PULSEQUANT: PROPAGATION-GUIDED SUBSPACE CORRECTION FOR 4-BIT VIDEO DIFFUSION TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Quantization errors in video diffusion transformers can be amplified or attenuated by subsequent denoising updates, making local reconstruction error an incomplete predictor of final impact. We introduce PulseQuant, a 4-bit post-training quantization method that combines trajectory sensitivity with activation geometry to guide offline calibration. Isolated block–step interventions estimate propagation risk, which prioritizes sensitive trajectory states during row-radius selection. With these radii fixed, response-subspace correction uses neighboring-code edits to reduce residual components along dominant activation directions. Both stages preserve the original 4-bit weight representation. Controlled interventions show that short-horizon propagated error predicts final latent error more reliably than immediate block-output error, supporting calibration beyond local reconstruction objectives. Evaluations on Wan models, Self Forcing, and MiniMax-H3 demonstrate improvements in key consistency and dense-reference metrics while remaining competitive on other attributes across model scales and generation paradigms.

1 INTRODUCTION

Video diffusion transformers (Peebles & Xie, 2023; Wan et al., 2025; Huang et al., 2026a; MiniMax Research, 2026) repeatedly execute a large network throughout denoising, making post-training quantization (PTQ) attractive for reducing storage and computation without retraining. Yet the same iterative process can amplify quantization errors: a small local disturbance may develop into spatial distortion or temporal inconsistency after subsequent updates. Video PTQ must therefore preserve the denoising trajectory, not just the accuracy of isolated operators.

Existing methods adapt to timestep-varying distributions, reshape tensors before quantization, or preserve difficult structure in higher precision (Li et al., 2023; He et al., 2023; Zhao et al., 2024; Xiao et al., 2023; Li et al., 2024). These strategies improve local quantization behavior, but local reconstruction quality alone does not establish how an error will affect the remaining denoising trajectory. Subsequent updates may attenuate or amplify similar local disturbances differently, creating a ***local-to-global misalignment*** between immediate error and final impact.

To examine this mismatch, we perform 80 isolated 4-bit block–step interventions on Wan 2.1-14B, each followed by dense continuation. Figure 1 shows that local block-output RMSE is nearly uncorrelated with final latent error ($\rho = -0.08$), whereas four-step propagated error better recovers its within-timestep ranking ($\rho = 0.70$); final error also varies strongly with trajectory position. This evidence motivates weighting calibration by downstream sensitivity. Within the fixed 4-bit representation, calibration must then choose code changes that effectively reduce response error, which depends on how weight residuals align with the activation distribution.

We introduce PulseQuant, a propagation-guided subspace correction method for 4-bit video diffusion transformers. As illustrated in Figure 2, ***propagation-aware calibration*** uses isolated block–step interventions to prioritize sensitive trajectory states during row-radius selection. With these radii fixed, ***response-subspace correction*** applies legal neighboring-code edits to reduce residual components along dominant activation directions. Both stages operate offline and preserve the existing 4-bit weight representation, without high-precision residual branches or online correction. Across model scales, generation paradigms, and activation precisions, PulseQuant improves key

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

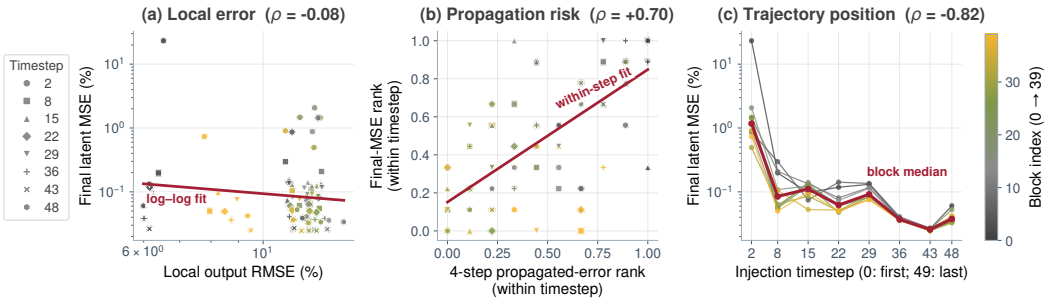


Figure 1: Propagation diagnostics from 80 isolated 4-bit interventions on Wan 2.1-14B (10 blocks $\times$ eight denoising steps), each followed by dense continuation. (a) Local block-output RMSE is weakly associated with final latent error. (b) Four-step propagated error E_H predicts final damage after ranking blocks within each timestep. (c) Final damage depends strongly on trajectory position. Colors encode block depth and markers encode injection timestep.

consistency and dense-reference metrics, leading all paired fidelity metrics at W4A4 among the tested PTQ baselines. For Wan 2.1-1.3B on an RTX 5080, packed NVFP4 more than halves DiT peak memory; combining it with low-bit attention (Li et al., 2026b; Zhang et al., 2024) yields up to a 3.51$\times$ DiT-step speedup over Dense BF16.

Our contributions are threefold:

- We quantify the local-to-global misalignment in video PTQ through isolated 4-bit block–step interventions. Short-horizon propagated error predicts final latent error more reliably than immediate reconstruction error, motivating calibration that accounts for downstream sensitivity across trajectory positions.
- We develop an offline calibration procedure that couples propagation-weighted radius selection with response-subspace code correction. The former prioritizes sensitive trajectory states, while the latter uses neighboring-code edits to reduce response error along dominant activation directions at fixed radii.
- We demonstrate gains in key consistency and dense-reference metrics across bidirectional and autoregressive video generators, model scales, and activation precisions. Native packed NVFP4 deployment on an RTX 5080 more than halves DiT peak memory and complements low-bit attention, delivering up to a 3.51$\times$ DiT-step speedup over Dense BF16.

2 Related Work

2.1 Quantization for diffusion and video transformers

Diffusion PTQ must accommodate activation distributions that change throughout denoising. Diffusion-specific calibration and error correction address this nonstationarity (Shang et al., 2023; Li et al., 2023; He et al., 2023), while timestep-aware quantization and reconstruction explicitly account for temporal variation (So et al., 2023; Huang et al., 2024a; Wang et al., 2024). For image and video transformers, this challenge extends to variation across spatial positions, channels, and tokens, motivating finer-grained quantization and calibration strategies (Chen et al., 2025; Zhao et al., 2024; Feng et al., 2025; Li et al., 2026c). Spatiotemporal redundancy offers another compression opportunity, exploited by DeltaQuant (Li et al., 2026a). Beyond PTQ, QVGen (Huang et al., 2026b) uses quantization-aware training with auxiliary correction modules that are progressively removed through rank decay. Alongside improving the low-bit approximation, calibration must account for how its errors evolve during iterative generation.

Error accumulation across denoising steps is an established concern (Li et al., 2023). AccuQuant (Lee et al., 2025) addresses it directly by matching full-precision and quantized outputs over multiple simulated denoising steps. PulseQuant instead uses isolated block–timestep pulses

followed by dense continuation to estimate downstream sensitivity and incorporates these measurements as weights in row-radius calibration.

2.2 Low-bit representations and response correction

Low-bit PTQ improves quantization quality by reshaping the representation or optimizing discrete reconstruction decisions. Channel scaling and learnable equivalent transformations redistribute quantization difficulty while preserving the dense function (Xiao et al., 2023; Shao et al., 2024), with AWQ (Lin et al., 2024) using activation-informed scaling to protect salient weight channels. Fixed or learned rotations further reduce outlier concentration (Ashkboos et al., 2024; Liu et al., 2025). Incoherence processing and structured codebooks support accurate low-bit representations (Tseng et al., 2024), while AQLM (Egiazarian et al., 2024) learns additive codebooks with input-adaptive reconstruction and block-wise joint optimization. For visual diffusion transformers, OrbitQuant combines normalization, RPBH rotation, and a shared spherical codebook (Lee et al., 2026).

Within a chosen representation, reconstruction-based methods optimize rounding or code decisions using weight-, layer-, or block-local objectives (Nagel et al., 2020; Li et al., 2021; Frantar et al., 2022). SVDQuant instead retains difficult structure in a high-precision low-rank branch (Li et al., 2024). Building on the OrbitQuant representation, PulseQuant uses activation-adapted response directions to guide neighboring-code edits at fixed calibrated row radii. The subspace guides offline index selection, while the exported weights retain the shared codebook and per-row scales.

3 Preliminaries

3.1 Quantized linear projections

Consider a linear layer $\mathbf{y} = \mathbf{W}\mathbf{x}$, where $\mathbf{x} \in \mathbb{R}^d$, $\mathbf{y} \in \mathbb{R}^m$, and $\mathbf{W} \in \mathbb{R}^{m\times d}$. We adopt the channel transform of OrbitQuant (Lee et al., 2026): a positive diagonal scaling $\mathbf{D}$ followed by an orthogonal randomized permuted block-Hadamard (RPBH) rotation $\mathbf{\Pi}_d$. Transforming the activations and applying the corresponding inverse transform to the weights gives

$$\mathbf{x}' = \mathbf{\Pi}_d\mathbf{D}\mathbf{x}, \qquad \mathbf{W}' = \mathbf{W}\mathbf{D}^{-1}\mathbf{\Pi}_d^\top, \qquad \mathbf{W}'\mathbf{x}' = \mathbf{W}\mathbf{x}. \tag{1}$$

These transforms reshape the quantization coordinates while preserving the dense layer output. For a token batch $\mathbf{X} \in \mathbb{R}^{N\times d}$ stored row-wise, the activation transform is $\mathbf{X}' = \mathbf{X}\mathbf{D}\mathbf{\Pi}_d^\top$. The low-bit layer computes

$$\widehat{\mathbf{y}} = \widehat{\mathbf{W}}'Q_A(\mathbf{x}'), \tag{2}$$

where Q_A denotes symmetric per-token activation quantization, expressed in dequantized values, and $\widehat{\mathbf{W}}'$ denotes the quantized weight matrix represented by 4-bit indices and per-row scales. Appendix B specifies the activation quantizer.

3.2 Spherical weight coding

OrbitQuant (Lee et al., 2026) normalizes each transformed weight row $\mathbf{w}'_i$ by its initial radius $r_i = \|\mathbf{w}'_i\|_2$. Each normalized coordinate is then assigned to its nearest entry in a shared, ordered 16-value codebook $\mathcal{C} = \{C_0, \ldots, C_{15}\}$:

$$c_{ij} = \arg\min_{k\in\{0,\ldots,15\}} |w'_{ij}/r_i - C_k|, \qquad q_{ij} = r_i C_{c_{ij}}. \tag{3}$$

Each row is therefore stored as one radius, serving as its reconstruction scale, and one 4-bit index per coordinate. PulseQuant calibrates the row radii and refines the discrete indices while keeping the codebook fixed; the resulting weights are $\widehat{w}'_{ij} = \bar{r}_i C_{c_{ij}}$.

4 Method

4.1 Overview

PulseQuant calibrates row radii and discrete weight indices in two offline stages, as illustrated in Figure 2. *Propagation-aware calibration* first profiles isolated 4-bit perturbations and uses their

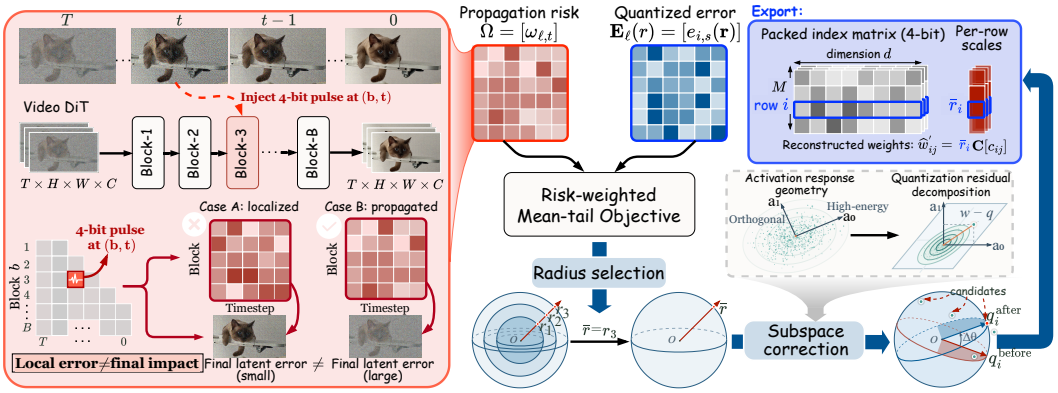


Figure 2: Overview of PulseQuant. Isolated 4-bit pulses reveal how local perturbations propagate through denoising (left). The resulting risk weights guide per-row radius selection through a mean–tail output-error objective (center). With the selected radii fixed, legal neighboring-code edits reduce residual components along high-energy activation directions (right). The calibrated weights are exported as packed 4-bit indices and per-row scales, with no online profiling or correction.

gains to weight row-radius selection. *Response-subspace correction* then holds the selected radii fixed and applies legal neighboring-code edits to reduce residual components along dominant activation directions. The resulting indices and row radii define the quantized weights used in Eq. 2. Algorithm 1 gives the complete procedure; Figure 7 in Appendix C.1 visualizes activation balancing, propagation risk, and residual reduction.

4.2 Propagation-aware calibration

Profiling trajectory sensitivity. We estimate downstream sensitivity with an isolated intervention at each sampled block–timestep anchor (b, t). Starting from the same latent state, we compare a dense trajectory with a pulsed trajectory that quantizes block b only at step t; both then continue in dense precision for H updates. The profiling pulse uses uniform W4A4 quantization–dequantization on the attention and FFN linear projections of the selected block. For CFG-based profiling, it covers both forward calls at the anchor step; Appendix B.1 details the pulse configuration. Let $\mathbf{z}^{\mathrm{p}}$ and $\mathbf{z}^{\mathrm{d}}$ denote their latent states. We measure the normalized discrepancy immediately after the pulse and after H further updates, and define the propagation gain from their relative change:

$$E_h(b,t) = \frac{\|\mathbf{z}^{\mathrm{p}}_{t+1+h} - \mathbf{z}^{\mathrm{d}}_{t+1+h}\|_2^2}{\|\mathbf{z}^{\mathrm{d}}_{t+1+h}\|_2^2 + \epsilon}, \quad h \in \{0, H\}, \qquad g_{b,t} = \sqrt{\frac{E_H(b,t) + \epsilon}{E_0(b,t) + \epsilon}}. \tag{4}$$

The gain measures the change in normalized latent discrepancy over H subsequent updates: $g_{b,t} > 1$ indicates amplification, whereas $g_{b,t} < 1$ indicates attenuation. We obtain a trajectory-wide profile by interpolating positive gains from a sparse anchor grid in log space over the scheduler noise coordinate and relative block depth. Each layer ℓ inherits the interpolated profile $\bar{g}_{\ell,t}$ of its block. We temper and clip the profile, then normalize the weights to have unit mean over all block–timestep locations:

$$\omega_{\ell,t} = \mathrm{MeanOne}\Big(\mathrm{clip}(\bar{g}^{\gamma}_{\ell,t}, a, b)\Big). \tag{5}$$

Here, γ controls the contrast between locations, $[a, b]$ limits the influence of extreme estimates, and mean-one normalization preserves the overall objective scale. The resulting $\omega_{\ell,t}$ assigns greater calibration weight to perturbations that are more strongly amplified downstream. Appendix A relates this finite-horizon measurement to first-order error propagation.

Risk-weighted radius selection. The propagation profile weights the contribution of each trajectory state to radius calibration. For layer ℓ, let $\mathbf{X}'_{\ell,s} \in \mathbb{R}^{N_s \times d}$ contain the transformed tokens recorded at calibration call s, and let $t(s)$ be the corresponding denoising step. For row i, the base indices define the fixed codebook direction $\mathbf{v}_i = [C_{c_{i1}}, \ldots, C_{c_{id}}]^\top$, giving the reconstruction $r\mathbf{v}_i$ at

radius r. For each candidate $r \in \mathcal{R}_i$, we compute the mean squared row-output error $e_{i,s}(r)$ and weight it by the propagation risk at that call:

$$\begin{aligned} e_{i,s}(r) &= \frac{1}{N_s}\|\mathbf{X}'_{\ell,s}(\mathbf{w}'_i - r\mathbf{v}_i)\|_2^2, \qquad u_{i,s}(r) = \omega_{\ell,t(s)} e_{i,s}(r), \\ \bar{r}_i &= \arg\min_{r\in\mathcal{R}_i} \left[(1-\lambda)\,\mathrm{Mean}_s\, u_{i,s}(r) + \lambda\,\mathrm{TopMean}_{\rho,s}\, u_{i,s}(r)\right]. \end{aligned} \tag{6}$$

Here, $\mathrm{TopMean}_\rho$ averages the largest ρ fraction of the risk-weighted call errors, and λ balances the mean and tail terms. The mean term controls average response distortion, while the tail term emphasizes calls with the largest risk-weighted errors. Propagation weighting and tail aggregation serve distinct roles: the former accounts for downstream sensitivity, while the latter prioritizes calls that remain difficult after this weighting. Because each weight row is reused throughout denoising, the objective selects a single radius across the sampled trajectory states. We hold this radius $\bar{r}_i$ fixed during the subsequent code correction.

4.3 Response-subspace correction

With the row radii calibrated, we refine the discrete indices using the activation geometry. Weight residuals affect the layer output through their interaction with activations: a residual aligned with a high-energy activation direction produces a larger mean squared output error than an equally sized residual along a low-energy direction. This motivates focusing code correction on a compact set of high-energy response directions. For layer ℓ, we retain an evenly spaced subset $\bar{\mathbf{X}}'_\ell \in \mathbb{R}^{M_\ell \times d}$ and define its uncentered second moment $\mathbf{C}_\ell = \bar{\mathbf{X}}'^{\top}_\ell \bar{\mathbf{X}}'_\ell / M_\ell$. The quantity $\mathbf{a}^\top \mathbf{C}_\ell \mathbf{a}$ measures activation energy along direction $\mathbf{a}$. We estimate two high-energy directions with a short orthogonal power iteration, initialized by the all-ones and alternating-sign Walsh patterns:

$$\begin{aligned} \mathbf{a}_0 &\leftarrow \mathrm{norm}(\mathbf{C}_\ell \mathbf{a}_0), & \mathbf{a}_0^{(0)} &\propto [1,1,\ldots], \\ \mathbf{a}_1 &\leftarrow \mathrm{norm}\big((\mathbf{I} - \mathbf{a}_0\mathbf{a}_0^\top)\mathbf{C}_\ell \mathbf{a}_1\big), & \mathbf{a}_1^{(0)} &\propto [1,-1,\ldots]. \end{aligned} \tag{7}$$

Multiplication by $\mathbf{C}_\ell$ adapts each seed toward a high-energy activation direction, while the second update orthogonalizes the second axis against the first. We use $M_{\max} = 512$ tokens and two iterations, evaluating each matrix product as $\bar{\mathbf{X}}'^{\top}_\ell(\bar{\mathbf{X}}'_\ell \mathbf{a})/M_\ell$ without materializing $\mathbf{C}_\ell$.

The resulting axes guide discrete correction at the selected radius $\bar{r}_i$. We partition row i into channel groups $\mathcal{G}$ of size G; within a group, let $\mathbf{w}$ and $\mathbf{q}$ denote the corresponding slices of the dense transformed row $\mathbf{w}'_i$ and its current quantized reconstruction, and restrict each $\mathbf{a}_u$ in the same way. We minimize

$$\mathcal{L}_{i,\mathcal{G}}(\mathbf{q}) = \underbrace{\frac{1}{2}\sum_{u=0}^{1} \frac{\langle \mathbf{w}-\mathbf{q}, \mathbf{a}_u\rangle^2}{\|\mathbf{a}_u\|_2^2 + \epsilon}}_{\text{subspace response}} + \underbrace{\frac{\tau}{G}\|\mathbf{w}-\mathbf{q}\|_2^2}_{\text{full-group distortion}}. \tag{8}$$

Here, $\mathbf{w} - \mathbf{q}$ is the group-wise residual between the dense transformed row and its current 4-bit reconstruction. The first term averages its squared projections onto the two response axes, normalized by their squared norms with ϵ for stability. The second term controls mean squared distortion across the full group, with weight τ. Each legal neighboring-code edit replaces an index with an adjacent codebook index at the fixed row scale. From the same current reconstruction, we evaluate both one-level moves for each coordinate and retain the better proposal. We rank these proposals by objective decrease and select the cumulative prefix with the lowest joint objective, including the empty prefix. Each prefix is evaluated jointly to account for interactions among coordinate edits. Appendix B.2 reports the default settings, while Appendix A.3 provides the exact updates, response-error bound, and prefix-selection guarantee.

Table 1: **Quantitative evaluations on VBench (T2V).** VBench attributes are percentages; PSNR, SSIM, and LPIPS compare matched quantized and Dense BF16 videos. We bold the **best results** and underline the second-best results.

Model	Method	W/A	VBench attributes (%)							Dense-reference fidelity		
			IQ ↑	AQ ↑	MS ↑	BC ↑	SC ↑	Scene ↑	OC ↑	PSNR ↑	SSIM ↑	LPIPS ↓
Wan 2.1-1.3B	Dense BF16	16/16	67.71	66.06	98.29	96.19	93.97	42.44	26.15	ref.	ref.	ref.
	SmoothQuant (Xiao et al., 2023)	4/6	59.83	58.12	98.12	94.77	91.14	30.96	24.20	14.78	0.464	0.499
	SVDQuant (Li et al., 2024)	4/6	63.68	63.57	97.83	95.20	92.62	35.61	25.42	15.36	0.513	0.433
	ViDiT-Q (Zhao et al., 2024)	4/6	63.60	63.68	97.78	95.03	92.70	38.08	25.73	14.86	0.503	0.447
	DVD-Quant (Li et al., 2026c)	4/6	63.21	64.04	**98.62**	95.23	94.38	41.86	25.60	14.96	**0.529**	0.430
	OrbitQuant (Lee et al., 2026)	4/6	63.78	64.72	98.54	95.37	**94.66**	42.01	25.57	15.30	0.521	0.424
	PulseQuant (ours)	4/6	**64.96**	**64.79**	98.21	**95.44**	93.10	**43.97**	**25.91**	**15.64**	0.527	**0.411**
	SmoothQuant (Xiao et al., 2023)	4/4	55.24	47.60	94.35	94.56	85.76	11.77	19.76	14.38	0.392	0.561
	SVDQuant (Li et al., 2024)	4/4	62.36	59.61	96.94	94.76	90.19	20.09	24.32	15.27	0.479	0.465
	ViDiT-Q (Zhao et al., 2024)	4/4	58.02	50.23	95.23	94.59	86.03	15.19	20.99	14.30	0.401	0.547
	DVD-Quant (Li et al., 2026c)	4/4	51.18	52.74	**98.28**	94.24	89.88	13.30	21.66	12.96	0.392	0.658
	OrbitQuant (Lee et al., 2026)	4/4	**63.14**	**63.33**	98.10	94.69	**93.76**	34.23	25.17	14.97	0.491	0.458
	PulseQuant (ours)	4/4	62.70	62.46	97.83	**94.84**	91.97	**36.70**	**25.44**	**15.42**	**0.503**	**0.443**
Self Forcing	Dense BF16	16/16	70.88	67.33	98.40	95.53	94.77	47.31	26.42	ref.	ref.	ref.
	SmoothQuant (Xiao et al., 2023)	4/6	70.40	66.15	98.12	94.80	94.01	46.29	26.21	13.71	0.402	0.464
	SVDQuant (Li et al., 2024)	4/6	70.33	66.05	98.50	95.33	95.04	**50.07**	26.17	13.81	0.401	0.469
	ViDiT-Q (Zhao et al., 2024)	4/6	70.40	65.97	97.94	94.91	93.60	43.53	26.20	13.51	0.392	0.495
	DVD-Quant (Li et al., 2026c)	4/6	**70.64**	66.02	98.52	95.60	**95.37**	47.97	26.06	13.69	0.409	0.457
	OrbitQuant (Lee et al., 2026)	4/6	70.08	**67.02**	98.40	95.51	94.57	49.49	26.26	14.02	0.408	0.454
	PulseQuant (ours)	4/6	70.46	66.74	**98.77**	**95.99**	95.11	45.64	**26.28**	**14.17**	**0.428**	**0.443**
	SmoothQuant (Xiao et al., 2023)	4/4	69.35	65.72	98.04	94.34	93.49	40.77	25.35	13.60	0.386	0.497
	SVDQuant (Li et al., 2024)	4/4	69.19	65.76	97.53	94.57	94.74	46.75	25.83	13.80	0.392	0.495
	ViDiT-Q (Zhao et al., 2024)	4/4	69.42	64.59	97.74	93.30	92.59	36.92	25.80	13.12	0.353	0.561
	DVD-Quant (Li et al., 2026c)	4/4	69.87	66.64	98.52	94.78	**95.58**	38.59	26.02	13.39	0.380	0.509
	OrbitQuant (Lee et al., 2026)	4/4	69.04	**67.02**	98.51	**95.18**	94.77	44.69	26.20	13.91	0.401	0.476
	PulseQuant (ours)	4/4	**69.96**	66.80	**98.85**	95.10	95.32	**47.60**	**26.33**	**14.07**	**0.418**	**0.461**

4.4 Calibration and deployment

Algorithm 1 summarizes offline calibration. Recorded activations support both radius selection and response-axis estimation, while the propagation profile weights calibration calls by downstream sensitivity. With K_{p} power iterations and M_ℓ retained tokens, axis estimation costs $O(2K_{\mathrm{p}}M_\ell d)$ per layer; correction for an $m \times d$ weight matrix costs $O(md \log G)$ per pass. These costs are incurred offline. The calibrated indices and row radii, together with the transform metadata, define the exported model used in Eq. 2. Each layer shares its calibrated weights across denoising steps, while propagation profiles and response axes guide offline calibration.

Algorithm 1 PulseQuant offline calibration

Require: Model f, calibration set $\mathcal{D}$, transforms $(\mathbf{D}, \mathbf{\Pi}_d)$, codebook $\mathcal{C}$, anchors $\mathcal{A}$
Ensure: Corrected indices $\{c_{ij}\}$, row radii $\{\bar{r}_i\}$
Run dense f on $\mathcal{D}$; record activations $\mathbf{X}'_{\ell,s}$
for each block–step anchor $(b, t) \in \mathcal{A}$ **do**
  Apply one 4-bit pulse; compute $g_{b,t}$ by Eq. 4
end for
Interpolate and normalize $\{g_{b,t}\}$ into $\omega_{\ell,t}$ by Eq. 5
for each target layer ℓ **do**
  Initialize $\{c_{ij}\}$ by Eq. 3; select $\{\bar{r}_i\}$ by Eq. 6
  Estimate $(\mathbf{a}_{\ell,0}, \mathbf{a}_{\ell,1})$ by Eq. 7
  for each row i and channel group $\mathcal{G}$ **do**
    Rank code moves by decrease in Eq. 8
    Apply the minimum-loss cumulative prefix
  end for
end for
return $\{c_{ij}\}$, $\{\bar{r}_i\}$, and transform metadata

5 Experiments

5.1 Implementation Details

Models. We evaluate Wan 2.1-1.3B and 14B Wan et al. (2025), Wan 2.2-A14B, Self Forcing Huang et al. (2026a), and MiniMax-H3 MiniMax Research (2026), covering bidirectional and autoregressive generation across model scales and architectures. We generate 81-frame videos for Wan models and Self Forcing, and 124-frame videos for MiniMax-H3. Wan models and MiniMax-H3 use classifier-free guidance (CFG), with default resolutions of 480×832 and 544×960, respectively.

Table 2: **Quantitative evaluations of scale and architecture transfer on VBench (T2V).**

Model	Method	W/A	VBench attributes (%)								Dense-reference fidelity		
			IQ ↑	AQ ↑	MS ↑	DD ↑	BC ↑	SC ↑	Scene ↑	OC ↑	PSNR ↑	SSIM ↑	LPIPS ↓
Wan 2.1-T2V (#Param:14B)	Dense BF16	16/16	67.14	67.12	98.48	66.67	96.69	94.36	39.46	26.46	ref.	ref.	ref.
	ViDiT-Q (Zhao et al., 2024)	4/6	**67.25**	66.23	98.58	56.94	96.30	94.10	41.72	26.31	15.89	0.545	0.378
	OrbitQuant (Lee et al., 2026)	4/6	67.18	67.14	98.60	**62.50**	96.44	95.01	37.86	26.22	**16.93**	**0.585**	**0.325**
	PulseQuant (ours)	4/6	67.22	**67.34**	**98.73**	61.11	**96.56**	**95.21**	**42.81**	**26.52**	16.90	0.579	0.331
	ViDiT-Q (Zhao et al., 2024)	4/4	59.41	58.36	98.13	30.56	95.00	91.59	25.29	23.60	13.27	0.415	0.567
	OrbitQuant (Lee et al., 2026)	4/4	66.55	67.31	98.53	**55.56**	**95.71**	94.44	43.60	26.17	16.25	0.554	0.360
	PulseQuant (ours)	4/4	**66.93**	**67.50**	**98.59**	**55.56**	95.46	**94.61**	**43.68**	**26.28**	**16.70**	**0.565**	**0.349**
Wan 2.2-T2V (#Param:27B)	Dense BF16	16/16	69.58	68.17	97.82	77.78	95.10	92.48	48.40	26.70	ref.	ref.	ref.
	ViDiT-Q (Zhao et al., 2024)	4/6	**69.95**	**68.06**	97.96	76.39	94.79	92.68	**48.47**	26.55	14.90	0.464	0.412
	OrbitQuant (Lee et al., 2026)	4/6	69.23	67.52	97.95	77.78	**95.03**	92.65	47.75	**26.65**	15.78	**0.509**	**0.358**
	PulseQuant (ours)	4/6	69.86	67.93	**97.98**	**79.17**	94.80	**92.81**	46.44	**26.65**	**15.93**	**0.509**	0.361
	ViDiT-Q (Zhao et al., 2024)	4/4	69.22	67.24	97.62	72.22	94.07	91.46	46.00	26.34	14.63	0.436	0.449
	OrbitQuant (Lee et al., 2026)	4/4	68.88	67.77	97.88	75.00	94.50	92.17	43.82	**26.72**	15.15	0.476	0.396
	PulseQuant (ours)	4/4	**69.24**	**67.94**	**97.95**	**76.39**	**94.82**	**92.61**	**46.58**	26.44	**15.47**	**0.487**	**0.389**
MiniMax-H3 (#Param:33B)	Dense BF16	16/16	68.57	65.38	99.01	75.00	96.08	92.55	47.60	26.64	ref.	ref.	ref.
	ViDiT-Q (Zhao et al., 2024)	4/6	67.25	65.54	98.54	61.11	**96.05**	**93.94**	41.06	25.67	12.72	0.376	0.569
	OrbitQuant (Lee et al., 2026)	4/6	68.35	66.20	98.97	**76.39**	95.54	93.27	45.64	**26.54**	14.04	0.420	0.499
	PulseQuant (ours)	4/6	**68.51**	**66.67**	**98.98**	75.00	95.47	92.50	**49.56**	26.26	**14.26**	**0.430**	**0.496**
	ViDiT-Q (Zhao et al., 2024)	4/4	59.93	53.94	97.82	43.06	95.08	91.10	20.35	22.99	12.57	0.349	0.646
	OrbitQuant (Lee et al., 2026)	4/4	66.48	65.35	98.72	70.83	**95.16**	**92.84**	42.30	26.14	13.86	0.409	0.518
	PulseQuant (ours)	4/4	**66.83**	**65.43**	**98.73**	70.83	94.97	92.23	**46.88**	**26.24**	**14.17**	**0.424**	**0.514**

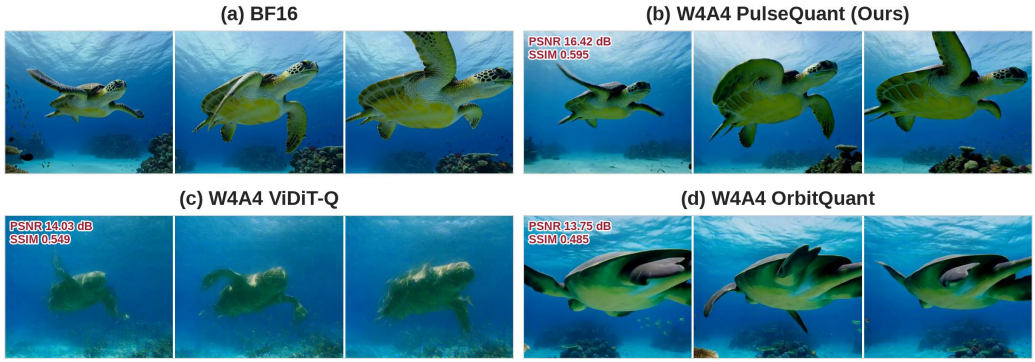
Baselines and metrics. We compare PulseQuant with SmoothQuant (Xiao et al., 2023), SVDQuant (Li et al., 2024), ViDiT-Q (Zhao et al., 2024), DVD-Quant (Li et al., 2026c), and OrbitQuant (Lee et al., 2026), using Dense BF16 as the reference. We report seven VBench attributes (Huang et al., 2024b) and paired PSNR, SSIM (Wang et al., 2004), and LPIPS (Zhang et al., 2018) against Dense BF16 videos generated from the same noise. VBench characterizes generated-video quality, while the paired metrics measure fidelity to the dense reference.

Precision and systems protocol. W4A4 and W4A6 use 4-bit weights and, respectively, 4- or 6-bit activations in every target linear layer. For all quality comparisons, we run every method on NVIDIA H200 using simulated quantization–dequantization (Q/DQ), with the same linear-layer coverage within each setting. Within each setting, methods share evaluation prompts, seeds, and scheduler settings. Calibration statistics are re-estimated for each checkpoint and precision. The systems evaluation is conducted on RTX 5080, where we deploy Wan 2.1-1.3B with packed NVFP4 weights (Alvarez et al., 2025) and native W4A4 Tensor Core kernels. We report synchronized DiT-step latency and DiT-only peak allocation, excluding offline calibration and packing, text encoding, and VAE decoding. Section 5.4 reports the deployment results, while Appendix D provides kernel coverage and backend settings.

5.2 Performance Analysis

Comparison with baselines. Table 1 compares methods on Wan 2.1-1.3B and Self Forcing. Across both models and precisions, PulseQuant achieves the highest Overall Consistency and PSNR and the lowest LPIPS among PTQ methods, with the best SSIM in three of four settings. The gains are most consistent in dense-reference fidelity and Overall Consistency, while individual appearance attributes remain mixed. On Wan 2.1-1.3B at W4A4, for example, OrbitQuant retains higher image and aesthetic quality scores, whereas PulseQuant better preserves the dense reference across all three paired metrics. This contrast highlights that preserving the dense model's output and maximizing individual appearance scores are related but distinct goals. On Self Forcing, the fidelity gains accompany the highest Motion Smoothness at both precisions, extending the benefit to autoregressive generation.

Transfer across scale and architecture. Table 2 evaluates transfer across models with 14B–33B parameters. At W4A4, PulseQuant achieves the best PSNR, SSIM, and LPIPS among PTQ methods on all three models, spanning bidirectional Wan models and autoregressive MiniMax-H3. At W4A6, results are more mixed: PulseQuant leads six VBench attributes on Wan 2.1-14B, while OrbitQuant retains the best paired fidelity; on MiniMax-H3, PulseQuant leads all paired metrics,



Text prompt: "CG game concept digital art, a majestic green turtle gracefully swimming through a vast, crystal-clear ocean. The turtle has smooth, almost iridescent green skin, large flippers, and a long, curved shell. It swims with fluid motions, its tail flicking gently as it propels itself forward. The ocean is deep blue, with gentle waves and schools of colorful fish swimming alongside. Sunlight filters down from above, casting shimmering patterns on the water's surface. The background features a serene, underwater landscape with coral reefs and sandy bottoms. Low-angle, slow-motion shot focusing on the turtle's perspective, highlighting its intricate details and the serene beauty of the ocean."

Figure 3: MiniMax-H3 W4A4 comparison at frames 5, 60, and 100 under the same condition.

Table 3: Component ablation on Wan 2.1-T2V-1.3B W4A6.

Configuration	IQ ↑	AQ ↑	OC ↑	PSNR ↑	LPIPS ↓
Uniform PTQ	56.90	57.29	23.43	14.00	0.549
+ Rotation	63.20	62.39	24.70	14.97	0.443
+ Risk	64.38	**64.97**	25.73	15.59	0.414
+ Subspace corr.	64.64	64.24	25.74	15.61	0.412
PulseQuant	**65.83**	64.93	**26.18**	**16.08**	**0.385**

Table 4: Ablation results of correction-direction ablation on Wan 2.1-T2V-1.3B W4A4 with other settings held fixed.

Direction	IQ ↑	AQ ↑	DD ↑	Scene ↑	OC ↑
None	62.17	**62.90**	51.39	35.17	25.50
Random	61.62	62.86	48.61	37.72	25.38
Fixed Walsh	61.96	62.24	51.39	34.52	25.08
Calibrated (ours)	**62.70**	62.46	**55.56**	**37.78**	**25.64**

but OrbitQuant achieves higher Overall Consistency. The clearest cross-model advantage is therefore dense-reference fidelity at W4A4, with more metric-dependent trade-offs at W4A6. Figure 3 presents a qualitative MiniMax-H3 comparison under matched generation conditions. ViDiT-Q preserves the coarse layout but suppresses fine texture, whereas OrbitQuant remains sharp while changing the subject scale and trajectory; PulseQuant more closely follows the dense composition and detail. Appendix E provides further qualitative comparisons across models and precisions.

5.3 Ablation Studies

We test the propagation premise, isolate the two calibration components, and analyze correction and profiling choices. Complete sweeps and secondary controls are reported in Appendix C.

Effect of each component. Table 3 separates the contributions of rotation and the two calibration modules. The risk-only and correction-only variants independently extend the same rotation baseline; the full method combines both modules. Rotation provides the initial quality recovery, and each calibration module further improves Overall Consistency and paired fidelity. Combining both modules gives the best Overall Consistency, PSNR, and LPIPS, supporting their complementary roles in trajectory prioritization and response preservation. Table 7 further shows that subspace correction reduces high-pass temporal error, linking response preservation to temporal fidelity.

Correction analysis. Activation-adapted directions improve correction quality at a comparable edited-code fraction. In Table 4, calibrated axes lead in four of five metrics over random, fixed Walsh, and no-correction controls. Figure 5 shows lower protected-subspace residuals with nearly unchanged edited-code fraction and global weight MSE. Together, these results support directing code edits toward calibration-relevant responses rather than minimizing weight error alone.

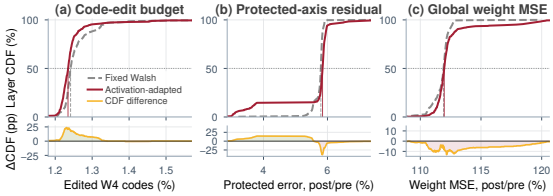


Figure 5: Response-axis adaptation lowers subspace error with code edits and weight MSE in Wan-14B.

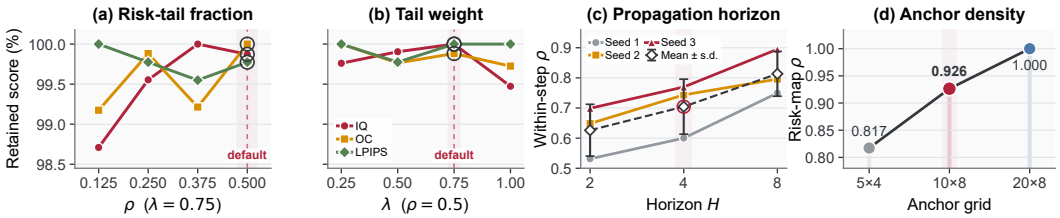


Figure 4: Sensitivity and profiling choices. (a–b) Risk-objective robustness on Wan 2.1-T2V-1.3B W4A4. (c–d) Propagation-horizon stability and anchor-grid agreement on Wan 2.1-T2V-14B.

Effect of propagation risk. Figure 1 exposes the limitation of local reconstruction: block-output RMSE is nearly uncorrelated with final latent error ($\rho = -0.08$), whereas four-step propagated error recovers the within-timestep ranking of final error ($\rho = 0.70$). Perturbations of similar local magnitude can therefore have different downstream effects, supporting propagation-aware calibration.

Sensitivity and profiling design. Figure 4 examines sensitivity to the risk-objective parameters and profiling granularity. Performance remains stable over the tested ρ and λ ranges. We use $\rho = 0.5$ and $\lambda = 0.75$ to balance average response error with emphasis on calls having large risk-weighted errors. For propagation profiling, $H = 4$ maintains a strong correlation with final error across seeds while requiring fewer continuation steps than $H = 8$. The 10×8 block–timestep grid also closely agrees with the 20×8 reference while halving the number of profiling calls. These settings provide the default balance between profiling fidelity and cost.

5.4 Efficiency Discussion

Native inference efficiency. Figure 6 reports native W4A4 inference on Wan 2.1-1.3B using one RTX 5080 under the measurement protocol above. Appendix D details the packed NVFP4 implementation and kernels.

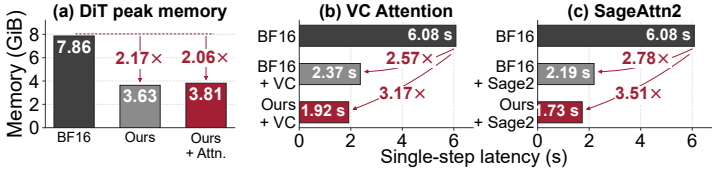


Figure 6: RTX 5080 W4A4 efficiency: (a) DiT memory; (b,c) VC/SageAttention2 per-step latency and speedup relative to BF16.

Packed NVFP4 reduces DiT peak memory from 7.86 to 3.63 GiB; combining it with low-bit attention brings the peak to 3.81 GiB. Low-bit linear layers also complement the latency savings from low-bit attention. VC-Attention (Li et al., 2026b) reduces DiT-step latency from 6.08 to 2.37 s; combining it with PulseQuant further lowers latency to 1.92 s, giving a 3.17× speedup over Dense BF16. SageAttention2 (Zhang et al., 2024) reaches 2.19 s alone and 1.73 s with PulseQuant, for a 3.51× speedup over the same baseline. Thus, native W4A4 linear kernels provide additional reductions in DiT-step latency on top of both low-bit attention backends.

6 Conclusion

We introduced PulseQuant, a 4-bit post-training quantization method that combines propagation-guided radius calibration with activation-informed code correction. It uses downstream sensitivity to prioritize calibration states and dominant activation directions to guide discrete weight updates, preserving the existing 4-bit representation. Controlled interventions show that short-horizon propagated error predicts final latent error more reliably than immediate block-output error. Evaluations across bidirectional and autoregressive video generators demonstrate gains in key quality metrics, with consistent dense-reference fidelity advantages over the state-of-the-art PTQ baselines. Native packed NVFP4 deployment reduces DiT memory and provides additional DiT-step acceleration when combined with low-bit attention. These results support calibration guided by both error propagation and activation response geometry. Reducing profiling cost and extending correction beyond a compact activation subspace remain important directions for future work.

REPRODUCIBILITY STATEMENT

Section 4 describes the calibration procedure and discrete correction algorithm, with derivations in Appendix A. Section 5 specifies the evaluated models, baselines, metrics, and comparison protocol; Appendix B details the quantization operators, profiling configuration, radius search, and default hyperparameters. Quality comparisons use matched prompts, seeds, and scheduler settings with quantization–dequantization on H200, while native deployment is evaluated separately on RTX 5080. Appendix D documents kernel coverage, software versions, and measurement boundaries for the systems results.

ETHICS STATEMENT

PulseQuant lowers the memory and computational requirements of video generation, enabling broader access to existing models on resource-constrained hardware. These efficiency gains can support research and creative applications with more modest computing resources. Our experiments evaluate the quality and efficiency of quantized models and involve no human-subject studies. Responsible use should follow the licenses and safeguards of the underlying models, respect privacy and consent, and clearly disclose synthetic content where appropriate to mitigate misleading or harmful uses.

AI USE STATEMENT

AI tools assisted with research ideation and execution, including discussing methodological alternatives, supporting code implementation and debugging, and inspecting experimental records. They also helped organize, revise, and polish the manuscript for clarity and readability. All AI-assisted text and analysis were reviewed by the authors, and every reported number, claim, and citation was verified against our own experimental logs and the cited sources. We take responsibility for the final content of this work.

REFERENCES

Eduardo Alvarez, Omri Almog, Eric Chung, Simon Layton, Dusan Stosic, Ronny Krashinsky, and Kyle Aubrey. Introducing nvfp4 for efficient and accurate low-precision inference. *NVIDIA Technical Blog*, 2025.

Saleh Ashkboos, Amirkeivan Mohtashami, Maximilian L Croci, Bo Li, Pashmina Cameron, Martin Jaggi, Dan Alistarh, Torsten Hoefler, and James Hensman. Quarot: Outlier-free 4-bit inference in rotated llms. *Advances in Neural Information Processing Systems*, 37:100213–100240, 2024.

Lei Chen, Yuan Meng, Chen Tang, Xinzhu Ma, Jingyan Jiang, Xin Wang, Zhi Wang, and Wenwu Zhu. Q-dit: Accurate post-training quantization for diffusion transformers. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28306–28315. IEEE, 2025.

Vage Egiazarian, Andrei Panferov, Denis Kuznedelev, Elias Frantar, Artem Babenko, and Dan Alistarh. Extreme compression of large language models via additive quantization. *arXiv preprint arXiv:2401.06118*, 2024.

Weilun Feng, Chuanguang Yang, Haotong Qin, Xiangqi Li, Yu Wang, Zhulin An, Libo Huang, Boyu Diao, Zixiang Zhao, Yongjun Xu, and Michele Magno. Q-VDiT: Towards accurate quantization and distillation of video-generation diffusion transformers. *arXiv preprint arXiv:2505.22167*, 2025.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. Gptq: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323*, 2022.

Yefei He, Luping Liu, Jing Liu, Weijia Wu, Hong Zhou, and Bohan Zhuang. Ptqd: Accurate post-training quantization for diffusion models. *Advances in neural information processing systems*, 36:13237–13249, 2023.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *Advances in Neural Information Processing Systems*, 38:167283–167308, 2026a.

Yushi Huang, Ruihao Gong, Jing Liu, Tianlong Chen, and Xianglong Liu. Tfmq-dm: Temporal feature maintenance quantization for diffusion models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7362–7371. IEEE, 2024a.

Yushi Huang, Ruihao Gong, Jing Liu, Yifu Ding, Chengtao Lv, Haotong Qin, and Jun Zhang. Qvgen: Pushing the limit of quantized video generative models. In *International Conference on Learning Representations*, volume 2026, pp. 155088–155121, 2026b.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21807–21818. IEEE, 2024b.

Donghyun Lee, Jitesh Chavan, Duy Nguyen, Sam Huang, Liming Jiang, Priyadarshini Panda, Timo Mertens, and Saurabh Shukla. Orbitquant: Data-agnostic quantization for image and video diffusion transformers. *arXiv preprint arXiv:2607.02461*, 2026.

Seunghoon Lee, Jeongwoo Choi, Byunggwan Son, Jaehyeon Moon, Jeimin Jeon, and Bumsub Ham. Accuquant: Simulating multiple denoising steps for quantizing diffusion models. *Advances in Neural Information Processing Systems*, 38:110108–110143, 2025.

Muyang Li, Yujun Lin, Zhekai Zhang, Tianle Cai, Xiuyu Li, Junxian Guo, Enze Xie, Chenlin Meng, Jun-Yan Zhu, and Song Han. Svdquant: Absorbing outliers by low-rank components for 4-bit diffusion models. *arXiv preprint arXiv:2411.05007*, 2024.

Xingyang Li, Samuel Tesfai, Zhekai Zhang, Haocheng Xi, Shuo Yang, Lvmin Zhang, Yufei Sun, Kelly Peng, Maneesh Agrawala, Ion Stoica, Kurt Keutzer, Jun-Yan Zhu, Song Han, Yujun Lin, and Muyang Li. DeltaQuant: 4-bit video diffusion models with spatiotemporal delta smoothing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 43578–43588, 2026a.

Xingyang Li, Dongyun Zou, Shining Zhang, Jiacheng Chen, Haocheng Xi, Lvmin Zhang, Jun-Yan Zhu, Song Han, Zhekai Zhang, Yujun Lin, and Muyang Li. Vc-attention: Value smoothing and softmax casting for low-bit attention. *arXiv preprint arXiv:2609.15810*, 2026b.

Xiuyu Li, Yijiang Liu, Long Lian, Huanrui Yang, Zhen Dong, Daniel Kang, Shanghang Zhang, and Kurt Keutzer. Q-diffusion: Quantizing diffusion models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17489–17499. IEEE, 2023.

Yuhang Li, Ruihao Gong, Xu Tan, Yang Yang, Peng Hu, Qi Zhang, Fengwei Yu, Wei Wang, and Shi Gu. Brecq: Pushing the limit of post-training quantization by block reconstruction. *arXiv preprint arXiv:2102.05426*, 2021.

Zhiteng Li, Hanxuan Li, Junyi Wu, Kai Liu, Haotong Qin, Linghe Kong, Guihai Chen, Yulun Zhang, and Xiaokang Yang. Dvd-quant: data-free video diffusion transformers quantization. In *International Conference on Learning Representations*, volume 2026, pp. 35463–35475, 2026c.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. Awq: Activation-aware weight quantization for on-device llm compression and acceleration. *Proceedings of machine learning and systems*, 6:87–100, 2024.

Zechun Liu, Changsheng Zhao, Igor Fedorov, Bilge Soran, Dhruv Choudhary, Raghuraman Krishnamoorthi, Vikas Chandra, Yuandong Tian, and Tijmen Blankevoort. Spinquant: Llm quantization with learned rotations. In *International Conference on Learning Representations*, volume 2025, pp. 92009–92032, 2025.

MiniMax Research. MiniMax H3: An open model breaking the boundaries between tasks and modalities, July 2026. URL `https://www.minimax.io/blog/minimax-h3`. Published July 31, 2026.

Markus Nagel, Rana Ali Amjad, Mart Van Baalen, Christos Louizos, and Tijmen Blankevoort. Up or down? adaptive rounding for post-training quantization. In *International conference on machine learning*, pp. 7197–7206. PMLR, 2020.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182. IEEE, 2023.

Yuzhang Shang, Zhihang Yuan, Bin Xie, Bingzhe Wu, and Yan Yan. Post-training quantization on diffusion models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1972–1981. IEEE, 2023.

Wenqi Shao, Mengzhao Chen, Zhaoyang Zhang, Peng Xu, Lirui Zhao, Zhiqian Li, Kaipeng Zhang, Peng Gao, Yu Qiao, and Ping Luo. Omniquant: Omnidirectionally calibrated quantization for large language models. In *International Conference on Learning Representations*, volume 2024, pp. 45472–45496, 2024.

Junhyuk So, Jungwon Lee, Daehyun Ahn, Hyungjun Kim, and Eunhyeok Park. Temporal dynamic quantization for diffusion models. *Advances in neural information processing systems*, 36:48686–48698, 2023.

Albert Tseng, Jerry Chee, Qingyao Sun, Volodymyr Kuleshov, and Christopher De Sa. QuIP#: Even better LLM quantization with hadamard incoherence and lattice codebooks. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 48630–48656, 2024.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Changyuan Wang, Ziwei Wang, Xiuwei Xu, Yansong Tang, Jie Zhou, and Jiwen Lu. Towards accurate post-training quantization for diffusion models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16026–16035. IEEE, 2024.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. Smoothquant: Accurate and efficient post-training quantization for large language models. In *International conference on machine learning*, pp. 38087–38099. PMLR, 2023.

Jintao Zhang, Haofeng Huang, Pengle Zhang, Jia Wei, Jun Zhu, and Jianfei Chen. Sageattention2: Efficient attention with thorough outlier smoothing and per-thread int4 quantization. *arXiv preprint arXiv:2411.10958*, 2024.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 586–595. IEEE, 2018.

Tianchen Zhao, Tongcheng Fang, Haofeng Huang, Enshu Liu, Rui Wan, Widyadewi Soedarmadji, Shiyao Li, Zinan Lin, Guohao Dai, Shengen Yan, Huazhong Yang, Xuefei Ning, and Yu Wang. ViDiT-Q: Efficient and accurate quantization of diffusion transformers for image and video generation. *arXiv preprint arXiv:2406.02540*, 2024.

Appendix

CONTENTS

A METHOD DERIVATIONS AND DISCRETE CORRECTION 13
A.1 FIRST-ORDER VIEW OF PROPAGATION RISK . . . 13
A.2 RESPONSE-ERROR DECOMPOSITION . . . 13
A.3 ADJACENT-CODE PREFIX SEARCH . . . 14
B EXPERIMENTAL PROTOCOL AND IMPLEMENTATION DETAILS 14
B.1 QUANTIZATION AND CALIBRATION IMPLEMENTATION . . . 14
B.2 DEFAULT HYPERPARAMETER SETTINGS . . . 15
B.3 NOTATION . . . 15
C ADDITIONAL ABLATION STUDIES 15
C.1 COMPLETE COMPONENT RESULTS . . . 16
C.2 CALIBRATION-SET SENSITIVITY . . . 17
C.3 PROPAGATION-RISK OBJECTIVE . . . 18
D NATIVE W4A4 DEPLOYMENT AND PROFILING 18
D.1 PACKED LINEAR PATH . . . 18
D.2 FUSION AND WORKSPACE REUSE . . . 19
D.3 ATTENTION BACKEND AND MEASUREMENT BOUNDARY . . . 19
E ADDITIONAL QUALITATIVE RESULTS 19
E.1 WAN 2.1-1.3B W4A4 COMPARISONS . . . 19
E.2 W4A6 AND CROSS-ARCHITECTURE COMPARISONS . . . 19

A METHOD DERIVATIONS AND DISCRETE CORRECTION

This section provides the derivations omitted from the main text. We first relate the measured propagation gain to first-order trajectory dynamics, then decompose the response error controlled by subspace correction, and finally specify the legal-code search used by PulseQuant.

A.1 FIRST-ORDER VIEW OF PROPAGATION RISK

Write one dense denoising update as $\mathbf{z}_{t+1} = F_t(\mathbf{z}_t)$, and let $\mathbf{e}_t$ be a small perturbation of the trajectory state. Linearizing each update gives

$$\mathbf{e}_{t+1} \approx \mathbf{J}_t \mathbf{e}_t + \mathbf{u}_t, \qquad \mathbf{e}_T \approx \sum_{t=0}^{T-1} \mathbf{\Phi}_{T,t+1} \mathbf{u}_t, \quad \mathbf{\Phi}_{T,t+1} = \mathbf{J}_{T-1} \cdots \mathbf{J}_{t+1}, \tag{9}$$

where $\mathbf{J}_t$ is the Jacobian of F_t and $\mathbf{u}_t$ is the local error introduced at update t. For an isolated error pulse, the ratio of its propagated error to its immediate error estimates the finite-horizon amplification induced by $\mathbf{\Phi}_{t+1+H,t+1}$. Equation 4 measures this amplification directly from latent trajectories and therefore does not require explicit Jacobian construction.

A.2 RESPONSE-ERROR DECOMPOSITION

We first consider the full layer-input space, where the two response axes are orthonormal. Write $\mathbf{x}' = \beta_0 \mathbf{a}_0 + \beta_1 \mathbf{a}_1 + \mathbf{x}_\perp$, with $\beta_u = \langle \mathbf{x}', \mathbf{a}_u \rangle$ and $\mathbf{x}_\perp$ orthogonal to both axes. For the full-row residual $\mathbf{r} = \mathbf{w}'_i - \mathbf{q}_i$, the triangle and Cauchy–Schwarz inequalities give

$$|\langle \mathbf{r}, \mathbf{x}' \rangle| \le \sum_{u=0}^{1} |\beta_u| \, |\langle \mathbf{r}, \mathbf{a}_u \rangle| + \|\mathbf{r}\|_2 \|\mathbf{x}_\perp\|_2. \tag{10}$$

The bound separates response error along the selected axes from the contribution outside their span. Estimating these axes from transformed calibration activations focuses correction on directions with high observed activation energy, while controlling the residual norm limits its interaction with the remaining component.

Equation 8 implements this principle group-wise. Each full-row response residual decomposes as $\langle \mathbf{r}, \mathbf{a}_u \rangle = \sum_{\mathcal{G}} \langle \mathbf{r}_{\mathcal{G}}, \mathbf{a}_{u,\mathcal{G}} \rangle$, where subscripts $\mathcal{G}$ denote channel slices. These slices generally have different norms and need not remain mutually orthogonal. The group objective therefore uses normalized directional penalties, together with full-group distortion, as a local surrogate for response preservation rather than an exact orthogonal decomposition of the full-row error.

A.3 Adjacent-code prefix search

For a coordinate with current code c_j, the two legal adjacent moves and their induced weight changes are

$$c_j^- = \max(c_j - 1, 0), \quad c_j^+ = \min(c_j + 1, 15), \quad \delta_j^s = \bar{r}_i (C_{c_j^s} - C_{c_j}), \quad s \in \{-, +\}. \tag{11}$$

All single-coordinate proposals in a correction pass are evaluated from the same current reconstruction $\mathbf{q}$. For each coordinate, we retain the sign s_j that produces the smaller single-move value of $\mathcal{L}_{i,\mathcal{G}}$. We sort the retained proposals by increasing objective change to obtain an order π. For every cumulative prefix, we recompute the full joint objective, including the cross terms induced by simultaneous coordinate changes in the protected responses, and select

$$K^* = \arg \min_{K \in \{0,\dots,G\}} \mathcal{L}_{i,\mathcal{G}} \left(\mathbf{q} + \sum_{k=1}^{K} \delta_{\pi(k)}^{s_{\pi(k)}} \mathbf{e}_{\pi(k)} \right). \tag{12}$$

The empty prefix $K = 0$ is evaluated first, and ties are resolved in favor of the earliest prefix. Thus, the original reconstruction is retained whenever no non-empty prefix improves the objective. Including $K = 0$ ensures $\mathcal{L}_{i,\mathcal{G}}(\mathbf{q}_{\text{new}}) \le \mathcal{L}_{i,\mathcal{G}}(\mathbf{q})$. Every accepted coordinate uses a legal codebook index at the calibrated row radius, preserving the deployment format.

B Experimental protocol and implementation details

This section specifies the quantization operators and calibration settings used in the experiments. Within each setting, methods share the model revision, quantized-layer coverage, evaluation prompts, seeds, scheduler, and evaluator. Calibration statistics are re-estimated for each checkpoint and precision.

B.1 Quantization and calibration implementation

Activation quantizer. For unclipped symmetric per-token quantization with activation bit width b_a, let $q_{\max} = 2^{b_a - 1} - 1$ and $\alpha = \max(\|\mathbf{x}'\|_\infty / q_{\max}, \epsilon_{\text{scale}})$. The dequantized output is

$$Q_A(\mathbf{x}') = \alpha \ \text{clip}\left(\text{round}(\mathbf{x}'/\alpha), -q_{\max}, q_{\max} \right). \tag{13}$$

Reduction and rounding are performed in FP32 before restoring the input dtype. The scale floor, rounding rule, and backend are held fixed across methods within each comparison.

Profiling pulse configuration. The CFG-based block interventions used to construct the propagation-risk profiles apply a profiling-only uniform W4A4 quantizer to ten principal linear projections: the Q, K, V, and output projections of self-attention and cross-attention, together with the FFN input and output projections. Weights use channel balancing, block-Hadamard coordinates, and groupwise calibrated uniform 4-bit quantization. Activations use dynamic per-token symmetric 4-bit quantization with a frozen clipping schedule. The pulse is active in both CFG forward calls at the selected timestep; all other blocks and the subsequent H denoising updates execute in dense BF16.

Radius candidate grid. For each transformed weight row, we initialize the radius as $r_i^{(0)} = \|\mathbf{w}_i'\|_2$ and construct a symmetric multiplicative grid:

$$\mathcal{R}_i = \left\{ r_i^{(0)} \left(1 - \eta + \frac{2\eta k}{K_r - 1} \right) \right\}_{k=0}^{K_r - 1}. \tag{14}$$

The main experiments on Wan 2.1-1.3B, Wan 2.1-14B, MiniMax-H3, and Self Forcing use $K_r = 5$ and $\eta = 0.08$, giving $\mathcal{R}_i = r_i^{(0)} \{0.92, 0.96, 1.00, 1.04, 1.08\}$. The grid searches relative changes around each row norm rather than a shared absolute offset.

Table 5: Default PulseQuant calibration configuration. Model-specific exceptions are reported with the evaluation protocol.

Stage	Setting	Default
Calibration data	Number of prompts	3
Calibration data	Prompt composition	Complementary coverage
Propagation profile	Horizon H	4 steps
Propagation profile	Anchor grid	10 blocks $\times$ 8 timesteps
Risk objective	Tail weight λ	0.75
Risk objective	Tail fraction ρ	0.5
Radius search	Number of candidates K_r	5
Radius search	Relative half-span η	0.08
Response axes	Token budget $M_{\max}$	512
Response axes	Power iterations K_{p}	2
Response axes	Number of protected axes	2
Code correction	Group width G	128
Code correction	Distortion weight τ	0.25

B.2 Default hyperparameter settings

Table 5 summarizes the shared calibration configuration. We use three complementary prompts covering object appearance, object motion, and camera motion; Appendix C.2 reports the corresponding prompt-composition study. These defaults are fixed across the main experiments unless a model-specific exception is stated in the evaluation protocol.

B.3 Notation

Table 6 lists the symbols used in the method and appendices.

Table 6: Notation used in the method and implementation details.

Symbol	Meaning
$\mathbf{W}, \mathbf{x}, \mathbf{X}; m, d, N$	Weight matrix, token, batch; output/input widths, tokens
$\mathbf{\Pi}_d, \mathbf{D}$	Orthogonal RPBH transform; positive channel scaling
$\mathcal{C}, C_k, c_{ij}$	Shared 16-value codebook, entry, and coordinate index
Q_A, b_a	Activation quantizer and activation bit width
$r_i^{(0)}, r_i, \bar{r}_i$	Initial row norm (also denoted r_i) and calibrated radius
$\mathcal{R}_i, K_r, \eta$	Radius candidates, grid size, and relative half-span
$\ell, i, j; s, t(s)$	Layer, row, coordinate; calibration call and its denoising step
t, T, H	Denoising step, total updates, and profiling horizon
$\mathbf{X}'_{\ell,s}, N_s$	Transformed calibration activations and token count at call s
$g_{b,t}, \bar{g}_{\ell,t}, \omega_{\ell,t}$	Anchor gain, interpolated gain, and normalized calibration weight
$\gamma, [a, b]$	Gain exponent and clipping interval
$e_{i,s}(r), u_{i,s}(r)$	Row-output error and its risk-weighted value
λ, ρ	Tail-term weight and retained fraction of calibration calls
$\mathbf{p}_u, \mathbf{a}_{\ell,u}$	Fixed native-input probe and activation-adapted response axis
$\bar{\mathbf{X}}'_\ell, M_\ell, M_{\max}$	Retained activations, retained token count, and token budget
$K_{\mathrm{p}}, \mathbf{C}_\ell$	Power-iteration count and uncentered second moment
$\mathcal{G}, G; \mathbf{w}, \mathbf{q}$	Channel group and width; dense and quantized group slices
δ_j^s, τ	Weight change for move $s \in \{-, +\}$ and full-group distortion weight

C Additional ablation studies

This section supplements the compact ablations in the main paper with complete metrics and controlled studies of the calibration data and propagation-risk objective.

C.1 COMPLETE COMPONENT RESULTS

Table 7 reports the complete metrics corresponding to Table 3. All variants quantize the same 300 linear layers under an identical W4A6 protocol. Their differences therefore isolate the effects of rotation, propagation weighting, and response-subspace correction.

Table 7: Complete component ablation on Wan 2.1-T2V-1.3B W4A6 under matched 300-layer coverage.

	Components			VBench attributes (%)					Dense-reference fidelity			
Configuration	Rotation	Risk	Subsp. corr.	IQ ↑	AQ ↑	MS ↑	BC ↑	OC ↑	PSNR ↑	SSIM ↑	LPIPS ↓	T-HP ↓
Uniform PTQ				56.90	57.29	97.38	95.01	23.43	14.00	0.421	0.549	0.161
Rotation only	✓			63.20	62.39	98.09	95.01	24.70	14.97	0.512	0.443	0.155
Risk only	✓	✓		64.38	**64.97**	98.18	95.58	25.73	15.59	0.529	0.414	0.150
Subspace corr. only	✓		✓	64.64	64.24	98.26	**95.69**	25.74	15.61	0.536	0.412	**0.142**
PulseQuant	✓	✓	✓	**65.83**	64.93	**98.47**	95.30	**26.18**	**16.08**	**0.546**	**0.385**	0.144

Rotation provides the initial quality recovery. Relative to this baseline, risk weighting and response-subspace correction each improve Overall Consistency and all three paired fidelity metrics. The full method achieves the best PSNR, SSIM, and LPIPS, whereas correction alone yields the lowest temporal high-pass error. Thus, combining trajectory prioritization with response correction improves aggregate fidelity, while the temporal diagnostic retains a small trade-off.

Fixed-axis control. The fixed Walsh variant in Table 4 transports native input probes $\mathbf{p}_u$ into quantizer coordinates:

$$\mathbf{a}_{\ell,u}^{\text{fixed}} = \mathbf{\Pi}_d \mathbf{D} \mathbf{p}_u. \tag{15}$$

The corresponding row-response residual is $(\mathbf{w}_i' - \mathbf{q}_i)^\top \mathbf{a}_{\ell,u}^{\text{fixed}}$. Activation-adapted axes are instead estimated directly from transformed calibration activations $\mathbf{X}'$. This comparison contrasts fixed input probes with data-adapted response directions within the same correction framework.

Trajectory-wide diagnostics. Figure 7 examines the self-attention output projection `blocks.0.attn1.to_out.0` on Wan 2.1-14B W4A4. The panels show activation magnitudes before and after balancing, block–timestep variation in propagation gain, and output residuals under naive quantization and PulseQuant. At this layer, PulseQuant lowers relative output RMSE by 41%, with improvements at 96% of the sampled channel–timestep locations. These diagnostics illustrate the activation and trajectory structure motivating calibration and complement the component comparisons in Tables 3 and 7.

Table 8: MiniMax-H3 W4A4 calibration-set sensitivity on the VBench.

Calibration set	# Prompts	VBench attributes (%)								Dense-reference fidelity		
		IQ ↑	AQ ↑	MS ↑	DD ↑	BC ↑	SC ↑	Scene ↑	OC ↑	PSNR ↑	SSIM ↑	LPIPS ↓
Dense BF16	–	68.57	65.38	99.01	75.00	96.08	92.55	47.60	26.64	ref.	ref.	ref.
Prompt count: shared ordered prefix												
Prefix-1	1	66.39	65.55	98.73	68.06	**95.33**	92.42	46.51	25.97	14.26	0.424	0.509
Prefix-3	3	**66.83**	65.43	98.73	**70.83**	94.97	92.23	**46.88**	**26.24**	14.17	0.424	0.514
Prefix-6	6	66.74	**65.75**	**98.74**	68.06	95.29	**92.65**	40.84	26.15	14.22	0.426	0.512
Prefix-10	10	65.87	65.37	**98.74**	**70.83**	95.12	92.42	44.62	26.11	**14.40**	**0.434**	**0.501**
Prefix-20	20	65.52	65.18	98.70	68.06	95.16	92.47	43.60	26.21	14.31	0.431	0.506
Prompt composition: fixed three-prompt budget												
Complementary coverage	3	66.83	65.43	98.73	**70.83**	94.97	92.23	**46.88**	**26.24**	14.17	0.424	0.514
Similar static	3	65.30	65.64	**98.79**	65.28	95.24	92.17	42.30	26.21	**14.39**	**0.430**	0.508
Heterogeneous	3	**67.52**	**66.18**	98.66	65.28	**95.45**	**92.77**	44.04	26.21	14.22	0.422	**0.507**
Similar dynamic	3	66.49	65.61	98.67	66.67	95.26	92.67	44.48	26.08	14.12	0.421	0.511

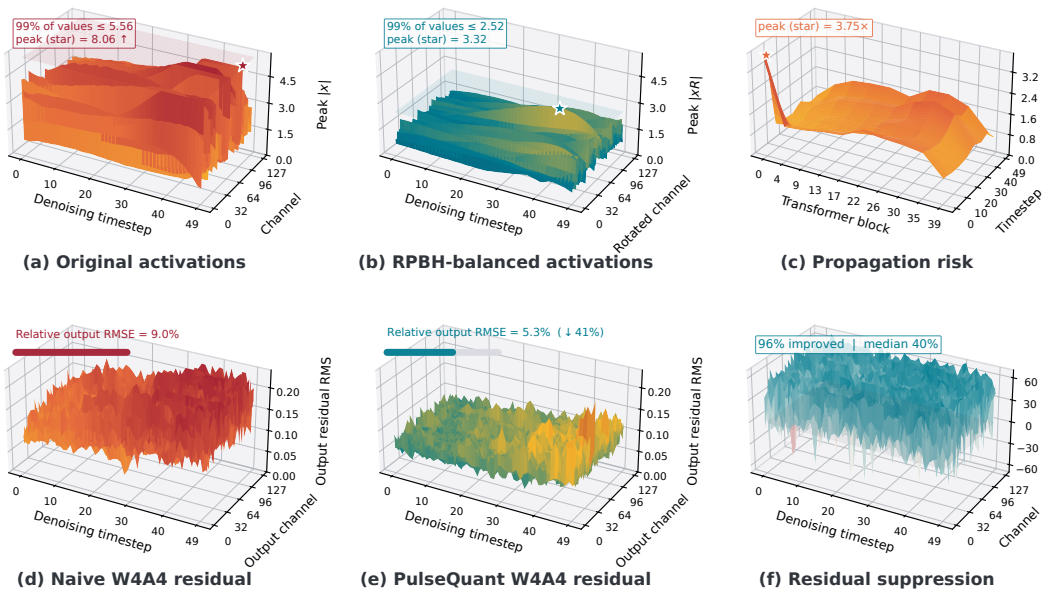


Figure 7: Layer-level activation and error diagnostics on Wan 2.1-14B W4A4 over one matched 50-step calibration trajectory. (a,b) Per-channel peak activation magnitudes before and after scaling and RPBH for the most outlier-heavy 128-channel group; planes and stars mark the 99th percentile and maximum, respectively. (c) Four-step propagation gains measured on 10 blocks $\times$ 8 timestep anchors; intermediate timesteps are interpolated only for visualization. (d,e) RMS output-response residuals relative to BF16 for naive W4A4 and PulseQuant, evaluated on 128 sampled output channels with a shared scale; bars report relative RMSE. (f) Pointwise residual reduction, where positive values favor PulseQuant. At this audited layer, relative RMSE decreases from 9.0% to 5.3%, and 96% of channel–timestep locations improve.

Table 9: Propagation-risk sensitivity on Wan 2.1-T2V-1.3B W4A4. Gray rows mark the default $(\lambda, \rho) = (0.75, 0.5)$.

λ	ρ	VBench attributes (%)									Dense-reference fidelity			N-Jerk ↓
		IQ ↑	AQ ↑	MS ↑	DD ↑	BC ↑	SC ↑	Scene ↑	OC ↑	TF ↑	PSNR ↑	SSIM ↑	LPIPS ↓	
ρ sweep with $\lambda = 0.75$														
0.75	0.125	61.97	62.03	97.78	51.39	94.62	**92.02**	**38.15**	25.23	**98.60**	**15.45**	**0.506**	**0.442**	1.429
0.75	0.250	62.50	62.40	97.80	52.78	94.73	91.89	35.61	25.41	98.59	15.43	0.503	0.443	1.429
0.75	0.375	**62.78**	62.37	97.79	**56.94**	**94.91**	91.76	35.03	25.24	98.57	15.38	0.503	0.444	1.429
0.75	0.500	62.70	**62.46**	**97.83**	55.56	94.84	91.97	36.70	**25.44**	98.57	15.42	0.503	0.443	**1.427**
λ sweep with $\rho = 0.5$														
0.00	0.50	**62.83**	**62.54**	97.80	58.33	94.76	91.69	34.38	25.42	98.56	15.39	**0.504**	**0.443**	1.428
0.25	0.50	62.55	62.47	97.82	56.94	94.65	91.66	35.47	**25.47**	98.55	**15.44**	0.503	**0.443**	1.428
0.50	0.50	62.64	62.50	**97.84**	**59.72**	**94.94**	**92.10**	35.47	25.41	**98.57**	15.38	0.503	0.444	1.428
0.75	0.50	62.70	62.46	97.83	55.56	94.84	91.97	36.70	25.44	**98.57**	15.42	0.503	**0.443**	**1.427**
1.00	0.50	62.37	62.32	97.80	56.94	94.91	92.04	**36.77**	25.40	**98.57**	15.40	0.503	**0.443**	1.428

C.2 Calibration-set sensitivity

We separately study calibration-set size and prompt composition on MiniMax-H3 W4A4 in Table 8. Dense BF16 is included as a common reference and is not part of either sweep.

Prompt count. The count sweep uses nested prefixes of one ordered pool, so larger sets retain all earlier prompts. Increasing the calibration set does not uniformly improve quality: ten prompts give the best paired fidelity, whereas three prompts achieve the highest Overall Consistency and Scene

scores in this sweep. The twenty-prompt set does not improve on ten prompts in any paired fidelity metric, indicating that a larger calibration budget alone does not ensure better reconstruction.

Prompt composition. At a fixed three-prompt budget, composition changes the balance between appearance, motion, and fidelity. Heterogeneous prompts achieve the best image and aesthetic quality and LPIPS, while complementary coverage leads Dynamic Degree, Scene, and Overall Consistency. Similar-static prompts yield the best PSNR, SSIM, and Motion Smoothness. We use complementary coverage to span object appearance, object motion, and camera motion, with strong scene and overall consistency in this comparison. Prefix-3 and complementary coverage use the same checkpoint and are repeated to make the two sweeps self-contained. Table 10 lists the exact prompts.

Table 10: Exact three-prompt calibration sets for the MiniMax-H3 composition study; prompts are reproduced verbatim.

Calibration set	ID	Calibration prompt
Complementary coverage	P1	A ceramic teapot rotating slowly on a wooden table in soft studio light
	P2	A glass marble rolling across a linen cloth while the camera tracks beside it
	P3	A small sailboat crossing a calm lake at dawn with gentle camera movement
Similar static	P1	A ceramic teapot resting on a wooden table under soft studio light, locked-off camera
	P2	A polished red apple resting on a linen-covered table under soft studio light, locked-off camera
	P3	A porcelain vase resting on a wooden shelf under soft studio light, locked-off camera
Heterogeneous	P1	A red fox trotting through a snowy pine forest while loose snow falls from the branches
	P2	A vintage red train crossing a stone bridge in heavy rain with a tracking camera
	P3	A hummingbird hovering beside red flowers in a sunlit rainforest, filmed with a macro lens
Similar dynamic	P1	A golden retriever walking from left to right through a green park while the camera tracks beside it
	P2	A young woman walking from left to right along a quiet garden path while the camera tracks beside her
	P3	A small delivery robot moving from left to right along a tree-lined walkway while the camera tracks beside it

C.3 Propagation-risk objective

Table 9 varies the tail fraction ρ at fixed $\lambda = 0.75$ and the tail weight λ at fixed $\rho = 0.5$. In the ρ sweep, the smallest fraction gives the best paired fidelity, while $\rho = 0.5$ yields the highest Overall Consistency. The λ sweep shows limited variation in paired fidelity, with PSNR between 15.38 and 15.44 dB and LPIPS between 0.443 and 0.444. Other attributes favor different settings, so no configuration dominates all metrics. We retain $(\lambda, \rho) = (0.75, 0.5)$ as a balanced default; it also gives the lowest reported normalized jerk in both sweeps.

D Native W4A4 deployment and profiling

This section provides the implementation details underlying the efficiency results in Section 5.4. We describe the packed linear path, the applied fusions, and the measurement boundary used for native W4A4 inference.

D.1 Packed linear path

We deploy Wan 2.1-1.3B on a single RTX 5080 (SM120) using PyTorch 2.11 and CUDA 12.8. Corrected weights and scales are exported offline and stored in packed NVFP4 format, while activations

are quantized dynamically during inference. Among the 300 packed linear modules, 210 supported modules dispatch to native W4A4 Tensor Core kernels. The remaining operators run at the original model precision and are included in the reported DiT-step latency.

D.2 Fusion and workspace reuse

We fuse the RPBH sign transform with activation packing, fuse the QKV projections, and reuse stream-ordered activation workspaces and invariant BF16 masks across calls. These optimizations avoid repeated allocation, host synchronization, and explicit materialization of the rotated activations. For a representative FFN shape of $(M, K, N) = (32760, 1536, 8960)$, the fused path takes 1.54 ms, compared with 7.74 ms for BF16 GEMM, corresponding to a 5.01$\times$ speedup. It is also 1.18$\times$ faster than the path that materializes the rotation. Across six sequences, workspace reuse produces zero maximum deviation from freshly allocated workspaces.

D.3 Attention backend and measurement boundary

SageAttention2 uses an SM120 CUDA kernel with INT8 $QK^\top$ and FP8 PV computation, while VC Attention additionally fuses V-Smooth into this path. The persistent workspaces add approximately 185 MiB. Before workspace reuse, the combined VC and SageAttention2 paths require 4.07 and 3.69 s per DiT step, respectively. Removing allocation and materialization overhead yields a 2.12$\times$ backend speedup. We synchronize all measurements and report single-video DiT-step latency and DiT-only peak allocation. Offline calibration and packing, text encoding, and VAE decoding are excluded from both measurements.

E Additional qualitative results

This section provides additional visual comparisons under controlled generation settings. All panels use matched prompts, seeds, schedulers, and initial noise. Red labels report full-video PSNR and SSIM against the corresponding BF16 trajectory. We sample frames at fixed temporal positions to reveal both appearance drift and long-range temporal inconsistency.

E.1 Wan 2.1-1.3B W4A4 comparisons

Figures 8, 9, and 10 compare six methods on three fixed prompts using frames 5, 40, and 75. We select the cases once according to the PulseQuant–ViDiT-Q paired-fidelity gap and then filter them for semantic diversity. None of these prompts is used for calibration.

The kitten sequence tests fine foreground geometry and foreground–background separation, whereas the lightning and beach scenes stress global composition and illumination. Across distant frames, PulseQuant better preserves the principal subject and scene layout. The competing methods more frequently alter silhouette, texture, contrast, or background structure. These errors persist across time, showing how trajectory-level quantization error develops into semantic drift.

E.2 W4A6 and cross-architecture comparisons

Figure 11 presents a six-method comparison on Wan 2.1-1.3B W4A6. Figures 12–15 apply the same matched protocol to MiniMax-H3 at W4A6 and W4A4. We show frames 5, 40, and 75 for Wan 2.1-1.3B and frames 5, 60, and 100 for MiniMax-H3. The examples cover rigid fine structure, human appearance, articulated animal motion, and repeated floral texture. Across these settings, PulseQuant remains closer to the dense subject geometry and composition, whereas competing outputs more often alter texture, scale, or pose.



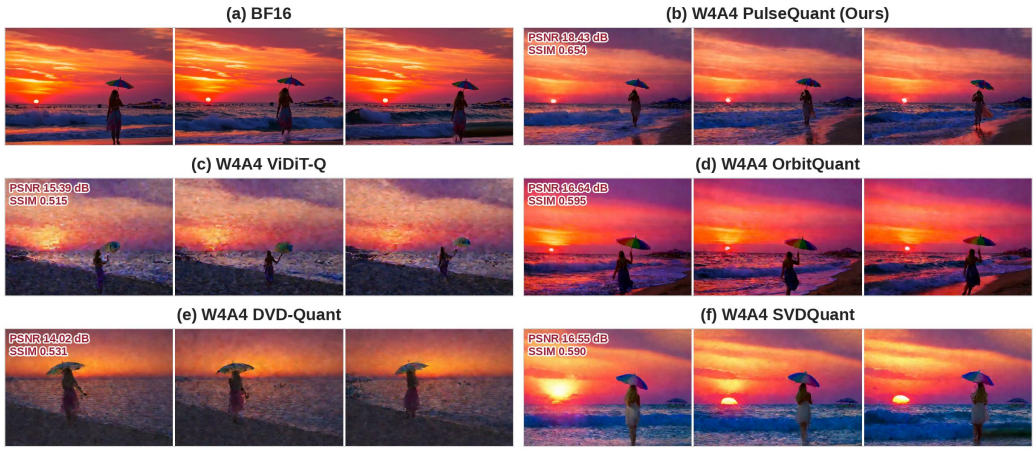
Text prompt: "A dramatic lightning strike illuminates the iconic Eiffel Tower against a backdrop of dark, ominous clouds in the sky. The lightning crackles and illuminates the structure, casting stark shadows and highlighting every intricate detail. Dark storm clouds swirl menacingly overhead, adding to the intense atmosphere. The Eiffel Tower stands tall and proud amidst the storm, its metal lattice frame gleaming in the electric flash. The scene captures the raw power and beauty of nature's fury, with the lightning slicing through the sky. The lightning bolt strikes the tower, sending a shockwave through the air as it leaves a trail of sparks. The image is captured from a low-angle perspective, emphasizing the grandeur and vulnerability of the tower during the storm."

Figure 8: Wan 2.1-1.3B W4A4 comparison for the Eiffel Tower under lightning.



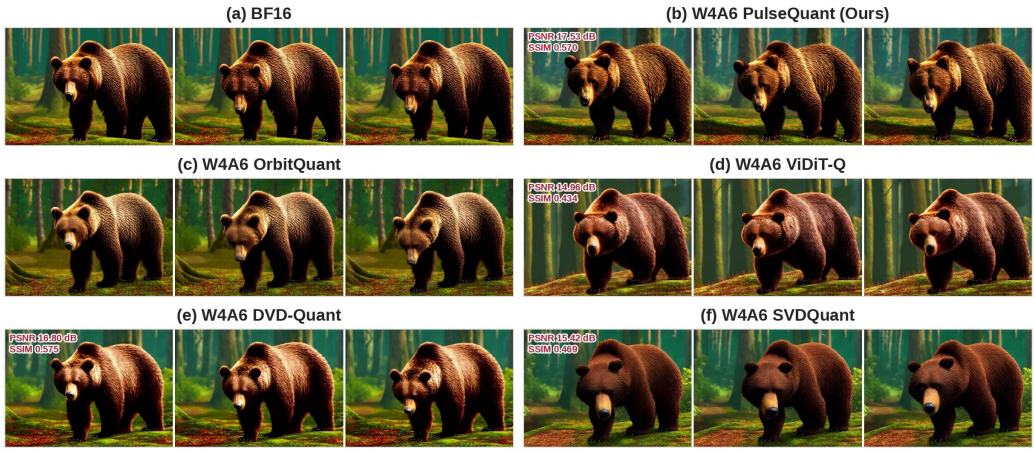
Text prompt: "A whimsical animated short film, a playful kitten with fluffy white fur and large expressive eyes, sitting contentedly on a small pedestal fountain in a quaint garden. The kitten gently laps at the water with its paw, its whiskers twitching with satisfaction. The garden is filled with vibrant greenery and colorful flowers, the sun casting a warm golden glow over everything. Soft, gentle lighting highlights the kitten's joyful expression. The background features a subtle, dreamy haze, adding depth and atmosphere. Hand-drawn cel animation style, slow-motion footage emphasizing the fluidity of the kitten's movements as it drinks. Wide shot of the garden with the kitten in the center, then transitioning to a medium shot focusing on the kitten's face and the fountain, capturing every detail of its expression and body language."

Figure 9: Wan 2.1-1.3B W4A4 comparison for an animated kitten at a garden fountain.



Text prompt: "Sunset time-lapse photography at the serene beach, capturing the vibrant colors and moving clouds in the sky. A young woman in a flowing sunset-colored sundress walks along the shoreline, her long blonde hair gently swaying with the gentle breeze. She holds a colorful beach umbrella aloft, its shade providing a cool spot under the scorching sun. The sky transitions from a deep orange to a fiery red, with wispy white clouds casting shadows across the ocean. The waves lap rhythmically against the sandy shore, creating a soothing sound. In the distance, a group of dolphins leaps gracefully in the water, adding a touch of joy to the serene scene. The beach is dotted with umbrellas and colorful beach chairs, and the sky is filled with warm hues of pink and purple as the sun begins its descent. The overall atmosphere is peaceful and beautiful, captured in a mesmerizing time-lapse sequence. Medium shot of the woman walking along the beach, wide shots of the sunset and the sea, and sweeping aerial views of the beach at dusk."

Figure 10: Wan 2.1-1.3B W4A4 comparison for a woman walking along a beach at sunset.



Text prompt: "CG game concept digital art, a large brown bear standing in a dense forest clearing, its fur textured to resemble moss-covered earth. The bear has a keen, focused expression, nostrils flaring as it sniffs the crisp autumn air, searching meticulously for any sign of nearby prey. Its body language is alert and ready, tail flicking gently from side to side. The forest background is lush and vibrant, with towering trees casting dappled shadows on the ground. In the distance, a few birds take flight. The bear's fur is covered in fine, realistic detailing, from the subtle patterns to the individual hairs. A low-angle, medium shot perspective captures the bear's determined gaze and the intricate textures of its fur. The overall scene exudes a sense of nature's mystery and the bear's innate instincts."

Figure 11: Wan 2.1-1.3B W4A6 comparison for a brown bear in a forest clearing.



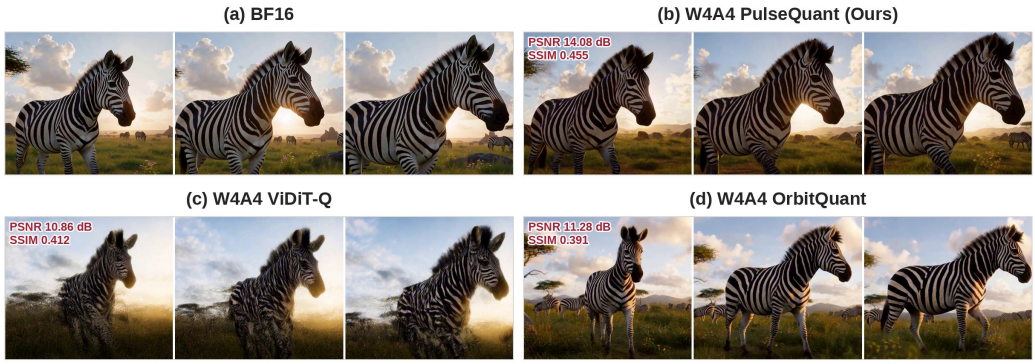
Text prompt: "CG game concept digital art, a tranquil tableau featuring a serene cup resting on a polished wooden surface. The cup is made of delicate porcelain, with intricate hand-painted designs in shades of pale blue and green. It sits gracefully on a small, intricately carved wooden coaster. The wooden surface is smooth and gleaming, with subtle grain patterns. Soft, warm ambient lighting casts gentle shadows, highlighting the intricate details of the cup and the coaster. The background is a minimalist, softly textured room with hints of pastel colors, creating a peaceful and serene atmosphere. Low-angle view, medium shot focusing on the detailed cup."

Figure 12: MiniMax-H3 W4A6 comparison for a porcelain cup on a carved wooden coaster.



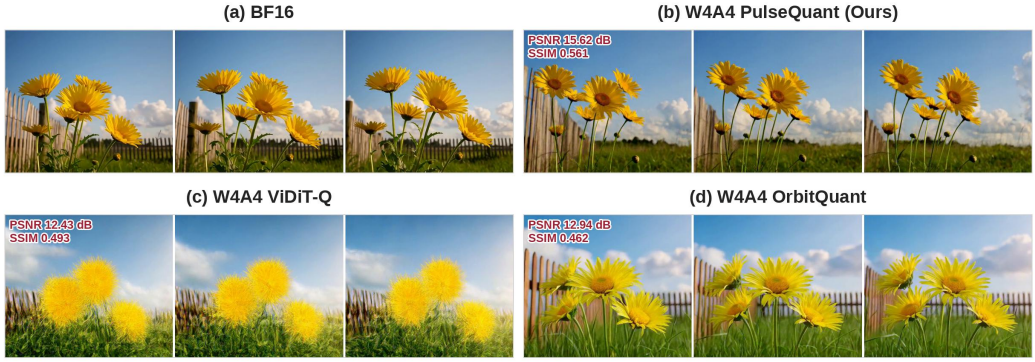
Text prompt: "Beachside scene captured during sunset, featuring a young woman with sun-kissed skin and tousled sandy blonde hair. She is wearing a flowing, turquoise-colored sundress with delicate floral patterns and a matching wide-brimmed straw hat adorned with a small flower. Her expressive brown eyes are framed by long lashes, and she has a serene yet playful smile. She is lounging on a plush, striped towel, positioned under a large umbrella providing shade. A calm ocean breeze rustles the palm leaves nearby. In the background, a picturesque coral reef can be seen reflecting the warm hues of the setting sun. The sky is painted with soft shades of orange and pink, blending seamlessly into the deep blue of the horizon. Soft, dreamy lighting and gentle waves create a tranquil atmosphere. The scene captures the essence of a leisurely summer day at the beach. Medium shot, half-body portrait, with focus on the woman's joyful expression and the serene environment. Gentle camera movement following the woman as she enjoys the moment."

Figure 13: MiniMax-H3 W4A6 comparison for a woman relaxing by the beach at sunset.



Text prompt: "CG game concept digital art, a majestic zebra with a sleek black and white striped pattern, taking a peaceful walk across a vast savanna. The zebra stands tall with its head held high, walking slowly through tall grass and sparse trees. The sun sets behind the zebra, casting a warm golden glow over the landscape. The sky is filled with fluffy white clouds, adding a soft and serene atmosphere. The zebra moves gracefully, with each step carefully measured, as it explores its surroundings. The savanna is dotted with small wildflowers and scattered with ancient-looking rocks. In the distance, a herd of other zebras can be seen grazing peacefully. The overall scene is tranquil and beautiful. Low-angle view, medium shot focusing on the zebra's face and posture."

Figure 14: MiniMax-H3 W4A4 comparison for a zebra walking across a sunlit savanna.



Text prompt: "Yellow flowers gently swaying in the breeze, their vibrant petals dancing with each gentle gust. The flowers are large and daisy-like, with soft yellow centers surrounded by feathery yellow petals. They sway gracefully in a field of green grass, with a slight breeze blowing from the right side. The sky is a clear, sunny blue, with fluffy white clouds floating lazily across the horizon. A small wooden fence frames the scene, adding a touch of rustic charm. The sun casts a warm glow over everything, creating a peaceful and serene atmosphere. Soft, natural lighting fills the frame, highlighting the beauty of the flowers and the lush green surroundings. The composition is balanced, with the flowers at the center, the fence on the left, and the sky and clouds on the right. The focus is on the movement of the flowers, captured with smooth, fluid camera movements. Medium shot, low-angle view."

Figure 15: MiniMax-H3 W4A4 comparison for yellow flowers swaying beside a wooden fence.