DeltaQuant: 4-Bit Video Diffusion Models with Spatiotemporal Delta Smoothing

Supplementary Material

1. Additional Implementation Details

For the NVFP4 setting, we employ per-group symmetric quantization with a group size of 16 for both weights and activations. Each group's scale is represented in FP8 format (E4M3). For weights, we further introduce a per-channel scaling factor in FP16 to enhance representational diversity with negligible computational overhead.

Calibration is performed by sampling only 16–32 activations per layer across various timesteps and prompts. The entire calibration process requires approximately 24 hours on a single H100 GPU for all three models. Notably, our DeltaQuant incurs no additional calibration cost beyond that of SVDQuant [2].

The calibration procedure consists of two stages. In the first stage, we calibrate the smoothing factor $\boldsymbol{\lambda}$, defined as a vector where the smoothing factor for channel i is computed as $\max(|\boldsymbol{X}_{:,i}|)^{\alpha}/\max(|\boldsymbol{W}_{i,:}|)^{\beta}$. The hyperparameters α and β are searched over the interval $[0, 1]$ with a stride of 0.1, selecting the pair that minimizes the layer output mean-squared error (MSE) after SVD on the calibration set. In the second stage, we calibrate the low-rank decomposition for all layers, usingemploying an iterative approach to identify the decomposition that yields the lowest output MSE.

2. Visualization of the Generated Videos

We present a comparative visualization of the original 16-bit model, W4A4 SVDQuant [2], and our W4A4 DeltaQuant. Evaluations are conducted on Wan2.2 [3] for both image-to-video and text-to-video tasks, as well as on LTX-Video [1] for the text-to-video task. As shown in Figure 2, 3, 4, extensive visualizations demonstrate that DeltaQuant consistently preserves the high pixel-level fidelity and overall visual quality of the 16-bit model, outperforming the W4A4 SVDQuant [2] baseline. Additional visual results, including videos and static web-based comparisons, are provided in the supplementary materials.

Importantly, we observe that in the image-to-video task, the first generated frame is particularly sensitive to quantization errors. This is because the input frame directly serves as the initial frame of the generated video, providing crucial guidance for subsequent generation. As shown in Table 1 and Figure 1, the initial frame generated by SVDQuant [2] has low similarity to ground truth, and suffers from noticeable distortions. In contrast, DeltaQuant leverages intrinsic spatiotemporal similarity in the activations, reliably generating high-quality initial frames that maintain strong similarity and fidelity to the input.

Table 1. Quantitative comparison of first frame quality. Our DeltaQuant outperforms SVDQuant in similarity and Image Reward (I.R.), achieving quality on par with the ground truth.

Method	LPIPS (↓)	PSNR (↑)	SSIM (↑)	I.R. (↑)
Ground Truth	–	–	–	1.025
SVDQuant	0.126	27.7	0.816	0.981
Ours	**0.058**	**30.1**	**0.903**	**1.021**

3. Theoretical Analasis

Following the main paper notations, under 4-bit quantization, the activation error is bounded by the activation norm, $\|Q_4(R(c)) - R(c)\|_F^2 \leq \gamma_4 \|R(c)\|_F^2$. The residual norm $\|R(c)\|_F^2 = \sum_{j,i}(X_{ij}^{(k)} - c_j)^2$ is minimized by the per-channel mean $c = \bar{X}^{(k)}$, yielding $\tilde{X}^{(k)}$. With in-cube similarity $\|\tilde{X}^{(k)}\|_F^2 \leq m\alpha D_k^2$, the quantization error is bounded by $\gamma_4 m\alpha D_k^2$ (smaller cubes are better), while the FP8 term is negligible since $\gamma_8 \ll \gamma_4$.

4. Kernel Efficiency Breakdown

Our nsys profiling on Wan2.2-I2V shows that FP8 core-token compute adds 6.9% overhead to the DeltaQuant kernel (Table 2), while unfused mean-delta incurs 105% overhead. Overall, DeltaQuant achieves $5.5\times$ speedup over BF16. Spatiotemporal reordering overhead is negligible, as only two reorders are required per inference due to the use of two cube sizes.

References

[1] Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, Poriya Panet, Sapir Weissbuch, Victor Kulikov, Yaki Bitterman, Zeev Melumian, and Ofir Bibi. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024. 1

[2] Muyang Li, Yujun Lin, Zhekai Zhang, Tianle Cai, Junxian Guo, Xiuyu Li, Enze Xie, Chenlin Meng, Jun-Yan Zhu, and Song Han. Svdquant: Absorbing outliers by low-rank component for 4-bit diffusion models. In *ICLR*, 2025. 1

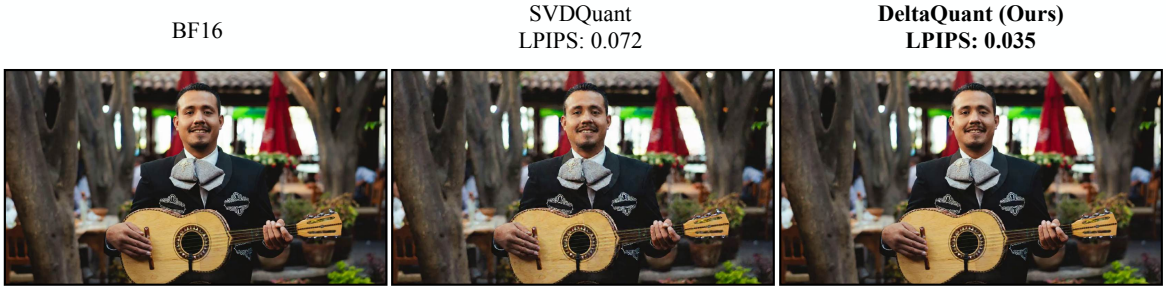
[3] Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang,

Operator	W4A8 Latency (s)	W4A8 TFLOPs	W4A4 Latency (s)	W4A4 TFLOPs	Quant. Latency (s)	W4A8 Percentage	DeltaQuant Latency (s)	BF16 Latency (s)	Speedup	Mean-Delta Overhead If not fused
QKV Proj.	0.998	554	12.07	1314	2.25	**6.5%**	15.32	79.3	**5.2×**	**14.54 (+95%)**
O Proj.	0.598	523	6.05	1313	2.26	**6.7%**	8.91	44.2	**5.0×**	**10.08 (+113%)**
FFN up	0.650	576	10.3	1039	1.12	**5.3%**	12.07	61.2	**5.1×**	**8.87 (+70.5%)**
FFN down	0.790	522	7.06	1516	0	**10.1%**	7.85	58.8	**7.5×**	**12.72 (+152%)**
Total	**3.04**	**544**	**35.48**	**1274**	**5.63**	**6.9%**	**44.15**	**243.5**	**5.5×**	**46.21 (+105%)**

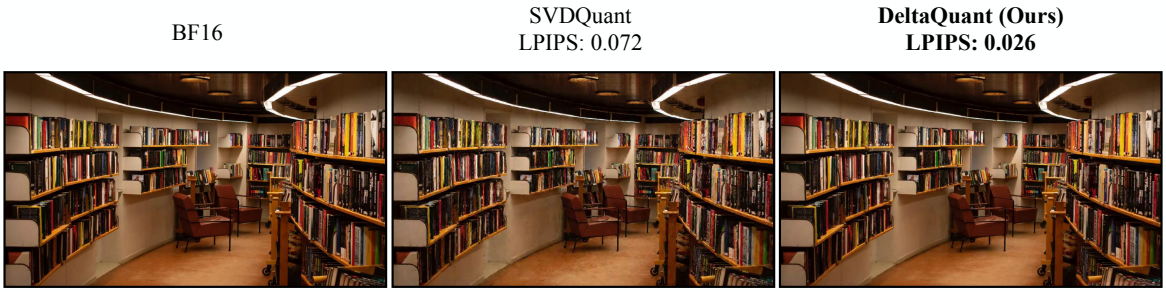
Table 2. Per-kernel latency/throughput breakdown.

Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 1

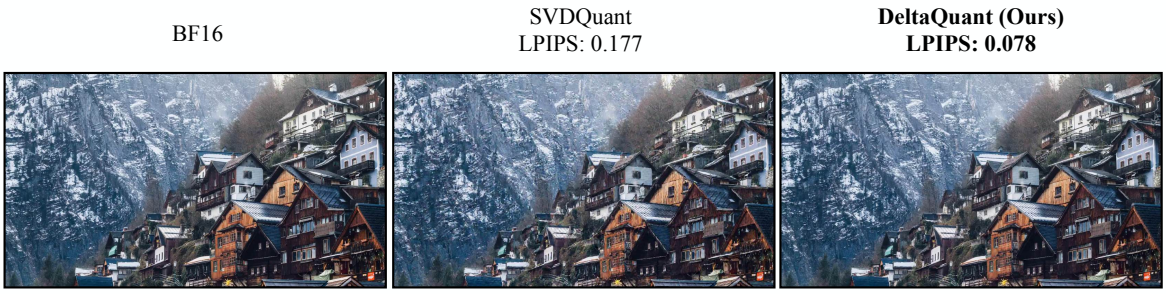
Prompt: *an aerial view of big ben and the houses of parliament in london, camera static.*



Prompt: *a man in a mexican outfit holding an acoustic guitar.*



Prompt: *a room filled with lots of shelves filled with books, camera pans right.*



Prompt: *a bunch of houses that are on a hillside, camera pans right.*

Figure 1. Visual comparison of the first generated frame on Wan2.2-I2V among the Ground Truth, SVDQuant, and DeltaQuant. Our method produces frames with significantly improved similarity and visual quality compared to SVDQuant, which exhibits noticeable distortions.

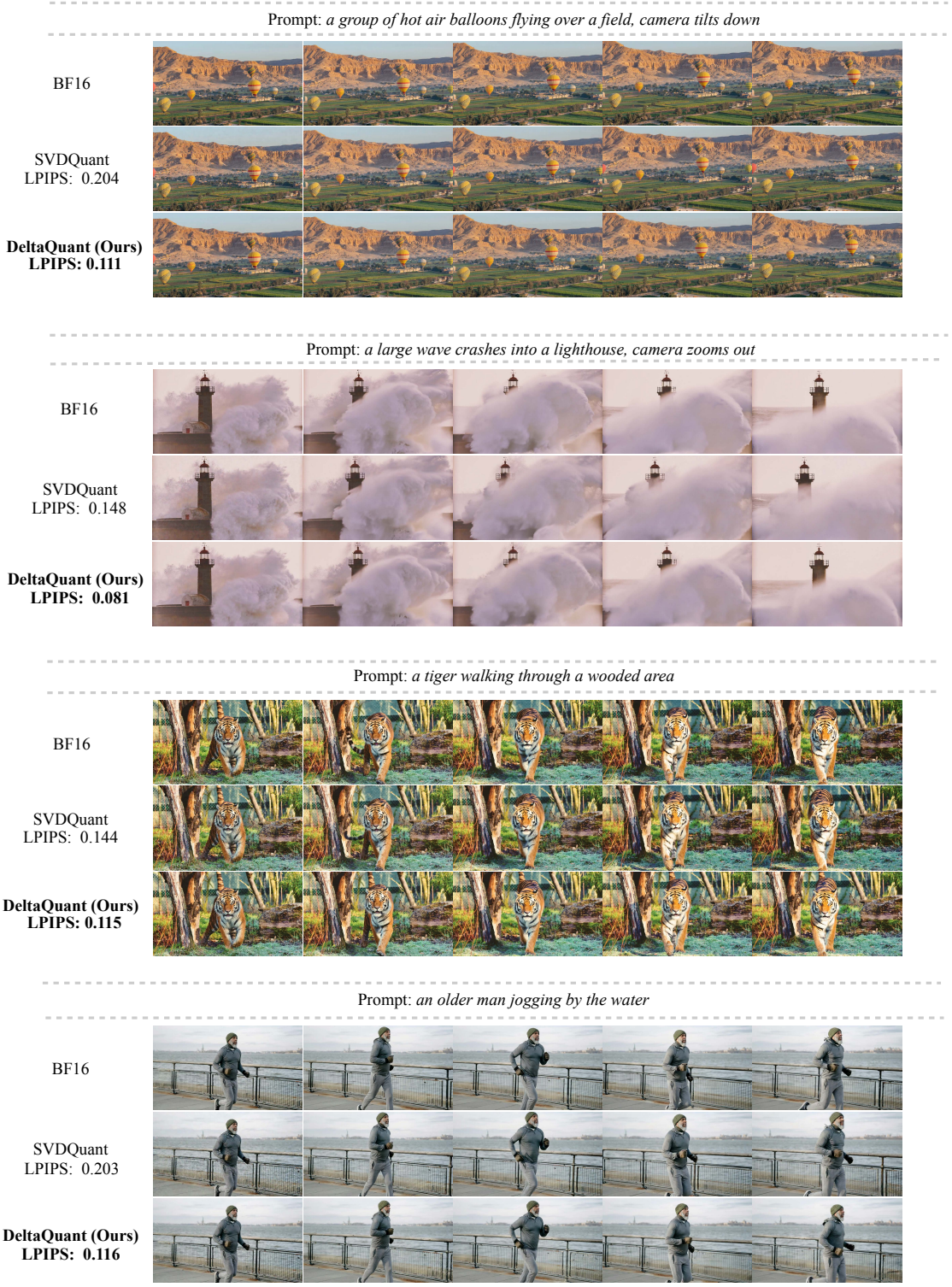


Figure 2. Comparison of Ground Truth, SVDQuant, and DeltaQuant on Wan 2.2 image-to-video generation.

Prompt: *the setting features a cinematic ambiance, with muted dusk tones lending an elegant, timeless quality to the scene. the view captures a graceful woman in her early 30s, singing under the warm glow of streetlights…*

BF16

SVDQuant
LPIPS: 0.310

DeltaQuant (Ours)
LPIPS: 0.207

Prompt: *soft and warm lighting sets the scene in a beautifully designed living room, where the natural stone fireplace, crafted from Charlotte stone, serves as the centerpiece. the flames crackle warmly, casting a cozy…*

BF16

SVDQuant
LPIPS: 0.363

DeltaQuant (Ours)
LPIPS: 0.340

Prompt: *the color palette is realistic, with lighting that is soft and warm, giving a natural, late afternoon feel. the shot captures a clean-shaven man approximately 55 years old, standing determinedly on a bustling city street…*

BF16

SVDQuant
LPIPS: 0.304

DeltaQuant (Ours)
LPIPS: 0.277

Prompt: *vibrant and colorful, with a dynamic, abstract background full of swirling patterns. a young asian girl, wearing glasses and a subtle smile, is the focus of the close-up shot. her slightly wavy hair frames her face softly….*

BF16

SVDQuant
LPIPS: 0.440

DeltaQuant (Ours)
LPIPS: 0.395

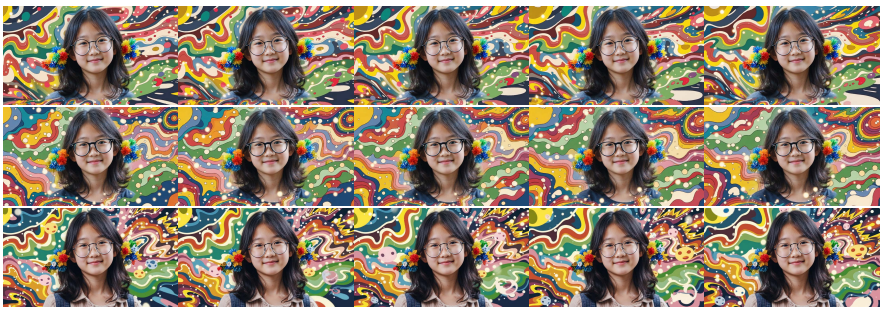


Figure 3. Comparison of Ground Truth, SVDQuant, and DeltaQuant on Wan 2.2 text-to-video generation.

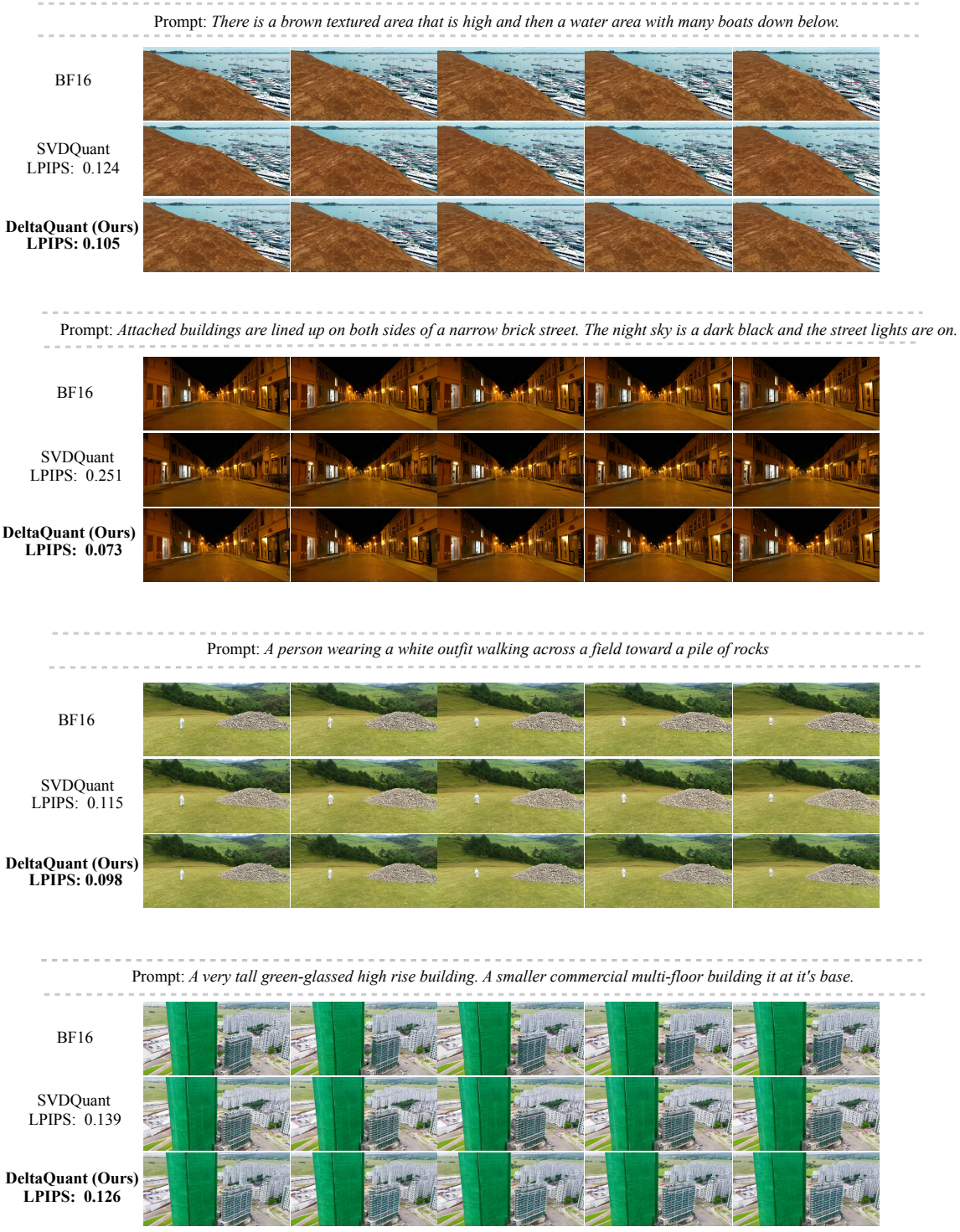


Figure 4. Comparison of Ground Truth, SVDQuant, and DeltaQuant on LTXVideo text-to-video generation.