \quad (10)$$

Therefore, combining Eq. 9 and 10, the subtraction $\nabla F(\mathbf{V}_n)$ is scaled by s . \square

This theorem indicates that the imbalance gradient descent occurs if applying origin AdaRound to FPQ, as shown in Figure 4b.

In the section 3, we propose a scale-aware version of AdaRound in Eq. 7 and 8. Ignoring the regularizer f_{reg} , we have the following theorem which provides theoretical evidence that the scale-aware AdaRound has balanced gradient descent's update across different scales s .

Theorem 2. *Let s be the quantization scale corresponding to the rounding mask V' . Then for gradient descent, given as $\mathbf{V}'_{n+1} = \mathbf{V}'_n - \alpha \nabla F(\mathbf{V}'_n)$, the subtraction $\nabla F(\mathbf{V}'_n)$ is independent of the scalar s .*

Proof. The learning objective does not change for scale-aware AdaRound. Hence, from Eq. 4 and 7, we give:

$$\nabla F(\mathbf{V}'_n) = \nabla_{\widetilde{W}} \|W\mathbf{x} - \widetilde{W}\mathbf{x}\|_F^2 \cdot s \cdot \frac{\partial h'(\mathbf{V}'_n)}{\partial \mathbf{V}'_n} \quad (11)$$

Refer to Eq. 8 and give:

$$\frac{\partial h'(\mathbf{V}')}{\partial \mathbf{V}'} = \frac{(\zeta - \gamma)}{s} \cdot \sigma(\mathbf{V}') \cdot (1 - \sigma(\mathbf{V}')) \quad (12)$$

Combining Eq. 11 and 12, we get:

$$\nabla F(\mathbf{V}'_n) = \nabla_{\widetilde{W}} \|W\mathbf{x} - \widetilde{W}\mathbf{x}\|_F^2 \cdot (\zeta - \gamma) \cdot \sigma(\mathbf{V}') \cdot (1 - \sigma(\mathbf{V}'))$$

This result shows that subtraction $\nabla F(\mathbf{V}'_n)$ is independent of scale s . \square

Therefore, the gradient descent is balanced and normalized among different scale s with our scale-aware AdaRound, as depicted in Figure 4c.

A.2 Additional Ablation Studies

In addition to the ablation studies in the main text, we include multiple ablation studies here to further demonstrate the design of FP4DiT.

A.2.1 Effect of Group Size.

Algorithm 1 MinMax quantization for FP format

Require: Full-precision array A_{FP} , number of bits n , number of exponent bits n_e , number of mantissa bits n_m , clipping value $maxval$

$A_{\text{abs}} \leftarrow \text{abs}(A_{\text{FP}})$

$bias \leftarrow 2^{n_e} - \log_2(A_{\text{abs}}) + \log_2(2 - 2^{-n_m}) - 1$

$A_{\text{clip}} \leftarrow \min(\max(A_{\text{FP}}, -maxval), maxval)$

$S_{\text{log}} \leftarrow \text{clamp}(\lfloor \log_2(\text{abs}(A_{\text{clip}})) \rfloor + bias), 1)$

$S \leftarrow 2.0^{(S_{\text{log}} - n_m - bias)}$

$result \leftarrow \text{round-to-nearest}(A_{\text{clip}}/S) \times S$

return $result$

We compare the group-wise weight quantization result with different group sizes in Table 9. Decreasing the group size g for weight quantization consistently improves the quantization performance down to $g = 128$. According to Park et al. (2022), group size g such as 128 can result in substantial improvement while maintaining low latency.

A.2.2 Effect of Calibration Dataset

We investigate FP4DiT’s effectiveness under different calibration dataset sizes N . Table 10 reports the HPSv2 score of the W4A8 PixArt- α quantized by our method with $N = 64, 128, 256$, and 512. The results show steady improvements as N increases, with performance saturating beyond $N = 256$. In particular, FP4DiT achieves a strong score of 27.43 with only 128 samples, and further gains to 28.21 at $N = 256$, while larger calibration does not bring additional benefits. This demonstrates that FP4DiT requires only a relatively small calibration set to achieve near-optimal performance, making it both efficient and effective compared to existing approaches.

A.2.3 Classifier-Free Guidance (CFG)

Precision	CFG=3.0	CFG=4.5	CFG=7.0
W4A8	27.22	27.43	26.74
Full Precision	30.97	31.01	31.26

Table 11: The quantization results for different CFG scales with PixArt- α .

Model	Group Size	FID ↓
PixArt- α	128	96.58
	256	100.60
	512	113.78
	1024	174.38

Table 9: The quantization results for different group sizes with PixArt- α .

N=64	N=64	N=64	N=64
25.09	27.43	28.21	28.20

Table 10: The quantization results for different sizes of calibration dataset with PixArt- α .

We investigate FP4DiT’s effectiveness under different CFG scales. Table 11 compares FP4DiT’s quantization performance with PixArt- α for three different CFG scales, which are 3.0, 4.5 (default), and 7.0. Although FP4DiT is calibrated at CFG = 4.5, it remains robust and exhibits only minor degradation under different CFG scales, which is still outstanding among the baseline methods as shown in Table 2.

A.3 Extended Experimental Settings**A.3.1 Floating Point Quantization Scheme**

Since Floating Point Quantization (FPQ) is not as straightforward as INT quantization, there has not been a simple and unified algorithm for performing FPQ yet. In this paper, we apply Algorithm 1 from (Kuzmin et al., 2022) to perform the FPQ. Unlike (Kuzmin et al., 2022) learning the clipping value $maxval$ and bit allocations between mantissa and exponent part, we use the absolute maximum of the tensor as $maxval$ and perform FPQ with a predetermined FP format.

In addition to the FPQ algorithm and the group-wise/token-wise quantization, we provide our quantization hyperparameters in Table 12. Note that the calibration size refers to the number of images we sampled from MS-COCO. For each image, we sample its input latent noise across 20 timesteps (50 for Hunyuan) as the calibration data for AdaRound. The activation FP format is for W4A6/W4A8 respectively.

A.4 Evaluation Settings

Model	Cali. Size	Cali. Step	Weight Format	Act. Format
PixArt- α	128	2500	E2M1	E2M3/E3M4
PixArt- Σ	128	2500	E1M2	E2M3/E3M4
Hunyuan	64	2500	E2M1	E2M3/E3M4

Table 12: Quantization calibration hyperparameters for FP4DiT.

the provided human preference predictor to estimate the performance. For the Fréchet Inception Distance (FID) (Heusel et al., 2017) measurement, we apply clean-fid ² library to measure the FID between generated images and ground-truth images.

For CLIP score (Hessel et al., 2021), we apply the openai-clip ¹ library to measure the CLIP score between prompts and generated images. We employ ViT-B/32 as the CLIP model. To measure HPSv2 (Wu et al., 2023a) score, we generate 3.2k images across the four categories (800 images per category) and use

A.5 Hardware and Software Resources

We execute our FP4DiT and baseline experiments on two rack servers. The first is equipped with 2 Nvidia A100 80 GPUs, an AMD EPYC-Rome Processor and 512GB RAM. The second server has 8 Nvidia V100 32GB GPUs, an Intel Xeon Gold GPU and 756GB RAM.

Our code is running under Python 3 using Anaconda virtual environments and open-source repository forks based on Q-Diffusion (Li et al., 2023). We modified the code to implement our FP4DiT and enable the interface between Quantization scripts and the Hugging Face Diffusers ³ library. The hardware cost measurement is conducted using the ptflops ⁴ library. We provide a code implementation with README listing all necessary details and steps.

A.6 Additional Visualization Results

In this section, we present the random samples derived from full-precision and W4A6/W4A8 PixArt- α , PixArt- Σ , and Hunyuan that are quantized by different baseline and FP4DiT. As depicted by Figures 11, 12, 13, 14, 15, and 16, FP4DiT generates results of impressive visual content. FP4DiT consistently shows superior performance across various DiT models. We list a detailed comparison between FP4DiT and other baselines as follows:

1. PixArt- α : On W4A8 (Fig. 11), FP4DiT generates more fine-grained images than all other baselines, such as the texture of macarons and waves. Besides, other baselines fail to generate a hedgehog while FP4DiT correctly depicts it. On W4A6 (Fig. 12), our method shows near W4A8 performance, while other baselines become noisy and lose details.
2. PixArt- Σ : On W4A8 (Fig. 13), there is almost no noise on FP4DiT’s generated images, which demonstrates the improvement of our method. As for the image details, FP4DiT has a more natural yoga posture, a more fine-grained model face, and more detailed cat hair and whiskers. Similar to PixArt- α , FP4DiT’s W4A6 (Fig. 14) images have the least noise and best image quality.
3. Hunyuan: On W4A8 (Fig. 15), FP4DiT’s generated images more closely align with the full precision model. For example, in the first image, FP4DiT has the same color right face while Q-Diffusion and TMFQ-DM alter the face’s color. In the third image, FP4DiT generates the most similar face

¹<https://github.com/openai/CLIP>

²<https://github.com/GaParmar/clean-fid>

³<https://huggingface.co/docs/diffusers/en/index>

⁴<https://github.com/sovrasov/flops-counter.pytorch>

decoration. In addition, FP4DiT also has the least noise on Hunyuan DiT. On W4A6 (Fig. 16), all methods generate blur images. However, the contour of images identified from FP4DiT’s visualization results matches the full precision images, while other baselines fail to produce a contour or the contour is not prompt-adherent.



Figure 11: More visualization result for W4A8 PixArt- α . Prompts: ‘A blue jay standing on a large basket of rainbow macarons.’; ‘Frog, in forest, colorful, no watermark, no signature, in forest, 8k.’; ‘Drone view of waves crashing against the rugged cliffs along Big Sur’s Garay Point beach. The crashing blue waters create white-tipped waves, while the golden light of the setting sun illuminates the rocky shore.’; ‘An ink sketch style illustration of a small hedgehog holding a piece of watermelon with its tiny paws, taking little bites with its eyes closed in delight. Photo of a lychee-inspired spherical chair, with a bumpy white exterior and plush interior, set against a tropical wallpaper.’

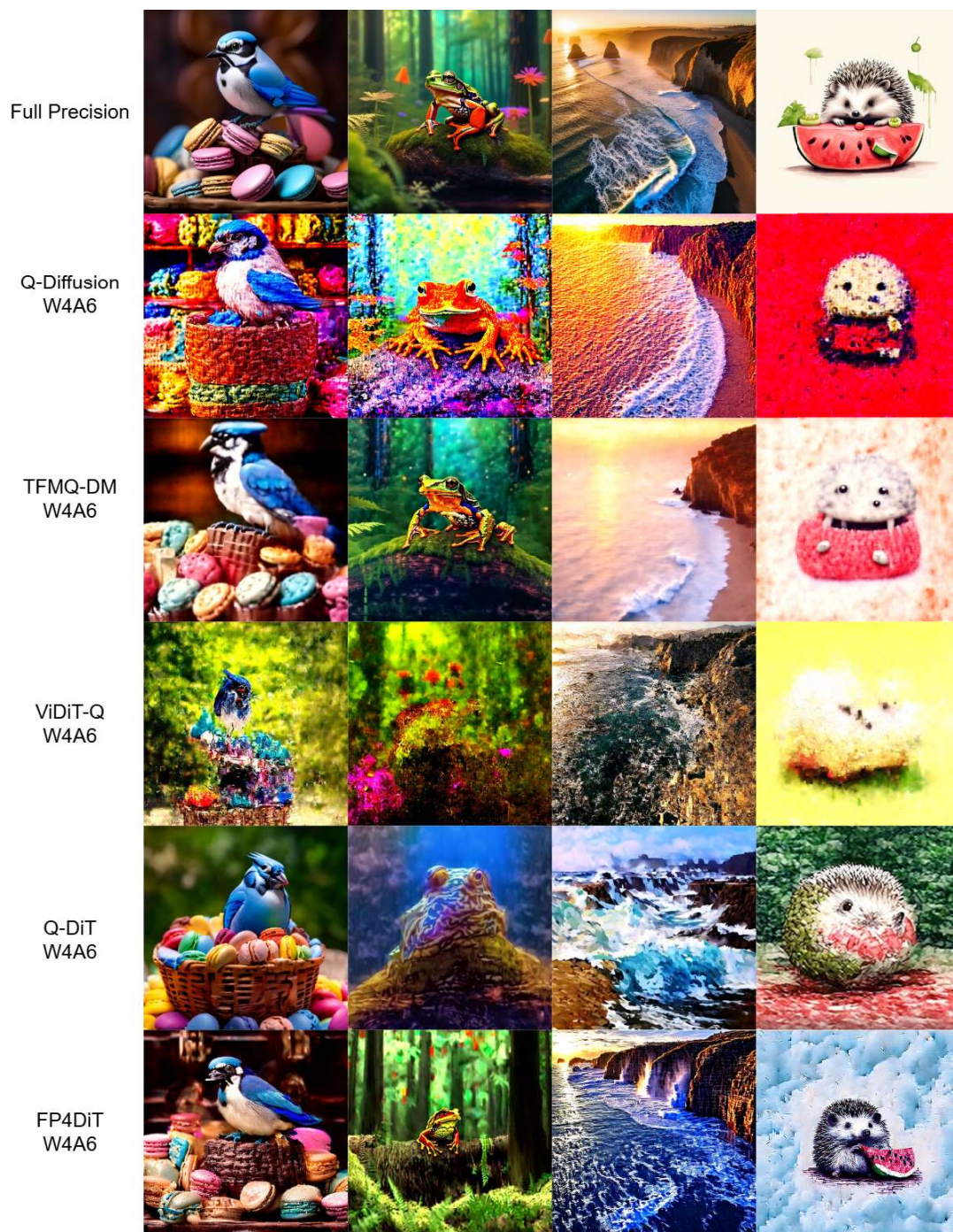


Figure 12: More visualization result for W4A6 PixArt- α . Prompts: ‘A blue jay standing on a large basket of rainbow macarons.’; ‘Frog, in forest, colorful, no watermark, no signature, in forest, 8k.’; ‘Drone view of waves crashing against the rugged cliffs along Big Sur’s Garay Point beach. The crashing blue waters create white-tipped waves, while the golden light of the setting sun illuminates the rocky shore.’; ‘An ink sketch style illustration of a small hedgehog holding a piece of watermelon with its tiny paws, taking little bites with its eyes closed in delight. Photo of a lychee-inspired spherical chair, with a bumpy white exterior and plush interior, set against a tropical wallpaper.’

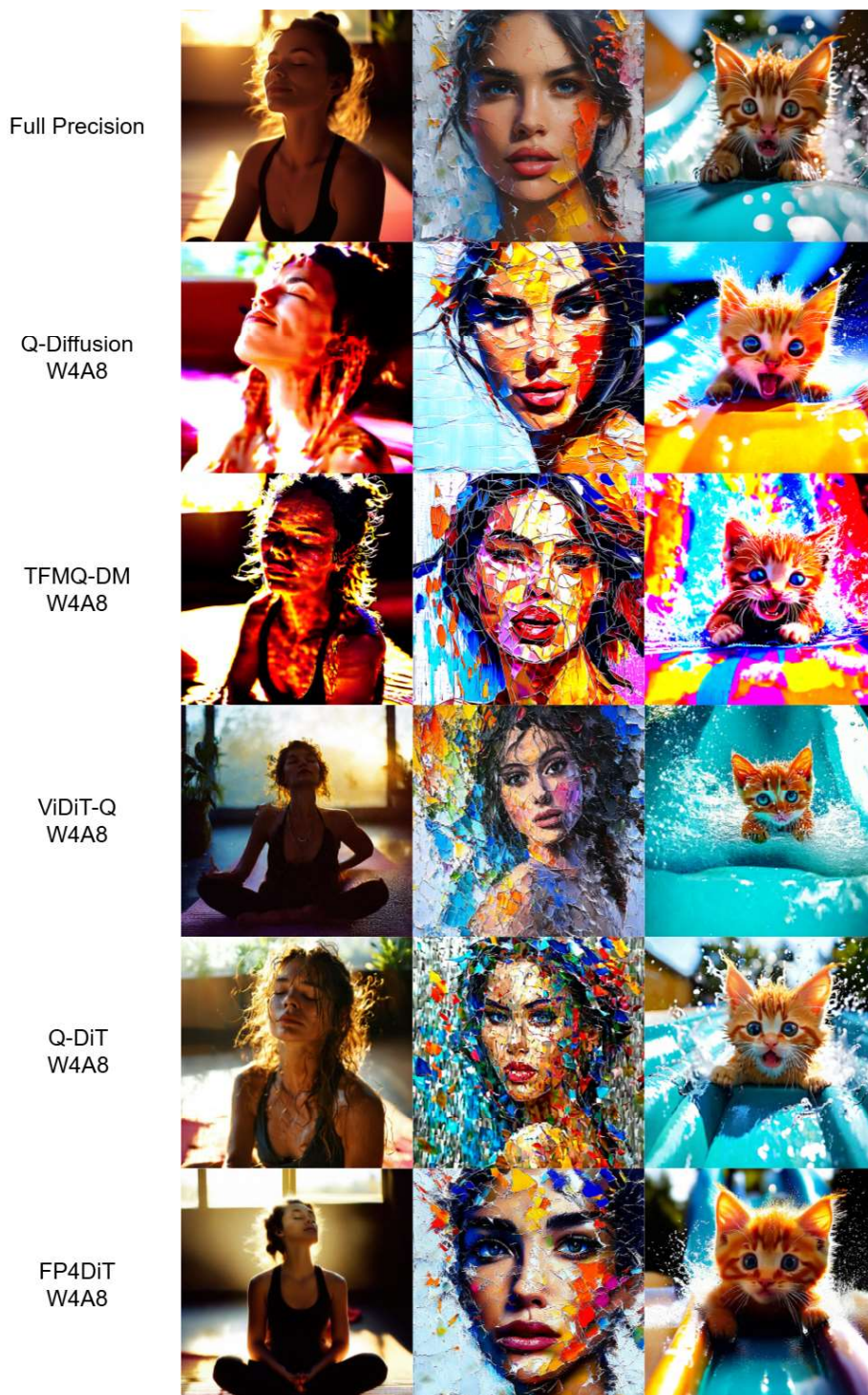


Figure 13: More visualization result for W4A8 PixArt-Σ. Prompts: ‘A very attractive and natural woman, sitting on a yoka mat, breathing, eye closed, no make up, intense satisfaction, she looks like she is intensely relaxed, yoga class, sunrise, 35mm.’; ‘Realistic oil painting of a stunning model merged in multicolor splash made of finely torn paper, eye contact, walking with class in a street.’; ‘A cute orange kitten sliding down an aqua slide. happy excited. 16mm lens in front. we see his excitement and scared in the eye. vibrant colors. water splashing on the lens.’

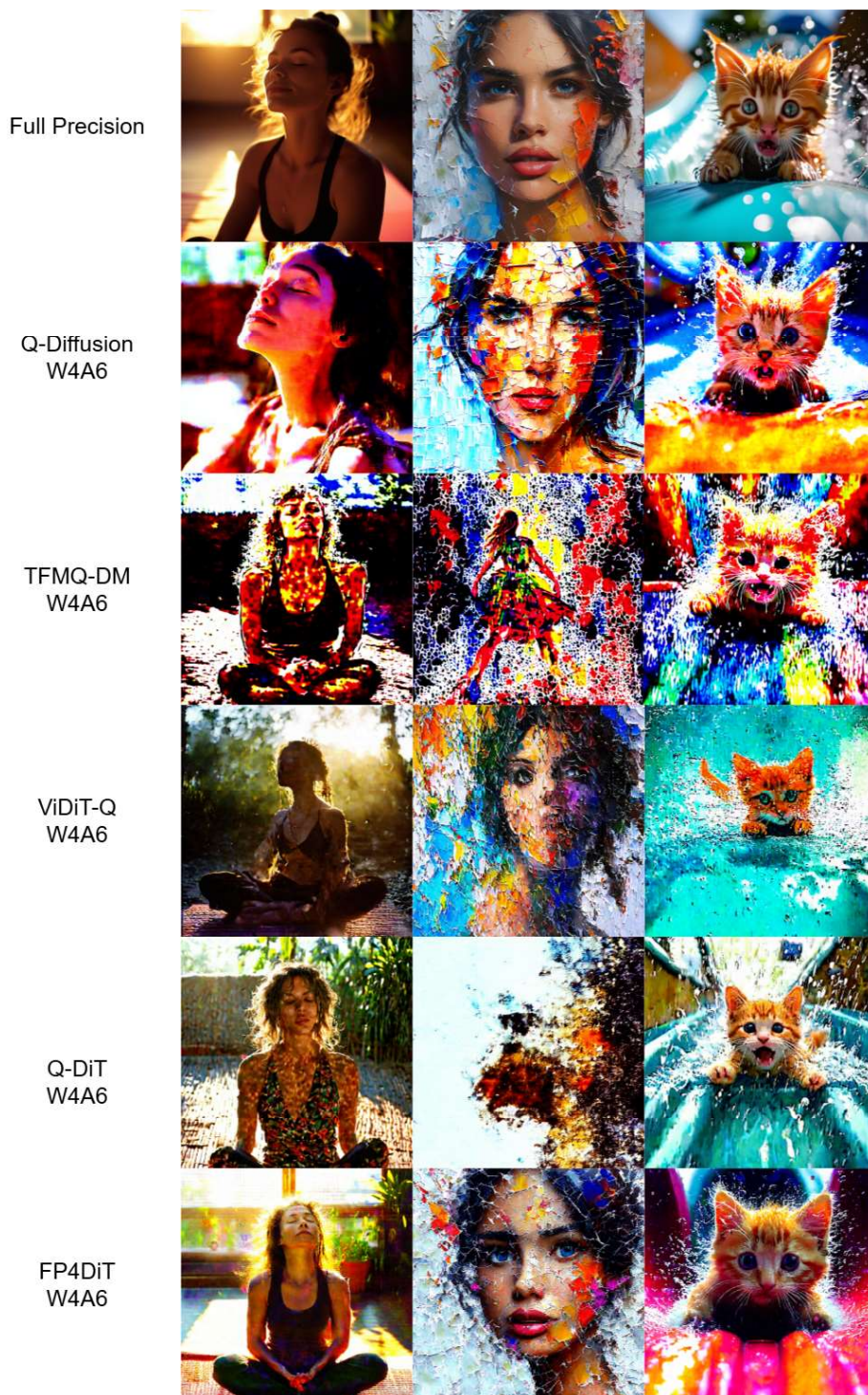


Figure 14: More visualization result for W4A6 PixArt-Σ. Prompts: ‘A very attractive and natural woman, sitting on a yoka mat, breathing, eye closed, no make up, intense satisfaction, she looks like she is intensely relaxed, yoga class, sunrise, 35mm.’; ‘Realistic oil painting of a stunning model merged in multicolor splash made of finely torn paper, eye contact, walking with class in a street.’; ‘A cute orange kitten sliding down an aqua slide. happy excited. 16mm lens in front. we see his excitement and scared in the eye. vibrant colors. water splashing on the lens.’



Figure 15: More visualization result for W4A8 Hunyuan. Prompts: ‘nature vs human nature, surreal, UHD, 8k, hyper details, rich colors, photograph.’; ‘A transparent sculpture of a duck made out of glass. The sculpture is in front of a painting of a landscape.’; ‘Steampunk makeup, in the style of vray tracing, colorful impasto, uhd image, indonesian art, fine feather details with bright red and yellow and green and pink and orange colours, intricate patterns and details, dark cyan and amber makeup. Rich colourful plumes. Victorian style.’

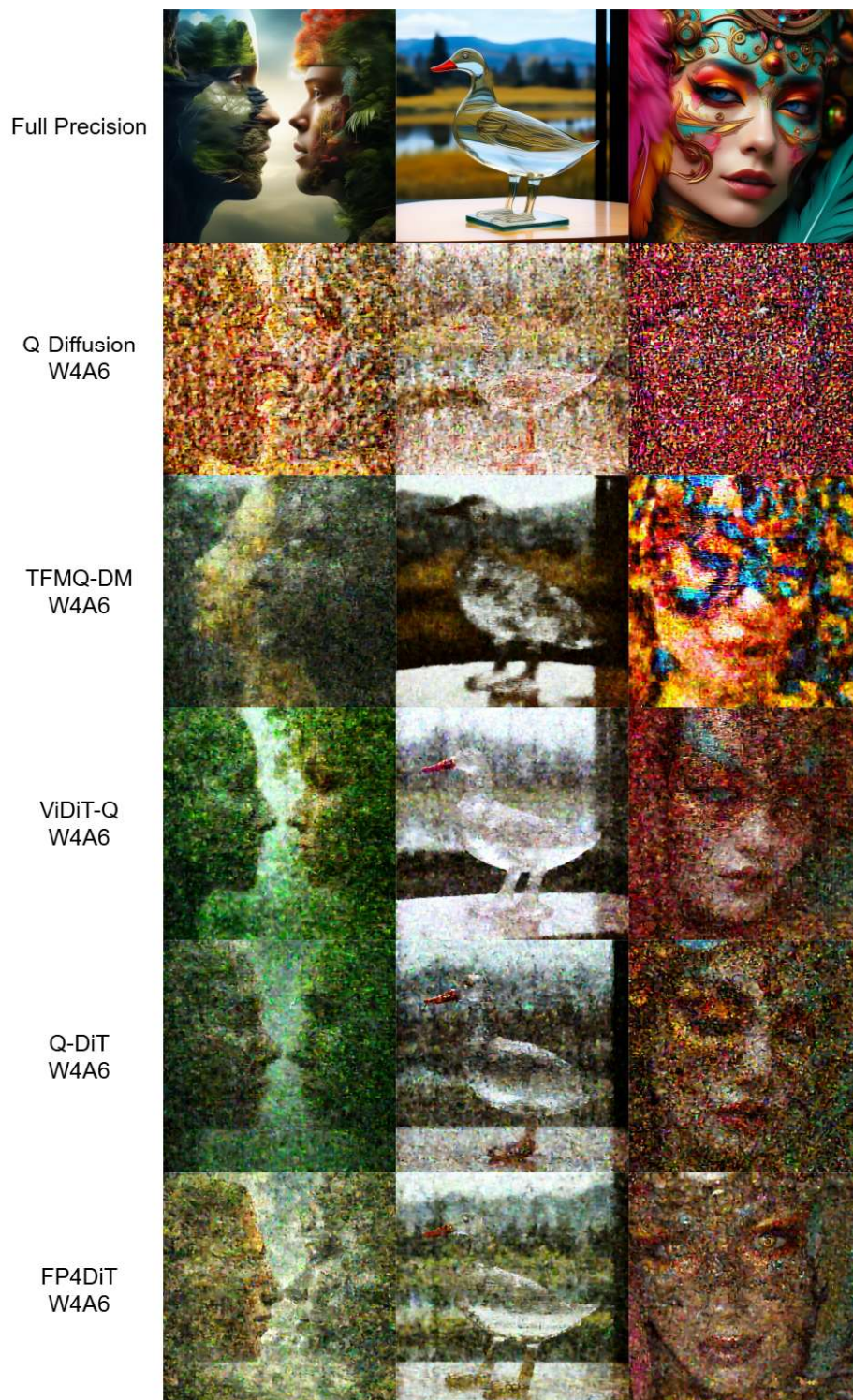


Figure 16: More visualization result for W4A6 Hunyuan. Prompts: ‘nature vs human nature, surreal, UHD, 8k, hyper details, rich colors, photograph.’; ‘A transparent sculpture of a duck made out of glass. The sculpture is in front of a painting of a landscape.’; ‘Steampunk makeup, in the style of vray tracing, colorful impasto, uhd image, indonesian art, fine feather details with bright red and yellow and green and pink and orange colours, intricate patterns and details, dark cyan and amber makeup. Rich colourful plumes. Victorian style.’