HadaNorm: Diffusion Transformer Quantization through Mean-Centered Transformations

Marco Federici [*1] **Riccardo Del Chiaro** [*1] **Boris van Breugel** [1] **Paul Whatmough** [1] **Markus Nagel** [1]

Abstract

Diffusion models represent the cutting edge in image generation, but their high memory and computational demands hinder deployment on resource-constrained devices. Post-Training Quantization (PTQ) offers a promising solution by reducing the bitwidth of matrix operations. However, standard PTQ methods struggle with outliers, and achieving higher compression often requires transforming model weights and activations before quantization. In this work, we propose HadaNorm, a novel linear transformation that extends existing approaches by both normalizing channels activations and applying Hadamard transforms to effectively mitigate outliers and enable aggressive activation quantization. We demonstrate that HadaNorm consistently reduces quantization error across the various components of transformer blocks, outperforming state-of-the-art methods.

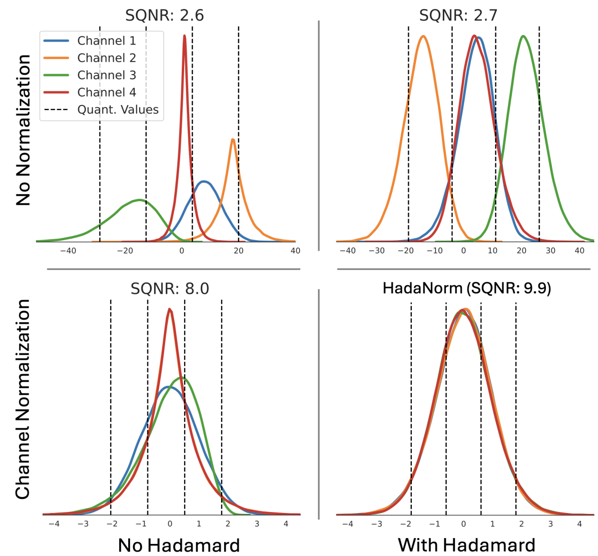

Figure 1. **HadaNorm reduces quantization error.** We take an illustrative setting of four channels with different distributions (top-left). Normalization (bottom-left) improves quantization, but it does not mix channels and hence cannot get rid of heavy tails. Hadamard transform (HT) (top-right) suffers when the channels have different means. HadaNorm (bottom-right) achieves better whitening, by both normalizing and applying the HT. The more Gaussian is easier to quantize.

1. Introduction

Diffusion models have emerged as the leading technique in deep learning for image generation, offering unparalleled visual realism. However, this advancement comes at a significant computational cost, primarily due to the large model sizes and the iterative denoising procedures required for each image generation. As the demand for scalable and efficient deployment of these models grows both on the cloud and on the edge where computational resources are scarce, optimizing inference efficiency has become a critical area of research.

Post-Training Quantization (PTQ) presents a promising solution for enhancing inference efficiency by quantizing weights and activations, especially for high-power demanding operations such as linear layers and large matrix multiplications. Despite its potential, PTQ faces substantial challenges, particularly when pushing activation bitwidth to 8-bits (A8) or weight to 4-bits (W4). Outliers are a major hurdle for low-bitwidth quantization: they can only be represented when a large quantization grid is used, leading to large bins and loss of precision for the majority of data. One approach to reduce outliers and consequently improve quantization is to apply invertible transformations to weights and activations, which do not alter the overall model output, but which do allow scaling outlier channels (Xiao et al., 2023). Follow-up works (Ashkboos et al., 2024; Liu et al., 2024; Ma et al., 2024; Zhao et al., 2025) use fast Hadamard transforms to mix channels, which effectively whitens the distribution and reduces outliers.

This paper extends previous work by introducing a sim-

[*]Equal contribution [1]Qualcomm AI Research. Correspondence to: Marco Federici <mfederic@qti.qualcomm.com>, Riccardo Del Chiaro <rdelchia@qti.qualcomm.com>.

Proceedings of the 42nd International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

ple yet effective centering transformation, which can be combined with existing approaches to further reduce quantization error.

Contributions. Our contributions are as follows:

1. We argue why the efficacy of the Hadamard transform for activation quantization is reduced due to mean and scale differences across channels (Figure 1).

2. We introduce HadaNorm, a simple transformation that applies centering and rescaling of channels together with a Hadamard transform to further improve performance of activation quantization.

3. Empirically, we show that HadaNorm significantly improves quantization performance in image diffusion transformers (DiT), resulting in a state-of-the-art CLIP score of 31.69 on PixArt-Sigma at W4A4.

2. Related Work

Transformations for better quantization are most explored in the Large Language Models (LLM) space. Xiao et al. (2023) find that activation quantization is harder than weight quantization, and propose per-channel scales to move outliers from activations to weights.

Ma et al. (2024); Shao et al. (2024) propose an affine transformation, that effectively adds a static bias term for LLM quantization. Ashkboos et al. (2024) propose to use Hadamard transforms, which are cheap at inference time, yet allow spreading of outliers across channels. (Zhao et al., 2025) combines channel scaling and mixing to achieve higher compression on DiT architectures, although mixed precision is required to address quantization-sensitive components in the architecture.

Li et al. (2024) further extends previous work on channel scaling Xiao et al. (2023) by introducing high-precision low-rank matrices to absorb the weight outliers prior to weight quantization, achieving higher compression levels at the cost of a small overhead. Shao et al. (2025) defines time-dependent transformation to address changes in the activation distribution during subsequent denoising steps, by applying transformation tailored to the activation distributions based on Lin et al. (2024), at the cost of additional overhead at inference time. Our method further extends the literature on transformation applied to DiT models by introducing a simple centering strategy that does not require additional calibration and introduces minimal overhead, yet consistently improves upon existing strategies.

3. Method

3.1. Distributional differences between channels reduces effectiveness of per-token quantization

Some of the related works use Hadamard transforms to mix channels, thereby spreading outliers over multiple channels and making each channel distribution approach a Gaussian distribution (Liu et al., 2025). This can be intuitively motivated by the central limit theorem: the Hadamard transform (HT) effectively adds up different channels (with signs flipped at times).

We make the simple observation that, after applying HT, channels may still exhibit substantially different mean. This is often the case in vision transformer models, as channels tend to have substantially different moments. Since all channels are quantized using the same quantization grid, this results in sub-optimal quantization.

We illustrate this using a toy example. In Figure 1 top-left, we plot the distribution of four channels. If we naively choose a quantization grid based on all four channels, we see that the quantization error is large, as many channels collapse to just one or two values. When we apply a HT (top-right), we find that indeed each channel is closer to a normal distribution. However, due to non-zero mean of the initial channels, we observe large difference in the means across channels after applying the HT. For example, channel 2 corresponds to the row $[+1, -1, +1, -1]$ of the HT, which due to the large positive mean of the initial channel 2, and large negative mean of the initial channel 3, leads to a large negative mean after the HT.

The solution is to apply both channel normalization and Hadamard transform. In the bottom-left figure, we show the effect of normalizing the channels. Although the SQNR is significantly better than before, channels are not mixed and outliers can heavily affect the scale of the quantization grid. Normalization alone does not affect the kurtosis of the channel distributions. We propose HadaNorm, which combines the dynamic channel normalization with the HT. Though simple, we get the best of both worlds: the HT mixes the channel distributions to reduce heavy tails, and because channel distributions are normalized, per-token quantization is more effective. Let us consider how this can be applied to a transformer model.

3.2. HadaNorm

To reduce the distributional differences between channels, ideally, we would want to both normalize the channels and apply the Hadamard transform. Let us consider what this means for a linear layer. Let us assume input vector $\mathbf{X} \in \mathbb{R}^{s\times d}$, a linear layer with weights $\mathbf{W} \in \mathbb{R}^{d\times m}$ and bias $\mathbf{b} \in \mathbb{R}^{m}$, a Hadamard transform represented as $\mathbf{H} \in \mathbb{R}^{d\times d}$,

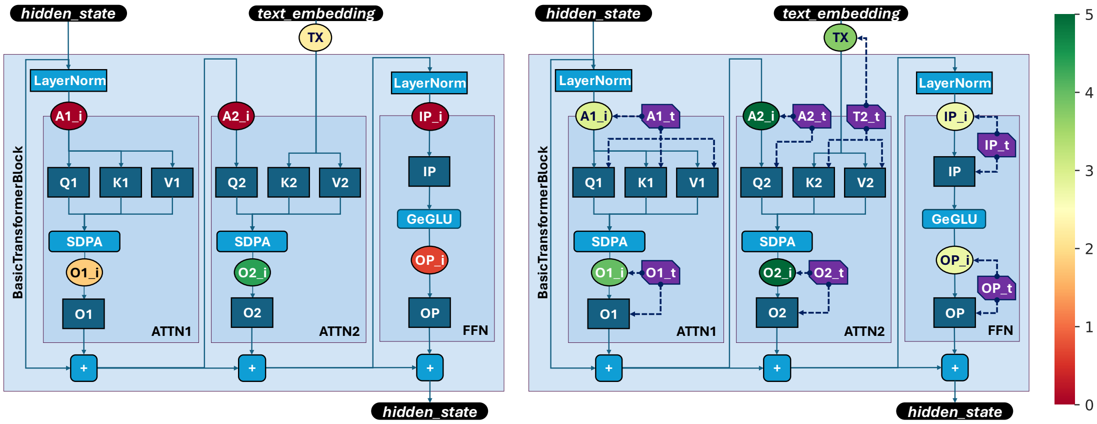

Figure 2. **HadaNorm reduces quantization error of all quantizers.** Effect of activation quantization for various components of the DiT architecture without (left) and with (right) HadaNorm transformations (indicated in purple). Activation Quantizers (circles) are colored by the corresponding impact on the SQNR. The dynamic centering is not applied on the textual quantizer.

and assume vectors $\boldsymbol{\mu}, \boldsymbol{\sigma} \in \mathbb{R}^d$. We can write:

$$\begin{aligned} \mathbf{XW} + \mathbf{b} &= (\mathbf{X}\mathrm{diag}\left(\boldsymbol{\sigma}^{-1}\right)\boldsymbol{H})(\boldsymbol{H}^T\mathrm{diag}\left(\boldsymbol{\sigma}\right)\mathbf{W}) + \mathbf{b} \\ &= (\mathbf{X}\mathrm{diag}\left(\boldsymbol{\sigma}^{-1}\right)\boldsymbol{H} - \boldsymbol{\mu} + \boldsymbol{\mu})(\boldsymbol{H}^T\mathrm{diag}\left(\boldsymbol{\sigma}\right)\mathbf{W}) + \mathbf{b} \\ &= \underbrace{(\mathbf{X}\mathrm{diag}\left(\boldsymbol{\sigma}^{-1}\right)\mathbf{H} - \boldsymbol{\mu})}_{\tilde{\mathbf{X}}}\underbrace{(\mathbf{H}^T\mathrm{diag}\left(\boldsymbol{\sigma}\right)\mathbf{W})}_{\tilde{\mathbf{W}}} + \underbrace{(\mathbf{b} + \boldsymbol{\mu}\tilde{\mathbf{W}})}_{\tilde{\mathbf{b}}}. \end{aligned} \tag{1}$$

Thus, we can maintain the original model output, but both $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{W}}$ have less outliers and will be quantized.

The per-channel mean $\boldsymbol{\mu}$ can be computed dynamically for each (transformed) batch element:

$$\boldsymbol{\mu} = \frac{1}{s}\mathbf{1}^T(\mathbf{X}\mathrm{diag}\left(\boldsymbol{\sigma}^{-1}\right)\boldsymbol{H}). \tag{2}$$

Note that the bias correction $\boldsymbol{\mu}\tilde{\mathbf{W}}$ can be computed efficiently in parallel, by appending a new token $\boldsymbol{\mu}$ to the transformed sequence $\tilde{\boldsymbol{X}}$ before performing the matrix multiplication with $\tilde{\boldsymbol{W}}$.

The scale $\boldsymbol{\sigma}$ cannot be chosen dynamically: this would require dynamically quantizing $\tilde{\mathbf{W}}$ depending on the current batch, which is undesirable. Instead, we follow Xiao et al. (2023), and statically determine $\boldsymbol{\sigma}$ based on the relative scales of $\mathbf{X}$'s channels and $\mathbf{W}$ input channels determined using a small calibration set:

$$\boldsymbol{\sigma}_i = \max(|\boldsymbol{X}_i|)^{\alpha} / \max(|\boldsymbol{W}_i|)^{1-\alpha}. \tag{3}$$

We place the HadaNorm layer throughout the network, before each linear layer (see Figure 2).

4. Experiments

Following Zhao et al. (2025); Shao et al. (2025); Li et al. (2024), we evaluate a quantized PixArt-Sigma (Chen et al., 2024) architecture on subset of captions from the COCO 2024 dataset (Lin et al., 2014) using 20 denoising steps. A disjoint calibration set is used to determine the activation statistics, which are used to tune the hyper-parameters α.

Quantization We quantize activations preceding each linear layer in the transformer blocks, which are indicated with ovals in Figure 2, at 4 bits precision (A4). Following Zhao et al. (2025), we dynamically compute a separate quantization grid for each token based on the minimum and maximum values. Weights in linear layers are also quantized at 4 bits in blocks of size 128.

Metrics We evaluate model performance using Signal to Quantized Noise Ratio (SQNR) computed in the latent space, the CLIP score (Hessel et al., 2021), and CLIP IQA (Wang et al., 2023) as a proxy of visual image quality based on CLIP features (Radford et al., 2021).

Baselines We compare HadaNorm against other quantization transformation strategies proposed in recent litterature including SmoothQuant (Xiao et al., 2023), QuaRot (Ashkboos et al., 2024), and the Static-Dynamic Channel Balancing (SDCB) method proposed in (Zhao et al., 2025), which combines HT with channel-wise scaling.

4.1. HadaNorm reduces outliers throughout the network

Set-up First we aim to isolate the source of quantization error, and see how HadaNorm may help. In all transformer blocks we switch off all quantizers except one activation quantizer (at 4 bit), measure the quantization error in terms of SQNR, and repeat this for all quantizers. Subsequently, we repeat this experiment, but with each transformation applied before each quantizer.

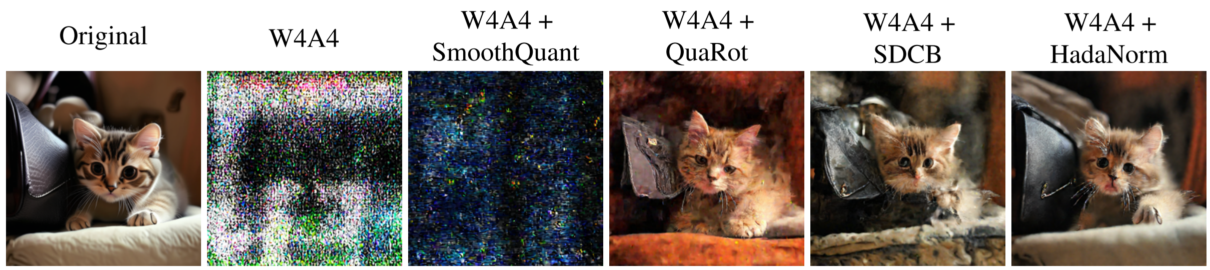

Figure 3. Visualization of the denoised images for the W4A4 quantized model starting from the same noise input and the prompt "An adorable cat attempts to hide in a purse to steal the persons identity".

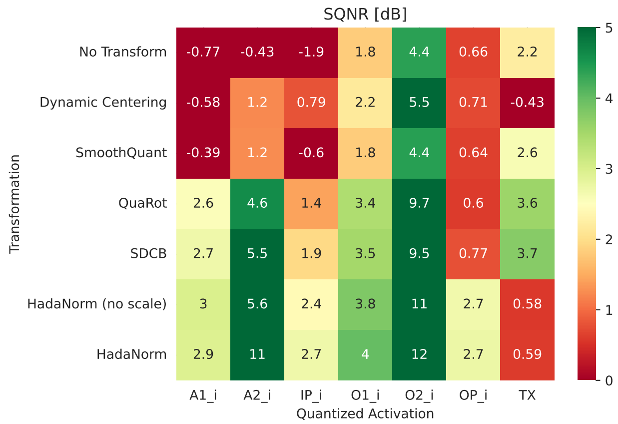

Figure 4. **HadaNorm's gain is mostly due to the combination of centering and the Hadamard transform.** Visualization of the SQNR resulting from the quantization of specific activations in the DiT architecture using various transformations.

Results. We see (Figure 2, left) that the largest quantization error originates from the quantization of image inputs to the attention and FFN blocks, and the `OP` layer. Adding HadaNorm (right) consistently improves SQNR with the exception of the Text quantizer (`TX`). Figure 4 reports the result of the same analysis for a range of transformations. We find that the source of improvements can not be explained by one individual transform, but rather by their composition. In particular, this study shows that centering significantly improves the performance of HT.

4.2. HadaNorm is SOTA at aggressive quantization

Set-up. In this experiment we compare the end-to-end performance of HadaNorm when quantizing Pixart-Sigma to W4A4. Transforms are applied as indicated in Figure 4 (right): each forward transform is applied online, while inverse channel scaling and HT are fused into the linear weights. Biases $\tilde{\boldsymbol{b}}$ are also computed dynamically following the expression in Equation 1 and Equation 2.

Results. We observe (Table 1) that quantization is hard: without any transforms, the negative SQNR indicates that the noise exceeds the signal. SmoothQuant and dynamic channel centering alone do not help much as they cannot reduce outliers by spreading them over multiple channels. HTs (QuaRot) results in a significant improvement by mitigating the effect of outliers, although SQNR is still poor. HadaNorm gives significant further improvements over SDCB (HT + channel scaling) thanks to the additional dynamic centering operation. The visual outputs corresponding to the results reported in the table are visualized in Figure 3. Additional results for 6-bits activation and weight quantization are reported in Appendix A.

Table 1. **HadaNorm outperforms all baselines.** Measure of SQNR [dB], CLIP score and CLIP IQA Pixart-Sigma models with 4 bits weights and activation quantization on a subset of the COCO 2024 dataset.

Transform	**SQNR** (↑)	**CLIP score**(↑)	**CLIP IQA**(↑)
Original	∞	31.66	0.90
No Transform	-2.88	19.22	0.11
SmoothQuant	-2.03	18.79	0.12
QuaRot	-0.39	30.88	0.76
SDCB	0.01	31.17	0.84
Dyn. Center	-2.32	19.68	0.14
HadaNorm	0.92	31.69	0.86

5. Conclusion

Although Hadamard transforms and channel scaling have been successfully used for improving quantization performance, we have shown that they are more effective when paired with a dynamic centering operation. The *HadaNorm* transform is a promising tool for more aggressive quantization, whilst being simple to implement and cheap to use during inference. Although this work has focused on the transformer blocks in diffusion models, future work may explore whether the HadaNorm transform provides the same benefit for quantizing other models (e.g. LLMs), and non-transformer layers (e.g. CNN blocks).

References

Ashkboos, S., Mohtashami, A., Croci, M. L., Li, B., Cameron, P., Jaggi, M., Alistarh, D., Hoefler, T., and Hensman, J. Quarot: Outlier-free 4-bit inference in rotated llms. In Globersons, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J. M., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/b5b939436789f76f08b9d0da5e81af7c-Abstract-Conference.html.

Chen, J., Ge, C., Xie, E., Wu, Y., Yao, L., Ren, X., Wang, Z., Luo, P., Lu, H., and Li, Z. Pixart-Σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. In Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., and Varol, G. (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XXXII*, volume 15090 of *Lecture Notes in Computer Science*, pp. 74–91. Springer, 2024. doi: 10.1007/978-3-031-73411-3_5. URL https://doi.org/10.1007/978-3-031-73411-3_5.

Hessel, J., Holtzman, A., Forbes, M., Bras, R. L., and Choi, Y. Clipscore: A reference-free evaluation metric for image captioning. In Moens, M., Huang, X., Specia, L., and Yih, S. W. (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pp. 7514–7528. Association for Computational Linguistics, 2021. doi: 10.18653/V1/2021.EMNLP-MAIN.595. URL https://doi.org/10.18653/v1/2021.emnlp-main.595.

Li, M., Lin, Y., Zhang, Z., Cai, T., Li, X., Guo, J., Xie, E., Meng, C., Zhu, J., and Han, S. Svdquant: Absorbing outliers by low-rank components for 4-bit diffusion models. *CoRR*, abs/2411.05007, 2024. doi: 10.48550/ARXIV.2411.05007. URL https://doi.org/10.48550/arXiv.2411.05007.

Lin, H., Xu, H., Wu, Y., Cui, J., Zhang, Y., Mou, L., Song, L., Sun, Z., and Wei, Y. Duquant: Distributing outliers via dual transformation makes stronger quantized llms. In Globersons, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J. M., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/9febda1c8344cc5f2d51713964864e93-Abstract-Conference.html.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In Fleet, D., Pajdla, T., Schiele, B., and Tuytelaars, T. (eds.), *Computer Vision – ECCV 2014*, pp. 740–755, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10602-1.

Liu, Z., Zhao, C., Fedorov, I., Soran, B., Choudhary, D., Krishnamoorthi, R., Chandra, V., Tian, Y., and Blankevoort, T. Spinquant: LLM quantization with learned rotations. *CoRR*, abs/2405.16406, 2024. doi: 10.48550/ARXIV.2405.16406. URL https://doi.org/10.48550/arXiv.2405.16406.

Liu, Z., Zhao, C., Fedorov, I., Soran, B., Choudhary, D., Krishnamoorthi, R., Chandra, V., Tian, Y., and Blankevoort, T. Spinquant: LLM quantization with learned rotations. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=ogO6DGE6FZ.

Ma, Y., Li, H., Zheng, X., Ling, F., Xiao, X., Wang, R., Wen, S., Chao, F., and Ji, R. Affinequant: Affine transformation quantization for large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=of2rhALq8l.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL http://proceedings.mlr.press/v139/radford21a.html.

Shao, W., Chen, M., Zhang, Z., Xu, P., Zhao, L., Li, Z., Zhang, K., Gao, P., Qiao, Y., and Luo, P. Omniquant: Omnidirectionally calibrated quantization for large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=8Wuvhh0LYW.

Shao, Y., Lin, D., Zeng, F., Yan, M., Zhang, M., Chen, S., Fan, Y., Yan, Z., Wang, H., Guo, J., Wang, Y., Qin, H., and Tang, H. TR-DQ: time-rotation diffusion quantization. *CoRR*, abs/2503.06564, 2025. doi: 10.48550/ARXIV.2503.06564. URL https://doi.org/10.48550/arXiv.2503.06564.

Wang, J., Chan, K. C. K., and Loy, C. C. Exploring CLIP for assessing the look and feel of images. In Williams, B., Chen, Y., and Neville, J. (eds.), *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pp. 2555–2563. AAAI Press, 2023. doi: 10.1609/AAAI.V37I2.25353. URL https://doi.org/10.1609/aaai.v37i2.25353.

Xiao, G., Lin, J., Seznec, M., Wu, H., Demouth, J., and Han, S. Smoothquant: Accurate and efficient post-training quantization for large language models. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 38087–38099. PMLR, 2023. URL https://proceedings.mlr.press/v202/xiao23c.html.

Zhao, T., Fang, T., Huang, H., Wan, R., Soedarmadji, W., Liu, E., Li, S., Lin, Z., Dai, G., Yan, S., Yang, H., Ning, X., and Wang, Y. Vidit-q: Efficient and accurate quantization of diffusion transformers for image and video generation. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=E1N1oxd63b.

A. Additional results

Table 2. Measure of SQNR, CLIP score and CLIP IQA Pixart-Sigma models with 6 bits weights and activation quantization on a subset of the COCO 2024 dataset.

Transform	**SQNR** (↑)	**CLIP score**(↑)	**CLIP IQA**(↑)
Original	∞	31.66	0.90
No Transform	0.5	32.39	0.92
SmoothQuant	1.46	31.95	0.91
QuaRot	2.32	31.85	0.91
SDCB	2.61	31.81	0.91
Dynamic Centering	1.56	31.87	0.91
HadaNorm	3.05	31.82	0.91

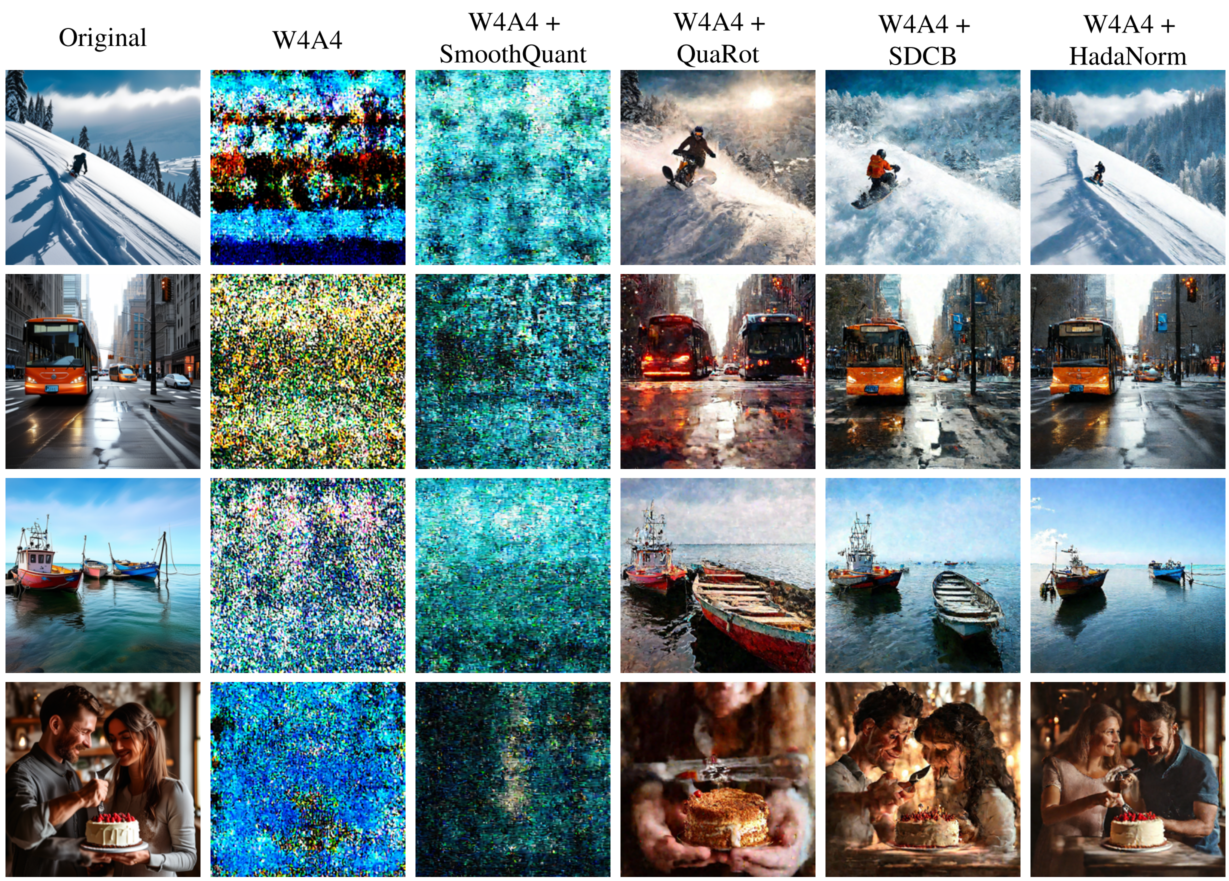

Figure 5. Additional mage generations for the W4A4 Pixart-Sigma model with several transforms and COCO prompts.

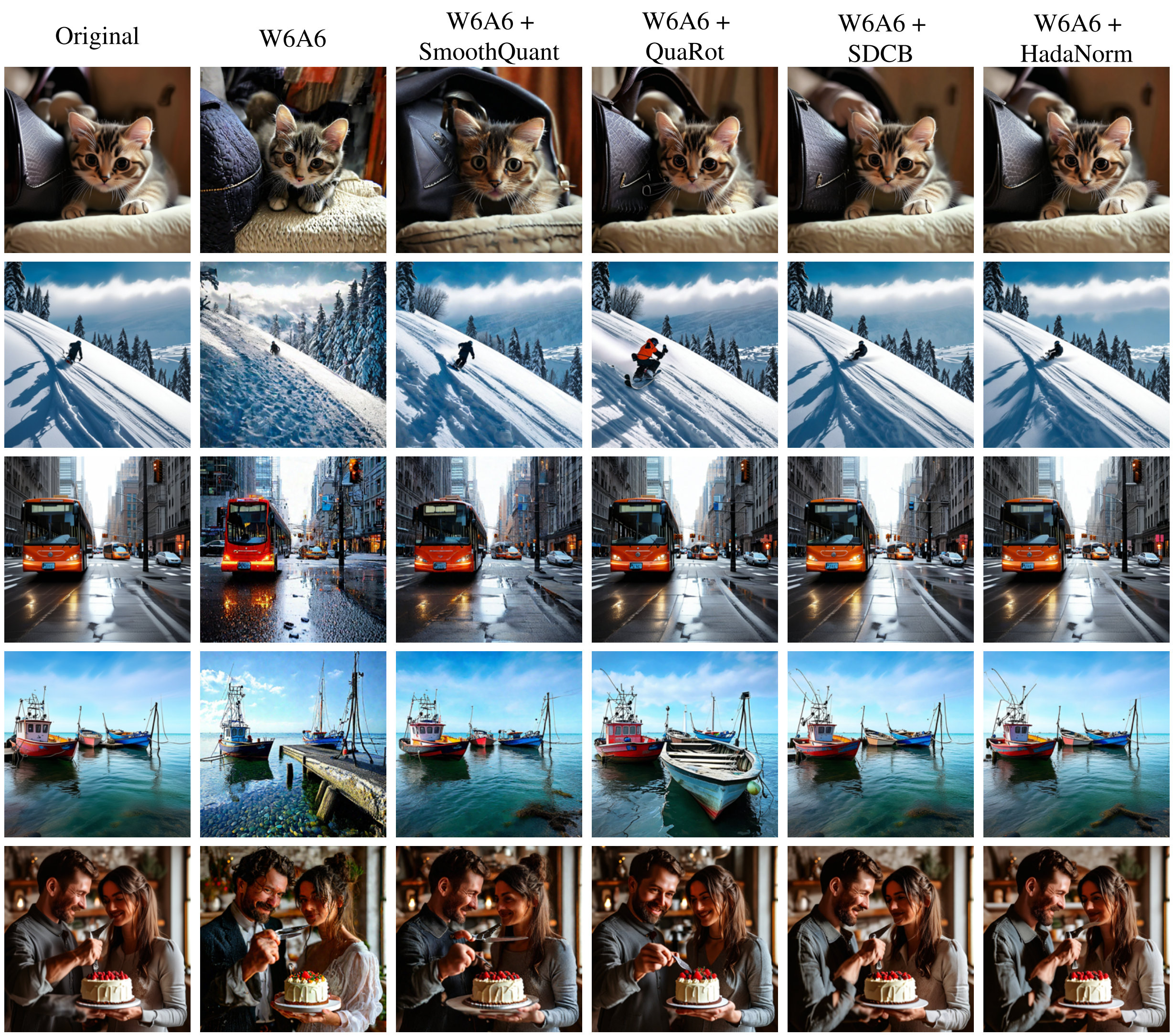

Figure 6. Image generations for the W6A6 Pixart-Sigma model with several transforms and COCO prompts.