Appendix for *Quantization without Tears*

A. Full Implementation Details

In this section, we present full implementation details of the different types of tasks in our experiments.

Image Classification. We selected RepQ-ViT [6], PTQ4ViT [14], and Percentile [5] as the primary baseline PTQ methods to integrate with our QwT modules. Following [6], we randomly sampled 32 images from the ImageNet [2] dataset as the calibration set to initialize the quantized weights for these baseline methods. Additionally, a separate set of 512 randomly selected images from the ImageNet training set was used to initialize the parameters of the QwT modules (excluding PTQ weights). For all networks, the affine transformation matrix W in QwT is implemented in FP16 format to reduce model size. In ResNet, W is further simplified as a group-wise convolution with a kernel size of 1 and 64 channels per group.

When finetuning the QwT modules along with the classification head for an additional epoch, we utilized AdamW [9] as the optimizer. The batch size was set to 32 per GPU (using a total of 4 GPUs), and weight decay was set to 0. The learning rate was configured to 1e-7 for ViT [3], 5e-6 for DeiT [12] and Swin [8], 1e-5 for ResNet [4].

In addition to the original classification loss, during 1-epoch finetuning we applied a simple distillation loss to minimize the squared L2-distance between the full-precision and quantized models—calculated on the output features before the classification head (*cls token* for ViT and DeiT, *global average feature* for Swin and ResNet), yielding the finetuning objective as $L_{cls} + L_{dis}$ (*i.e.*, without the combination weight hyperparameter.)

This distillation strategy is the feature mimicking method [13], which only utilizes the penultimate features and argues that features or activations from intermediate layers are not necessary or even harmful. It is also worth noting since only the penultimate features are required, feature mimicking is unsupervised.

Object Detection & Instance Segmentation. Following [6], we randomly sampled a single image from the COCO dataset [7] to initialize the quantized weights for baseline PTQ methods. All other details are consistent with the image classification case.

Image Generation. Consistent with the experimental setup of Q-DiT [1], we selected the DiT architecture and employed pretrained DiT-XL/2 models at a resolution of 256×256. Our experiments were extended to a broader range of settings, including varying the number of sampling steps (50 and 100) and classifier-free guidance (CFG) scales (0 and 1.5). These results are presented in the next section.

B. More Experimental Results

In this section, we provide more comprehensive quantization results across a range of backbones [3, 4, 8, 12] on the ImageNet dataset [2], as summarized in Tables 1, 2, and 3.

We also show the full results for large language models in Table 4. The main text only reports the overall average accuracy on eight zero-shot commonsense QA datasets. Table 4 lists the accuracy for each dataset separately. We also include the results on the MMLU benchmarks, tested in both zero-shot and five-shot modes.

The full results of image generation are summarized in Table 5. As shown in the table, our method consistently enhances the performance of the generative model across all tested configurations. To provide a more intuitive understanding, we visualize the generated images under each setting in Figure 1. Similar to the main paper, we ensured that the noise during the generation process remains consistent across all models. The visualizations further confirm that our method reliably improves the quality of the generated images. As a further illustration, we provide several representative images generated by our method in Figure 2.

References

[1] Lei Chen, Yuan Meng, Chen Tang, et al. Q-DiT: Accurate post-training quantization for diffusion Transformers. *arXiv:2406.17343*, 2024. 1

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 1

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, pages 1–21, 2021. 1, 4

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Con-*

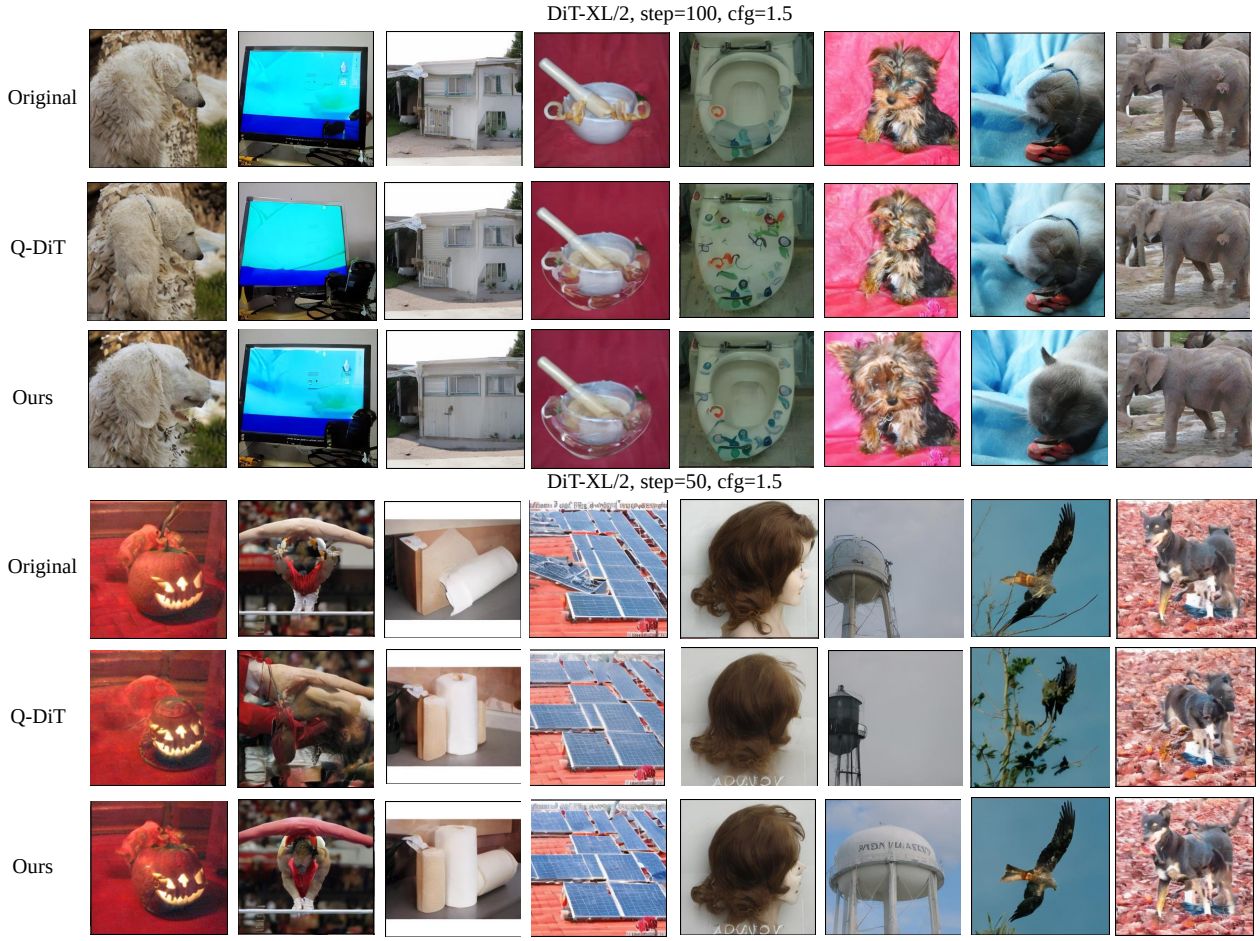

Figure 1. Qualitative visualization results of quantizing DiT-XL/2 on different settings.

ference on Computer Vision and Pattern Recognition, pages 770–778, 2016. 1, 4

[5] Rundong Li, Yan Wang, Feng Liang, Hongwei Qin, Junjie Yan, and Rui Fan. Fully quantized network for object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2805–2814, 2019. 1, 4

[6] Zhikai Li, Junrui Xiao, Lianwei Yang, and Qingyi Gu. RepQ-ViT: Scale reparameterization for post-training quantization of Vision Transformers. In *IEEE/CVF International Conference on Computer Vision*, pages 17181–17190, 2023. 1, 4

[7] Tsung-Yi Lin, Michael Maire, Serge Belongie, et al. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, pages 740–755, 2014. 1

[8] Ze Liu, Yutong Lin, Yue Cao, et al. Swin Transformer: Hierarchical Vision Transformer using shifted windows. In *IEEE/CVF International Conference on Computer Vision*, pages 9992–10002, 2021. 1, 4

[9] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, pages 1–18, 2019. 1

[10] Jaehyeon Moon, Dohyung Kim, Junyong Cheon, and Bumsub Ham. Instance-aware group quantization for Vision Transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16132–16141, 2024. 4

[11] Yuzhang Shang, Gaowen Liu, Ramana Rao Kompella, and Yan Yan. Enhancing post-training quantization calibration through contrastive learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15921–15930, 2024. 4

[12] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357, 2021. 1, 4

[13] Guo-Hua Wang, Yifan Ge, and Jianxin Wu. Distilling knowledge by mimicking features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):8183–8195, 2021. 1

[14] Zhihang Yuan, Chenhao Xue, Yiqi Chen, Qiang Wu, and Guangyu Sun. PTQ4ViT: Post-training quantization for Vision Transformers with twin uniform quantization. In *European Conference on Computer Vision*, pages 191–207, 2022. 1

Figure 2. More qualitative visualization results of our method on quantized DiT-XL/2.

Table 1. Full results on ViT [3] and DeiT [12] backbones.

Network	Method	#Bits	Size	Top-1
DeiT-T	Full-precision	32/32	22.9	72.2
	IGQ-ViT† [10]	4/4	-	62.5
	RepQ-ViT [6]	4/4	3.3	58.2
	RepQ-ViT + QwT	4/4	4.2	61.4
	RepQ-ViT + QwT*	4/4	4.2	**64.8**
	IGQ-ViT† [10]	6/6	-	71.2
	RepQ-ViT [6]	6/6	4.6	71.0
	RepQ-ViT + QwT	6/6	5.5	71.2
	RepQ-ViT + QwT*	6/6	5.5	**71.6**
DeiT-S	Full-precision	32/32	88.2	79.9
	IGQ-ViT† [10]	4/4	-	74.7
	RepQ-ViT [6]	4/4	11.9	69.0
	RepQ-ViT + QwT	4/4	15.4	71.5
	RepQ-ViT + QwT*	4/4	15.4	**75.2**
	IGQ-ViT† [10]	6/6	-	79.3
	RepQ-ViT [6]	6/6	17.2	78.9
	RepQ-ViT + QwT	6/6	20.7	79.1
	RepQ-ViT + QwT*	6/6	20.7	**79.3**
ViT-S	Full-precision	32/32	88.2	81.4
	IGQ-ViT† [10]	4/4	-	**73.6**
	RepQ-ViT [6]	4/4	11.9	65.8
	RepQ-ViT + QwT	4/4	15.4	70.8
	RepQ-ViT + QwT*	4/4	15.4	72.9
	IGQ-ViT† [10]	6/6	-	80.8
	RepQ-ViT [6]	6/6	17.2	80.5
	RepQ-ViT + QwT	6/6	20.7	80.7
	RepQ-ViT + QwT*	6/6	20.7	**80.8**
ViT-B	Full-precision	32/32	346.3	84.5
	IGQ-ViT† [10]	4/4	-	**79.3**
	RepQ-ViT [6]	4/4	44.9	68.5
	RepQ-ViT + QwT	4/4	59.1	76.3
	RepQ-ViT + QwT*	4/4	59.1	78.5
	IGQ-ViT† [10]	6/6	-	83.8
	RepQ-ViT [6]	6/6	66.2	83.6
	RepQ-ViT + QwT	6/6	80.4	83.9
	RepQ-ViT + QwT*	6/6	80.4	**84.0**

Table 2. Full results on the Swin [8] backbone.

Network	Method	#Bits	Size	Top-1
Swin-T	Full-precision	32/32	113.2	81.4
	IGQ-ViT† [10]	4/4	-	77.8
	RepQ-ViT [6]	4/4	14.9	73.0
	RepQ-ViT + QwT	4/4	19.2	75.5
	RepQ-ViT + QwT*	4/4	19.2	**79.3**
	IGQ-ViT† [10]	6/6	-	80.9
	RepQ-ViT [6]	6/6	21.7	80.6
	RepQ-ViT + QwT	6/6	26.0	80.7
	RepQ-ViT + QwT*	6/6	26.0	**80.9**
Swin-S	Full-precision	32/32	198.4	83.2
	IGQ-ViT† [10]	4/4	-	81.0
	RepQ-ViT [6]	4/4	25.8	80.2
	RepQ-ViT + QwT	4/4	33.7	80.4
	RepQ-ViT + QwT*	4/4	33.7	**81.9**
	IGQ-ViT† [10]	6/6	-	82.9
	RepQ-ViT [6]	6/6	38.0	82.8
	RepQ-ViT + QwT	6/6	45.9	82.9
	RepQ-ViT + QwT*	6/6	45.9	**82.9**

Table 3. Full results on the ResNet [4] backbone.

Network	Method	#Bits	Size	Top-1
ResNet-18	Full-precision	32/32	46.8	71.0
	CL-Calib† [11]	4/4	-	69.4
	Percentile[5]	4/4	6.1	58.3
	Percentile + QwT	4/4	6.4	68.9
	Percentile + QwT*	4/4	6.4	**69.4**
	CL-Calib† [11]	6/6	-	-
	Percentile[5]	6/6	8.9	70.7
	Percentile + QwT	6/6	9.2	71.0
	Percentile + QwT*	6/6	9.2	**71.1**
ResNet-50	Full-precision	32/32	102.2	76.6
	CL-Calib† [11]	4/4	-	75.4
	Percentile[5]	4/4	14.0	68.4
	Percentile + QwT	4/4	16.0	74.5
	Percentile + QwT*	4/4	16.0	**75.8**
	CL-Calib† [11]	6/6	-	-
	Percentile[5]	6/6	19.9	76.0
	Percentile + QwT	6/6	21.9	76.8
	Percentile + QwT*	6/6	21.9	**76.8**
ResNet-101	Full-precision	32/32	178.2	77.3
	CL-Calib† [11]	4/4	-	-
	Percentile[5]	4/4	23.7	74.7
	Percentile + QwT	4/4	28.0	76.4
	Percentile + QwT*	4/4	28.0	**76.7**
	CL-Calib† [11]	6/6	-	-
	Percentile[5]	6/6	34.3	77.1
	Percentile + QwT	6/6	38.6	77.2
	Percentile + QwT*	6/6	38.6	**77.2**

Table 4. Detailed quantization results among the MMLU dataset and eight zero-shot commonsense QA datasets using LLaMA3-8B as the backbone.

Method	#Bits	MMLU (0-shot)	MMLU (5-shot)	BoolQ	PIQA	SIQA	HLSW	WG	ARC-e	ARC-c	OBQA	QA. Avg
Full-precision	16	63.39	65.30	82.17	81.18	32.91	78.93	73.95	81.14	53.50	45.00	66.10
GPTQ	4	61.40	63.94	81.25	81.39	32.91	78.28	72.77	78.03	50.60	44.00	64.90
GPTQ + QwT	4	**61.57**	**64.25**	81.22	81.45	32.91	77.77	73.40	79.21	50.68	44.80	**65.18**

Table 5. Quantitative results of quantizing DiT-XL/2 on ImageNet 256×256.

Model	**Bit-width (W/A)**	**Method**	**Size (MB)**	**FID (↓)**	**sFID (↓)**	**IS (↑)**	**Precision (↑)**	**Recall (↑)**
DiT-XL/2 (steps = 100)	16/16	FP	1349	12.40	19.11	116.68	0.6605	-
	4/8	PTQ4DM	339	252.31	82.44	2.74	0.0125	-
		RepQ-ViT	339	315.85	139.99	2.11	0.0067	-
		GPTQ	351	25.48	25.57	73.46	0.5392	-
		Q-DiT	347	15.76	19.84	98.78	**0.6395**	-
		Q-DiT + QwT	361	**15.35**	**19.63**	**104.04**	0.6373	**0.7478**
DiT-XL/2 (steps = 100, cfg = 1.5)	16/16	FP	1349	5.31	17.61	245.85	0.8077	-
	4/8	PTQ4DM	339	255.06	84.63	2.76	0.0110	-
		RepQ-ViT	339	311.31	138.58	2.18	0.0072	-
		GPTQ	351	7.66	20.76	193.76	0.7261	-
		Q-DiT	347	6.40	18.60	211.72	0.7609	-
		Q-DiT + QwT	361	**5.86**	**18.29**	**221.66**	**0.7678**	**0.6915**
DiT-XL/2 (steps = 50)	16/16	FP	1349	13.47	19.31	114.71	0.6601	-
	4/8	PTQ4DM	339	256.15	83.45	2.73	0.0150	-
		RepQ-ViT	339	324.25	142.98	2.12	0.0062	-
		GPTQ	351	26.31	25.54	69.73	0.5388	-
		Q-DiT	347	17.42	19.95	97.52	0.6219	-
		Q-DiT + QwT	361	**17.02**	**19.57**	**99.62**	**0.6302**	**0.7582**