TRAJEQ: TRAJECTORY-PRESERVING FLOW MATCHING QUANTIZATION FOR LOW-BIT VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

While flow matching (FM) has emerged as a prominent paradigm for high-quality video generation, its significant computational overhead presents a major obstacle to practical implementation. Although post-training quantization (PTQ) is an established approach for efficient generation, existing PTQ methods primarily minimize reconstruction error in a magnitude-centric manner. This objective is misaligned with deterministic FM trajectories, where low-bit quantization introduces velocity-field orientation errors that can compound across integration steps and cause deviations from the full-precision trajectory. To address this issue, we introduce ***TrajeQ*** to preserve the geometric integrity of the velocity field and reduce trajectory drift under low-bit quantization. Our approach introduces *Curvature-Adaptive Quantization (CAQ)* to detect angular shifts across timesteps and mitigate geometric distortion via adaptive precision. We also integrate *Directional Bottleneck Preservation (DBP)*, a sensitivity-guided precision assignment scheme that retains only the most vulnerable layers in full precision. Experiments demonstrate that ***TrajeQ*** achieves state-of-the-art VBench performance and maintains stable video generation under extreme W4A4 and W3A6 quantization regimes where existing methods degrade severely. We provide the code in the supplementary material and additional video results on the Project Page.

1 INTRODUCTION

Text-to-Video generation has emerged as a prominent frontier in multimodal AI, yet the prohibitive computational costs render on-device deployment a critical challenge. Recently, Ordinary Differential Equation (ODE)-based Flow Matching (FM) has emerged as a dominant paradigm for high-quality video generation (Lipman et al., 2022; Jin et al., 2025). FM learns continuous-time velocity fields from prescribed probability paths, supporting high-quality generation through efficient ODE sampling. However, implementing billion-parameter FM models (Black Forest Labs et al., 2025) on edge devices necessitates Post-Training Quantization (PTQ) (Banner et al., 2019) for efficient model compression without retraining.

Crucially, we observe that naively transferring state-of-the-art (SOTA) PTQ methods designed for convolutional diffusion (Shang et al., 2023; Li et al., 2023; Ryu et al., 2025; Lee et al., 2025) or Diffusion Transformers (DiTs) (Wu et al., 2024; Chen et al., 2025) to FM causes severe performance degradation. Recent diffusion quantization methods, including temporal or mixed-precision strategies (Zhao et al., 2025; Huang et al., 2024a; Zhao et al., 2024) and dynamic efficiency approaches (Wu et al., 2025), still fail to mitigate the error accumulation of deterministic FM integration. Distillation-based Q-VDiT (Feng et al., 2025) follows a different optimization protocol and is discussed in Appendix A.2. These methods adapt precision using reconstruction- or sensitivity-centric criteria, whereas FM requires preserving the trajectory geometry of the velocity field.

We attribute this failure to a fundamental misalignment between standard quantization assumptions and the geometric nature of FM. Magnitude-centric quantization prioritizes minimizing Euclidean error (Krishnamoorthi, 2018; Xiao et al., 2023; Ashkboos et al., 2024), inducing angular distortions that deviate from the deterministic path.

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.

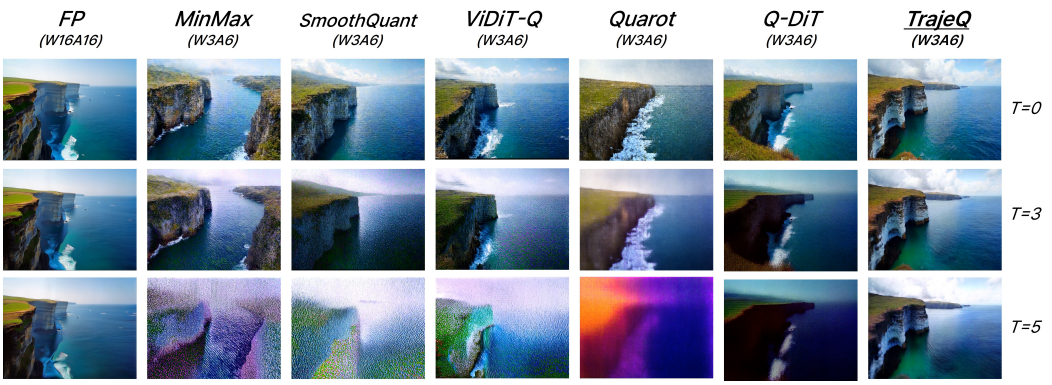


Figure 1: Qualitative comparison of sampled frames from a single video sequence under extreme W3A6 quantization.

Based on this analysis, we propose ***TrajeQ*** (Trajectory-Preserving Flow Matching Quantization), a novel framework designed to improve trajectory stability in video FM models. ***TrajeQ*** comprises two key techniques to suppress trajectory deviation as shown in Fig. 2.

First, we propose ***Curvature-Adaptive Quantization (CAQ)*** to preserve flow geometry through timestep-wise precision adaptation. Motivated by the relation between velocity turning and trajectory curvature, CAQ measures the cosine distance between consecutive activations and adds one bit when this score exceeds a threshold. We evaluate its agreement with high-turning FP timesteps and compare it with alternative allocation rules at the same selection count. These experiments support using activation-direction changes to guide temporal precision allocation.

Second, we introduce ***Directional Bottleneck Preservation (DBP)***, which converts *Directional Sensitivity* into a budgeted precision assignment problem. Instead of manually specifying sensitive layer types, DBP scores every quantizable layer using Directional Sensitivity and preserves only the top-ranked layers under a small FP16 parameter budget. Our analysis shows that embedding and adaLN layers are consistently selected by this ranking, indicating that they act as geometric bottlenecks for FM quantization.

Our key contributions are summarized as follows:

- **Trajectory-geometric analysis of FM quantization:** We introduce an analysis of PTQ for deterministic FM sampling by decomposing velocity perturbations into magnitude-aligned and orthogonal components. This analysis characterizes local directional distortion and the norm weighting of angular error in standard reconstruction-based PTQ objectives.
- **Curvature-motivated temporal precision allocation:** We propose *Curvature-Adaptive Quantization (CAQ)*, which uses inter-step activation-direction changes to allocate additional precision. We relate its geometric motivation to velocity turning and assess how closely the selected timesteps match high-turning FP timesteps. We also conduct allocation ablations while keeping the number of boosted timesteps fixed.
- **Direction-aware bottleneck preservation:** We define *Directional Sensitivity*, a geometry-aware metric for ranking layer vulnerability. It measures the stabilized ratio of mean angular error to mean squared error. Using this metric, we propose *Directional Bottleneck Preservation (DBP)* to retain original precision only for critical high-sensitivity layers under a small FP16 budget.
- **FM-tailored PTQ for video generation:** To the best of our knowledge, ***TrajeQ*** is the first PTQ framework designed around the trajectory geometry of deterministic FM sampling in video models. Trajectory diagnostics demonstrate reduced velocity angular error and trajectory deviation, while controlled ablations quantify the contributions of CAQ and DBP. It achieves SOTA VBench performance with stable generation under extreme low-bit W4A4 and W3A6 regimes.

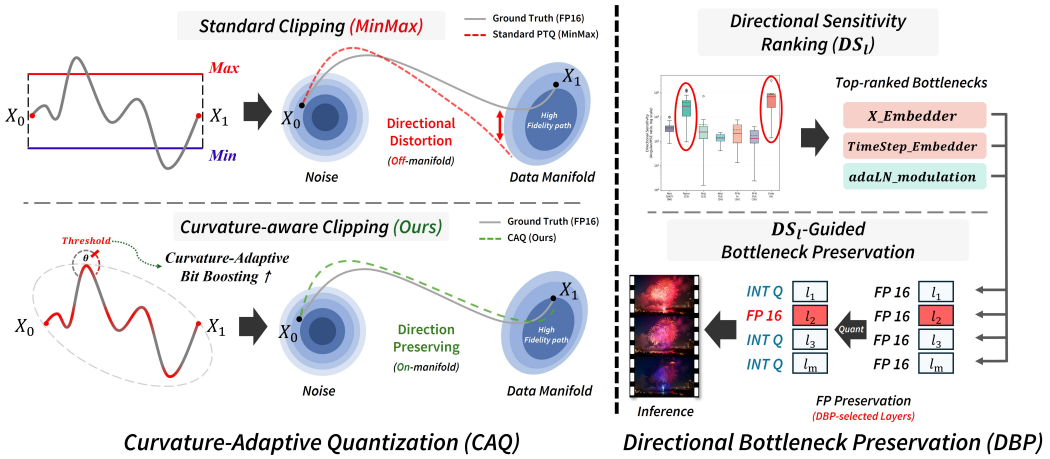


Figure 2: Overview of ***TrajeQ***. **Left (CAQ):** *Curvature-Adaptive Quantization* uses the activation-based score S_{dir} to select timesteps for bit boosting ($b \uparrow$). This selective allocation reduces directional distortion and trajectory drift. **Right (DBP):** *Directional Bottleneck Preservation* ranks quantizable layers using *Directional Sensitivity* (DS_l) and retains selected high-sensitivity layers in FP16 to mitigate trajectory drift.

2 Rethinking Quantization for Flow Matching

This section revisits PTQ for FM from a trajectory-geometric perspective. We first analyze how parallel and orthogonal velocity perturbations affect local updates under deterministic sampling. We then examine how reconstruction-based PTQ objectives weight angular error and motivate precision allocation using directional information.

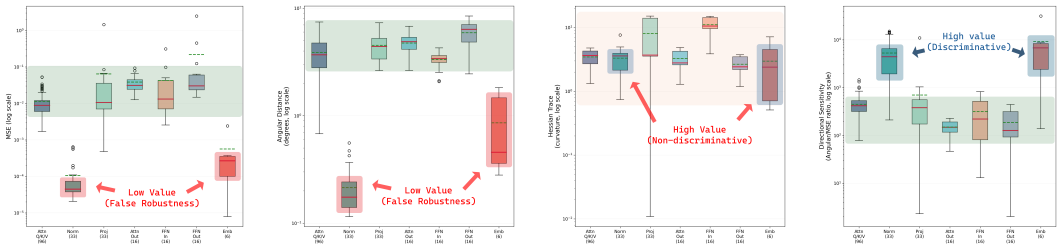
2.1 Directional Error Accumulation in Deterministic Sampling

FM generates samples by integrating a learned velocity field through deterministic ODE dynamics, a regime also adopted by deterministic samplers such as DDIM and DPM-Solver. In this regime, each state update depends directly on the quantized velocity prediction. For an Euler-type step, $\mathbf{x}_{t+\Delta t} = \mathbf{x}_t + \Delta t(\mathbf{v}_t + \boldsymbol{\delta}_v)$, the perturbation $\boldsymbol{\delta}_v$ shifts the current state and consequently changes the point at which subsequent velocities are evaluated. Quantization error is therefore propagated through the integration path rather than remaining a local one-step perturbation. Our analysis focuses on error propagation under deterministic sampling.

To characterize which propagated errors alter the trajectory geometry, we decompose $\boldsymbol{\delta}_v$ into components parallel and orthogonal to the full-precision velocity $\mathbf{v}$:

$$\boldsymbol{\delta}_v = \underbrace{\boldsymbol{\delta}_v^{\parallel}}_{\text{Magnitude}} + \underbrace{\boldsymbol{\delta}_v^{\perp}}_{\text{Direction}} \quad (\text{where } \langle \boldsymbol{\delta}_v^{\perp}, \mathbf{v} \rangle = 0). \tag{1}$$

For sufficiently small perturbations, the parallel component $\boldsymbol{\delta}_v^{\parallel}$ modulates the velocity magnitude along $\mathbf{v}$, primarily changing the step length while preserving the local update direction. To first order, this affects how far the sample moves along the current velocity direction, rather than rotating the local update. Thus, $\boldsymbol{\delta}_v^{\parallel}$ mainly induces a local step-length error. In contrast, the orthogonal component $\boldsymbol{\delta}_v^{\perp}$ lies perpendicular to $\mathbf{v}$ and directly changes the velocity orientation. This rotates the local update direction and shifts the next state laterally away from the full-precision trajectory. Because subsequent velocity predictions are evaluated at this shifted state, the deviation changes not only the current update but also the future vector field encountered by the sampler. In deterministic ODE-based sampling, such orientation errors can persist and compound across the integration horizon, with their effect growing over sampling steps. This accumulation exposes a limitation of standard PTQ objectives, which weight angular error by the vector norms and may leave velocity orientation insufficiently controlled. Accordingly, the quantization objective should explicitly preserve the direction of the velocity field.



(a) MSE (b) Angular Distance (c) Hessian Trace (d) DS_l (Ours)

Figure 3: Layer-wise sensitivity analysis comparing **(a-c) conventional metrics** against our **(d) Directional Sensitivity** (DS_l). `Norm` and `Embedding` layers have low absolute errors in (a,b) but high DS_l, which uses the stabilized ratio of mean cosine error to mean squared L_2 error. Preservation effects are evaluated in Appendix E.4.

2.2 Magnitude vs. Direction: A Quantization Strategy Gap

We now examine how standard PTQ objectives account for the orthogonal component identified in Sec. 2.1. Most PTQ methods optimize quantized weights by reconstructing the full-precision layer output, minimizing $\min_{\hat{\mathbf{W}}} \|(\mathbf{W}-\hat{\mathbf{W}})\mathbf{X}\|_F^2$ for input activations $\mathbf{X}$. While effective for reconstructing layer outputs, this objective does not explicitly distinguish velocity-aligned scaling from orientation-changing perturbations. In other words, two quantization errors with similar reconstruction loss can have very different projections onto the parallel and orthogonal components of the velocity field. This reconstruction loss can be expressed in covariance-weighted form:

$$\|(\mathbf{W}-\hat{\mathbf{W}})\mathbf{X}\|_F^2 = \mathrm{Tr}\Big((\mathbf{W}-\hat{\mathbf{W}})\mathbf{\Sigma}_{\mathbf{X}}(\mathbf{W}-\hat{\mathbf{W}})^\top\Big), \quad \mathbf{\Sigma}_{\mathbf{X}} = \mathbf{X}\mathbf{X}^\top. \tag{2}$$

This form implies that reconstruction-based PTQ mainly penalizes errors along high-energy input directions, which correspond to dominant eigenspaces of $\mathbf{\Sigma}_{\mathbf{X}}$. However, the trajectory impact of a quantization error depends not only on its absolute energy, but also on how much it rotates the velocity direction. As a result, directional drift can remain insufficiently controlled when low-norm activations undergo large orientation changes. We identify two limitations when applying reconstruction-based PTQ metrics to ODE-sampled models:

- **Directional Blindness:** Reconstruction-based and Hessian-based metrics do not generally isolate velocity-orthogonal error $\boldsymbol{\delta}_v^\perp$ from velocity-aligned scaling $\boldsymbol{\delta}_v^\parallel$. Although these metrics can depend on perturbation direction, MSE weights angular error by the vector norms, while Hessian-based scores reflect the chosen local objective (Appendix B.4). Angular distance and Hessian trace in Figs. 3b and 3c rank layers differently from DS_l.
- **Misidentification of Sensitive Layers:** Eq. (2) weights reconstruction error by feature magnitude. Conversely, our *Directional Sensitivity* ranks layers by the stabilized ratio of mean angular error to mean squared error. Thus, high DS_l can indicate severe directional distortion even for low-magnitude features, resulting in a distinct layer ranking compared to MSE (Figs. 3a and 3d).

3 Methodology

We propose ***TrajeQ*** (Trajectory-Preserving Flow Matching Quantization), a PTQ framework to preserve the geometric trajectory of FM. ***TrajeQ*** integrates two core components, namely *Curvature-Adaptive Quantization (CAQ)* to minimize directional distortion and *Directional Bottleneck Preservation (DBP)* to preserve structural bottlenecks (Fig. 2, details in Appendix C).

3.1 Curvature-Adaptive Quantization (CAQ)

To mitigate the directional blindness identified in Sec. 2.2, we introduce *Curvature-Adaptive Quantization (CAQ)*. We hypothesize that, given a limited precision budget, prioritizing higher activation bit-widths at sharply turning timesteps improves generation quality. CAQ targets these steps via a directional score, S_{dir}, assigning higher precision and keeping the rest at base precision.

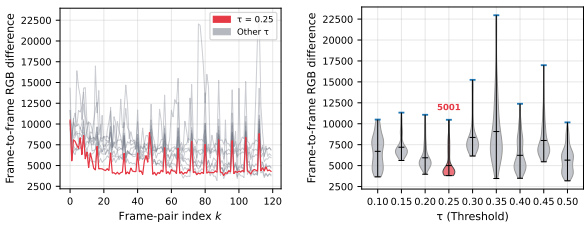
3.1.1 Activation Direction for Timestep Selection

We first explain why trajectory turning is relevant to precision allocation, then examine whether activation-direction changes can identify sharply turning timesteps. The following lemma relates velocity turning to trajectory curvature.

Lemma 1 (Curvature Consistency of the Turning Angle). Let $\mathbf{x}(t)$ be a C^3 reference trajectory within a fixed sampling stage, with $\mathbf{v}(t) = \dot{\mathbf{x}}(t)$ and nonzero velocity in a neighborhood of t. Let $\kappa(t)$ be the trajectory curvature and θ_t represent the angle between consecutive velocity vectors $\mathbf{v}(t)$ and $\mathbf{v}(t - \Delta t)$. As $\Delta t \to 0$, we have:

$$1 - \cos\theta_t = \frac{1}{2}\kappa(t)^2\|\mathbf{v}(t)\|_2^2(\Delta t)^2 + \mathcal{O}((\Delta t)^3). \tag{3}$$

Lemma 1 establishes the link between the angle formed by consecutive velocity vectors and trajectory curvature. Assuming comparable velocity magnitudes and step sizes, a greater directional shift indicates higher local curvature. Given that directional errors can propagate across sampling steps (Sec. 2.1), we allocate additional activation precision to these regions of high curvature to better preserve the trajectory geometry.



(a) RGB Differences (b) Threshold Distributions

Figure 4: Frame-to-frame RGB differences. (a) Differences between consecutive decoded frames. (b) Distributions across frame pairs for each threshold. Among the tested thresholds, $\tau = 0.25$ has the lowest mean difference.

However, because FP reference velocities are not directly available during quantized inference, we instead compare the current activation with its cached value from the preceding timestep. This layer-wise directional shift serves as an effective proxy for precision allocation. We therefore propose S_{dir} as a proxy for timestep selection, computed as the cosine distance between consecutive layer activations:

$$S_{dir}^{(l)}(t) = 1 - \frac{\langle \mathbf{h}_l(t), \mathbf{h}_l(t - \Delta t)\rangle}{\|\mathbf{h}_l(t)\|_2\|\mathbf{h}_l(t - \Delta t)\|_2}. \tag{4}$$

Here, $\mathbf{h}_l(t)$ denotes the flattened activation at layer l and timestep t. The link between this activation score and trajectory evolution is locally characterized in Lemma 3, which details how ODE state dynamics and timestep conditioning affect activation direction. Empirically, S_{dir} aligns strongly with the highest-turning FP timesteps, outperforming random selection under an equivalent budget (Tab. 6 in Appendix B.7). This allocation strategy yields superior generation quality compared to alternative rules (Tab. 3). Beyond these quality improvements, the full ***TrajeQ*** framework also reduces velocity angular error and trajectory deviation (Tab. 5).

3.1.2 Curvature-Adaptive Bit Boosting

At each quantized layer, CAQ uses the activation-direction score $S_{dir}(t)$ to select timesteps for *bit boosting*, increasing the activation bit-width when this score exceeds the threshold τ:

$$B_t = b_{\text{base}} + \Delta b \cdot \mathcal{H}(S_{dir}(t) - \tau), \quad \text{where } \mathcal{H}(z) = \mathbb{I}(z > 0). \tag{5}$$

Here, B_t is the selected activation bit-width, b_{base} is the base bit-width, and $\Delta b = 1$. The bit-width is determined independently for each layer and timestep and applies uniformly to the entire activation tensor. Under a fixed-range, high-resolution uniform-error model, one additional bit reduces granular noise variance by approximately 75% (Appendix B.5).

We use $\tau = 0.25$ based on threshold analyses across multiple bit-widths, sampling regimes, and output resolutions (Appendix E.1). This threshold achieves the highest CLIPSIM, VQA-Aes, and VQA-Tech scores in the W3A6 sweep and maintains competitive generation quality under 4-step distilled sampling and across the tested resolutions without additional tuning. As a complementary diagnostic, Fig. 4 reports L_2 differences between consecutive decoded RGB frames. Among the tested thresholds, $\tau = 0.25$ gives the lowest mean difference for the evaluated prompt.

3.2 DIRECTIONAL BOTTLENECK PRESERVATION (DBP)

While CAQ adapts bits across timesteps, *Directional Bottleneck Preservation (DBP)* protects structural bottlenecks in the vector field. We introduce *Directional Sensitivity* (DS), a metric capturing angular distortion relative to squared error. DBP ranks layers by DS and preserves only high-sensitivity layers in FP16, maintaining quality with minimal overhead.

3.2.1 SENSITIVE-LAYER ANALYSIS VIA DIRECTIONAL SENSITIVITY

To pinpoint structural bottlenecks, we analyze the sensitivity of feature orientations relative to their activation magnitudes. Standard quantization typically relies on Mean Squared Error (MSE). However, the same small orthogonal perturbation causes larger local angular changes in features with smaller activation norms. We formalize this dependence as follows.

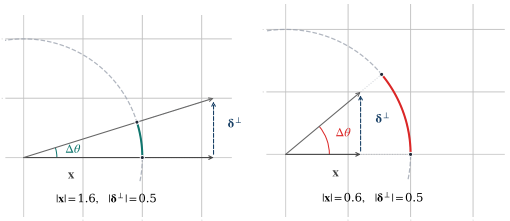
Lemma 2 (Magnitude-Induced Instability). Let $\mathbf{x} \in \mathbb{R}^d \setminus \{0\}$ be a fixed feature activation and $f(\mathbf{x}) = \mathbf{x}/\|\mathbf{x}\|$ denote its projection onto the unit hypersphere.

For the orthogonal component $\boldsymbol{\delta}^\perp$ of a perturbation $\boldsymbol{\delta} = \boldsymbol{\delta}^\parallel + \boldsymbol{\delta}^\perp$, with $\|\boldsymbol{\delta}\|/\|\mathbf{x}\| \to 0$, the normalized displacement satisfies:

$$\|f(\mathbf{x} + \boldsymbol{\delta}^\perp) - f(\mathbf{x})\| = \frac{1}{\|\mathbf{x}\|}\|\boldsymbol{\delta}^\perp\| + \mathcal{O}(\|\boldsymbol{\delta}^\perp\|^2). \tag{6}$$

Lemma 2 shows that inverse magnitude $1/\|\mathbf{x}\|$ scales the local directional response to a small absolute perturbation. For nonzero FP and layer-wise quantized activations, let $a_l^{t,c} = 1 - \cos(\mathbf{x}_l^{t,c}, \hat{\mathbf{x}}_l^{t,c})$ and $e_l^{t,c} = \|\hat{\mathbf{x}}_l^{t,c} - \mathbf{x}_l^{t,c}\|_2^2$ denote angular and squared errors, respectively. We define *Directional Sensitivity* as:

$$DS_l = \frac{\mathbb{E}_{c,t}[a_l^{t,c}]}{\mathbb{E}_{c,t}[e_l^{t,c}] + \epsilon}. \tag{7}$$



(a) Large $|\mathbf{x}|$ *(Stable)* (b) Small $|\mathbf{x}|$ *(Unstable)*

Figure 5: Impact of Vector Magnitude on Directional Sensitivity. Identical orthogonal perturbation $\boldsymbol{\delta}^\perp$ causes a substantially larger angular deviation $\Delta\theta$ in the smaller vector (b) than in (a).

Here, $\mathbb{E}_{c,t}$ denotes the empirical mean over profiling prompts $\mathcal{C}$ and timesteps $\mathcal{T}$, and $\epsilon > 0$ stabilizes the denominator. As derived in Appendix B.6, local angular error depends on the orthogonal perturbation relative to activation magnitude. We compute the mean angular and squared errors over a small profiling set before inference and use their stabilized ratio to score each layer. We then rank layers by DS_l and preserve layers under a fixed FP16 budget, rather than manually specifying layer types. As illustrated in Fig. 5, the same small orthogonal perturbation induces greater angular shifts in low-magnitude features, while DS_l is computed from observed quantization errors.

3.2.2 PRESERVATION OF GEOMETRIC BOTTLENECKS

Our DS_l-based scheme formulates Directional Bottleneck Preservation as a budgeted precision assignment problem. Given the layer-wise scores, we rank all quantizable layers by DS_l and retain only the top-sensitive subset under a fixed FP16 parameter budget. As standard mitigations (e.g., clamping, mean imputation) fail to correct this directional distortion (Appendix E.4), DBP preserves the layers with the highest DS_l scores. In our analysis, the selected layers are consistently concentrated in two geometric bottlenecks:

- *Embedding Layers:* Their high DS_l reflects a high stabilized ratio of mean angular error to mean squared error. Errors in these input representations can affect the predicted velocity field and subsequent trajectory evolution.
- *Normalization Layers:* Operating as continuous steering mechanisms, `adaLN` layers govern flow trajectories. Low-bit quantization in these modules noticeably degrades conditional guidance.

By allocating full precision only to the high-DS_l layers selected under this budget, which account for less than 1% of the total parameters, we suppress quantization-induced distortion in these layers while enabling extreme quantization elsewhere with minimal computational overhead. Appendix E.3 compares layer-selection rules under the same FP16 parameter budget.

Table 1: Evaluation of text-to-video generation on VBench with FLUX. The "16" bit-width represents full precision (FP16). **Bold** is the best result, and underline is second-best result. "N/A" denotes measurement failure due to severe noise (see Appendix G.2).

Method	Bit-width (W/A)	Imaging Quality	Aesthetic Quality	Motion Smooth.	Dynamic Degree	BG. Consist.	Subject Consist.	Scene Consist.	Overall Consist.
-	16/16	64.64	64.97	99.18	54.00	97.64	95.58	44.97	26.69
MinMax	6/6	63.94	63.91	98.92	45.79	96.53	92.95	43.16	**26.39**
SmoothQuant	6/6	63.79	62.95	98.97	45.83	96.86	93.76	43.29	25.95
QuaRot	6/6	60.93	62.86	98.18	41.66	94.96	92.85	44.18	26.17
Q-DiT	6/6	63.36	64.74	99.01	48.64	96.88	94.26	41.74	26.20
ViDiT-Q	6/6	62.77	64.06	99.04	45.85	96.93	94.12	42.15	**26.39**
TrajeQ	6/6	**64.08**	**64.90**	**99.08**	**52.77**	**96.98**	**94.51**	**44.84**	26.32
MinMax	4/6	61.88	63.73	98.86	41.38	96.00	92.14	41.58	26.26
SmoothQuant	4/6	59.80	59.62	98.90	33.33	94.26	90.67	42.24	25.32
QuaRot	4/6	55.03	59.50	98.41	35.27	91.02	91.32	42.65	25.77
Q-DiT	4/6	61.51	62.70	99.02	41.61	96.12	93.09	43.84	25.72
ViDiT-Q	4/6	61.40	63.43	98.92	**43.05**	96.75	93.92	41.07	25.98
TrajeQ	4/6	**62.39**	**64.05**	**99.04**	42.50	**97.13**	**94.37**	**44.14**	**26.28**
MinMax	4/4	59.81	31.51	N/A	N/A	N/A	N/A	00.00	03.29
SmoothQuant	4/4	59.30	30.60	N/A	N/A	N/A	N/A	00.21	03.28
QuaRot	4/4	51.56	40.11	94.14	11.44	86.02	64.71	23.47	19.69
Q-DiT	4/4	60.19	56.51	96.27	20.27	93.83	85.22	38.73	25.02
ViDiT-Q	4/4	61.17	55.80	97.72	24.72	93.05	86.64	40.84	24.56
TrajeQ	4/4	**63.23**	**59.10**	**97.85**	**26.38**	**94.10**	**87.21**	**40.91**	**25.30**
MinMax	3/6	54.36	44.54	97.60	12.50	88.98	78.03	28.63	21.56
SmoothQuant	3/6	53.37	44.86	97.57	08.33	88.81	79.24	26.52	21.53
QuaRot	3/6	33.72	39.67	98.82	10.13	91.57	78.39	10.17	18.28
Q-DiT	3/6	49.02	48.82	98.83	19.21	88.68	78.93	31.06	22.74
ViDiT-Q	3/6	53.64	47.60	97.93	16.44	89.74	81.39	25.36	22.39
TrajeQ	3/6	**58.06**	**57.00**	**98.98**	**19.44**	**93.48**	**90.23**	**38.29**	**24.91**

4 EXPERIMENTS

4.1 EXPERIMENTAL SETTINGS AND IMPLEMENTATIONS

Implementation Details. We conduct our primary experiments using Pyramid Flow miniFLUX (768p variant) (Jin et al., 2025) on an NVIDIA A100 GPU. To demonstrate generalizability, we also extend our evaluation to the MMDiT architecture (Esser et al., 2024), with detailed results provided in Appendix F.1. Implementation details are listed in Appendix D.

Baselines. To evaluate ***TrajeQ***, we compare it against two categories of quantization frameworks: general-purpose PTQ methods (MinMax (Krishnamoorthi, 2018), SmoothQuant (Xiao et al., 2023), and QuaRot (Ashkboos et al., 2024)) and SOTA diffusion-specific methods (Q-DiT (Chen et al., 2025) and ViDiT-Q (Zhao et al., 2025)). In the main comparisons, all baselines retain the same layers in FP16 as ***TrajeQ***, including the DBP-selected set.

Quantization and Evaluation. We implement PTQ in PyTorch (Paszke et al., 2019), with quantization scales computed on-the-fly during inference without offline calibration. DBP uses a small offline profiling set prior to inference. We adopt symmetric per-channel quantization for weights and symmetric per-token quantization for activations (Yao et al., 2022; Dettmers et al., 2022). We evaluate simulated quantization using floating-point operations. We evaluate W6A6, W4A6, W4A4, and W3A6 configurations with a CAQ threshold of $\tau = 0.25$ (Appendix E.1). Generation quality is measured on VBench (Huang et al., 2024b) and EvalCrafter (Liu et al., 2024), alongside efficiency metrics (Appendix D).

4.2 QUANTITATIVE EVALUATION

Tab. 1 summarizes performance across bit-widths. At W6A6, ***TrajeQ*** closes most of the gap to FP16 (Overall Consistency 26.32 vs. 26.69), indicating near-lossless compression in the moderate low-bit regime. Under W4A4, where standard PTQ methods often collapse or suffer severe degradation (N/A; Appendix G.2), ***TrajeQ*** preserves usable generation quality, retaining 59.10 Aesthetic Quality and 25.30 Overall Consistency. At W3A6, ***TrajeQ*** provides the largest gains on perceptual and semantic consistency metrics, improving Subject Consistency from 81.39 to 90.23 and Scene Consistency from 31.06 to 38.29 over the strongest baselines.

Tab. 2 reports performance on OpenSora prompts, further demonstrating the robustness of ***TrajeQ*** under diverse text-to-video scenarios. ***TrajeQ*** improves text-video alignment, perceptual quality, and temporal stability in most evaluated low-bit settings. At W6A6, it records the highest CLIPSIM

Table 2: Evaluation of text-to-video generation on OpenSora prompts with FLUX. The "16" bit-width represents full-precision (FP16). **Bold** indicates the best result, and underline indicates the second-best result. The symbol ↓ indicates that lower values are better.

Method	Bit-width (W/A)	CLIPSIM	CLIP-Temp	VQA-Aesthetic	VQA-Technical	△ Flow Score (↓)	Warping Error (↓)
-	16/16	0.3394	0.9983	64.96	64.08	-	0.022
MinMax	6/6	0.3275	0.9973	60.97	62.68	0.222	0.031
SmoothQuant	6/6	0.3322	0.9956	59.60	61.43	0.928	0.039
QuaRot	6/6	0.3300	0.9967	61.81	62.19	0.505	0.026
Q-DiT	6/6	0.3309	0.9975	62.95	63.26	0.489	0.033
ViDiT-Q	6/6	0.3244	0.9975	61.95	62.29	0.140	0.025
TrajeQ	6/6	**0.3351**	**0.9979**	**64.41**	**64.05**	**0.137**	**0.023**
MinMax	4/6	0.3227	0.9964	60.45	62.19	0.895	0.032
SmoothQuant	4/6	0.3304	0.9954	60.97	59.00	0.086	0.026
QuaRot	4/6	0.3250	0.9973	59.57	56.61	0.243	0.030
Q-DiT	4/6	0.3302	0.9974	60.89	57.31	0.169	0.026
ViDiT-Q	4/6	0.3210	0.9970	61.91	58.80	0.096	0.026
TrajeQ	4/6	**0.3313**	**0.9977**	**63.32**	**63.91**	**0.076**	**0.025**
MinMax	4/4	0.2009	0.9912	32.38	41.80	0.176	0.090
SmoothQuant	4/4	0.2048	0.9928	33.21	40.60	0.219	0.097
QuaRot	4/4	0.3240	0.9939	54.50	46.06	0.371	0.067
Q-DiT	4/4	0.3173	0.9938	56.69	48.99	0.297	0.036
ViDiT-Q	4/4	0.3176	0.9944	55.66	47.62	0.101	**0.033**
TrajeQ	4/4	**0.3256**	**0.9951**	**57.05**	**56.54**	**0.068**	0.035
MinMax	3/6	0.2907	0.9929	48.70	51.93	0.150	0.057
SmoothQuant	3/6	0.2823	0.9942	47.87	52.67	0.317	0.054
QuaRot	3/6	0.2379	0.9717	35.67	40.44	1.141	0.090
Q-DiT	3/6	0.2739	0.9901	47.63	42.26	0.212	0.035
ViDiT-Q	3/6	0.2706	0.9921	49.47	48.80	0.445	0.059
TrajeQ	3/6	**0.3186**	**0.9971**	**61.40**	**59.29**	**0.092**	**0.029**

Table 3: Ablation of W3A6 bit-boosting on OpenSora prompts. All boosting strategies select the same number of timesteps. S_{dir}-guided allocation achieves the highest scores on all four metrics.

Strategy	Bit-Boosting	CLIPSIM	CLIP-Temp	VQA-Aes	VQA-Tech
No-boosting (base bit only)	X	0.3043	0.9949	57.02	53.26
Random boosting (+1bit)	Random	0.3159	0.9968	56.34	53.18
Δ-Magnitude	$\|v_t - v_{t-\Delta t}\|_2$	0.3166	0.9968	59.17	56.20
Quant-MSE	$\|v - \hat{v}\|^2$	0.3155	0.9966	57.67	55.02
S_{dir} **guided (*TrajeQ*)**	**Based on** S_{dir}	**0.3186**	**0.9971**	**61.40**	**59.29**

(0.3351) and VQA-Aesthetic (64.41), effectively matching FP quality. In the W4A4 setting, while baselines suffer significant degradation, our method maintains robust generation quality with a VQA-Technical score of 56.54. Notably, in the extreme W3A6 configuration, ***TrajeQ*** outperforms the nearest competitor by a wide margin in both *Warping Error* (0.029 vs. 0.035) and *VQA-Aesthetic* (61.40 vs. 49.47). Therefore, ***TrajeQ*** maintains visual coherence and motion stability even under severe bit-width constraints, where competing methods exhibit pronounced degradation. ***TrajeQ*** at nominal W3A6 outperforms ViDiT-Q at W3A8 on all four metrics (Appendix E.5).

4.3 Ablation Study

Quantitative Component Analysis.

As shown in Tab. 4, standalone **CAQ** (Row 2) yields substantial gains, especially in *Dynamic Degree* ($+1.34$) and *Imaging Quality* ($+1.10$), mitigating directional distortions in visual and motion fidelity. However, this gain trades off contextual stability (-1.51 *Scene Consist.*). Conversely, applying **DBP** independently (Row 3) directly targets structural integrity. By guarding sensitive layers, DBP stabilizes global structure against trajectory distortion, ensuring robust consistency ($+1.30$ *Scene Consist.*). This structural stability enables aggressive temporal dynamics without collapse, driving a substantial *Dynamic Degree* gain of $+4.12$. Detailed analyses of DBP are provided in Appendix E.2 and Tab. 11.

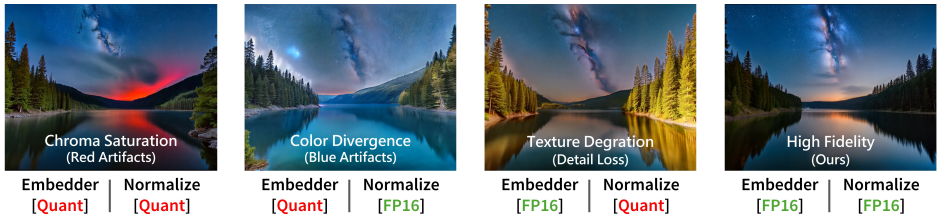
The complete **CAQ + DBP** framework integrates these complementary strengths. By mitigating CAQ's contextual trade-off while maximizing visual and dynamic fidelity ($+6.89$ *Dyn. D.*), the combined approach achieves SOTA results across seven of eight metrics. Overall, these results validate that directional correction and structural preservation are both essential for robust low-bit video generation. Comparisons with alternative allocation heuristics illustrate that S_{dir}-guided allocation is central to CAQ's effectiveness (Tab. 3).

Table 4: W4A6 VBench ablation over CAQ and DBP. The first row denotes Base PTQ. Top and bottom rows report quality and consistency metrics. Superscripts denote gains over Base PTQ.

Method		Img. Q BG C.	Aes. Q Sub. C	Mot. S Scn. C	Dyn. D All. C
CAQ	DBP	*(Top: Quality / Bottom: Consistency)*			
		61.03 96.99	63.25 93.95	98.69 41.45	35.61 26.09
✓		$\underline{62.13}^{+1.10}$ $96.38^{-0.61}$	$63.66^{+0.41}$ $94.12^{+0.17}$	$99.02^{+0.33}$ $39.94^{-1.51}$	$36.95^{+1.34}$ $26.17^{+0.08}$
	✓	$61.52^{+0.49}$ $\underline{97.02}^{+0.03}$	$\underline{63.72}^{+0.47}$ $\mathbf{94.89}^{+0.94}$	$\mathbf{99.17}^{+0.48}$ $\underline{42.75}^{+1.30}$	$\underline{39.73}^{+4.12}$ $\mathbf{26.53}^{+0.44}$
✓	✓	$\mathbf{62.39}^{+1.36}$ $\mathbf{97.13}^{+0.14}$	$\mathbf{64.05}^{+0.80}$ $\underline{94.37}^{+0.42}$	$\underline{99.04}^{+0.35}$ $\mathbf{44.14}^{+2.69}$	$\mathbf{42.50}^{+6.89}$ $\underline{26.28}^{+0.19}$

Table 5: Velocity angular error and normalized trajectory deviation on FLUX/Pyramid (510 ODE steps $\times$ 10 OpenSora prompts). Lower is better.

Bit-width	Method	Velocity angular error (°)	Mean trajectory deviation	Final-step deviation
W4A4	MinMax	37.59 ± 10.36	0.624	1.457 ± 0.219
W4A4	Q-DiT	33.72 ± 12.15	0.572	1.306 ± 0.234
W4A4	ViDiT-Q	27.91 ± 10.79	0.537	1.298 ± 0.276
W4A4	***TrajeQ(w/o CAQ)***	24.82 ± 9.34	0.467	1.226 ± 0.246
W4A4	***TrajeQ(w/o DBP)***	22.86 ± 9.12	0.399	1.166 ± 0.282
W4A4	***TrajeQ***	20.13 ± 8.63	0.385	1.143 ± 0.224
W3A6	MinMax	34.62 ± 11.00	0.749	1.351 ± 0.238
W3A6	Q-DiT	28.64 ± 9.72	0.813	1.299 ± 0.271
W3A6	ViDiT-Q	19.33 ± 12.61	0.755	1.307 ± 0.216
W3A6	***TrajeQ(w/o CAQ)***	16.72 ± 5.23	0.523	1.265 ± 0.201
W3A6	***TrajeQ(w/o DBP)***	14.82 ± 5.15	0.490	1.243 ± 0.223
W3A6	***TrajeQ***	10.96 ± 4.60	0.452	1.119 ± 0.178



(a) (b) (c) (d)

Figure 6: Visual ablation of structural bottlenecks presenting a qualitative comparison under W4A6 quantization: (a) Quantizing both Embedder and Normalization layers, (b) Quantizing Embedder only, (c) Quantizing Normalization only, and (d) Ours (Preserving both in FP16).

Velocity Angular Error and Trajectory Deviation.

To isolate local angular error from rollout drift, we compare FP and quantized velocities at identical FP states. We also measure normalized trajectory deviation, $\|\mathbf{x}_t^Q - \mathbf{x}_t^{\mathrm{FP}}\|_2 / \|\mathbf{x}_t^{\mathrm{FP}}\|_2$, at each step of independent rollouts with identical prompts and seeds. Tab. 5 demonstrates that ***TrajeQ*** achieves the lowest angular error and trajectory deviations at both bit-widths. Together with the generation quality assessments in Sec. 4.2, these observations provide empirical support for the trajectory-preserving motivation of ***TrajeQ***. See Appendix B.7 for S_{dir} selection agreement with FP velocity turning.

Qualitative Analysis and Efficiency.

Fig. 6 visually corroborates the necessity of DBP. Quantizing Embedding layers (X_{emb}) triggers severe chromatic divergence, whereas quantizing Normalization layers ($adaLN$) induces texture washout. These examples suggest that preserving embedding and normalization layers helps maintain color fidelity and texture detail, respectively. ***TrajeQ***'s DBP strategy effectively protects both, ensuring chromatic fidelity and structural sharpness. Further analyses are detailed in Appendices G.2 and G.3.

Regarding computational efficiency, further analysis in Appendix G.1 notes that ***TrajeQ*** requires higher peak memory and a 5% generation-time overhead compared to ViDiT-Q. Despite these costs, ***TrajeQ*** rigorously preserves generation fidelity under W3A6, a challenging regime where existing methods struggle to maintain generation quality.

5 Conclusion

We propose ***TrajeQ***, a quantization framework that preserves trajectory stability in FM-based video generation. To achieve this, our approach uses directional information to guide precision allocation during deterministic ODE sampling. Our analysis reveals the impact of orthogonal velocity perturbations on the local update direction. This insight directly motivates our two complementary mechanisms. *CAQ* performs curvature-adaptive bit boosting, while *DBP* provides sensitivity-guided structural preservation. Through this, ***TrajeQ*** maintains generation stability under aggressive quantization regimes where conventional baselines suffer severe degradation. Our current analysis does not fully characterize how quantization alters internal feature geometry. Future work exploring this aspect could guide the development of methods that jointly preserve trajectory stability and representation structure.

AI USE STATEMENT

We used generative AI tools to assist with translation and improve the clarity of the manuscript. We also sought feedback on our methodology and experimental design and used these tools to review our theoretical analysis and proofs for consistency. The theoretical framework, mathematical claims, and proofs were developed by the authors, with AI involvement confined to consistency checks and feedback. We did not use AI to originate research concepts or hypotheses, implement methods, run experiments, interpret results, or formulate conclusions. Nor did we use AI for synthetic dataset generation, dataset cleaning or reformatting, or qualitative or thematic data analysis. All AI-assisted suggestions were evaluated by the authors before adoption.

REPRODUCIBILITY STATEMENT

Appendix D contains comprehensive experimental details. The source code is provided in the supplementary material and will be publicly released on GitHub.

REFERENCES

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.

Ron Banner, Yury Nahshan, and Daniel Soudry. Post training 4-bit quantization of convolutional networks for rapid-deployment. *Advances in neural information processing systems*, 32, 2019.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. FLUX.1 Kontext: Flow matching for in-context image generation and editing in latent space, 2025. URL https://arxiv.org/abs/2506.15742.

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 *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 28306–28315, 2025.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. Gpt3. int8 (): 8-bit matrix multiplication for transformers at scale. *Advances in neural information processing systems*, 35: 30318–30332, 2022.

Zhen Dong, Zhewei Yao, Amir Gholami, Michael W Mahoney, and Kurt Keutzer. Hawq: Hessian aware quantization of neural networks with mixed-precision. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 293–302, 2019.

Zhen Dong, Zhewei Yao, Daiyaan Arfeen, Amir Gholami, Michael W Mahoney, and Kurt Keutzer. Hawq-v2: Hessian aware trace-weighted quantization of neural networks. *Advances in neural information processing systems*, 33:18518–18529, 2020.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Weilun Feng, Chuanguang Yang, Haotong Qin, Xiangqi Li, Yu Wang, Zhulin An, Libo Huang, Boyu Diao, Zixiang Zhao, Yongjun Xu, et al. 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. In *International Conference on Learning Representations (ICLR)*, 2023.

Ernst Hairer, Syvert P. Nørsett, and Gerhard Wanner. *Solving Ordinary Differential Equations I: Nonstiff Problems*. Springer, 2 edition, 1993.

Babak Hassibi and David Stork. Second order derivatives for network pruning: Optimal brain surgeon. *Advances in neural information processing systems*, 5, 1992.

Yefei He, Jing Liu, Weijia Wu, Hong Zhou, and Bohan Zhuang. Efficientdm: Efficient quantization-aware fine-tuning of low-bit diffusion models. *arXiv preprint arXiv:2310.03270*, 2023a.

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, 2023b.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

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, 2026.

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

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21807–21818, 2024b.

Yang Jin, Zhicheng Sun, Ningyuan Li, Kun Xu, Hao Jiang, Nan Zhuang, Quzhe Huang, Yang Song, Yadong Mu, and Zhouchen Lin. Pyramidal flow matching for efficient video generative modeling. In *International Conference on Learning Representations*, volume 2025, pp. 23378–23402, 2025.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5148–5157, 2021.

Raghuraman Krishnamoorthi. Quantizing deep convolutional networks for efficient inference: A whitepaper. *arXiv preprint arXiv:1806.08342*, 2018.

LAION-AI. Aesthetic-predictor. `https://github.com/LAION-AI/aesthetic-predictor`, 2022. Accessed: 2022-04-16.

Dongyeun Lee, Jiwan Hur, Hyounguk Shon, Jae Young Lee, and Junmo Kim. Dmq: Dissecting outliers of diffusion models for post-training quantization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18510–18520, 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 *Conference on Computer Vision and Pattern Recognition 2026*, 2026. URL `https://openreview.net/forum?id=vFRiUoE3UX`.

Xiuyu Li, Yijiang Liu, Long Lian, Huanrui Yang, Zhen Dong, Daniel Kang, Shanghang Zhang, and Kurt Keutzer. Q-diffusion: Quantizing diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17535–17545, 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.

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.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22139–22149, 2024.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in neural information processing systems*, 35:5775–5787, 2022.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Hyogon Ryu, NaHyeon Park, and Hyunjung Shim. Dgq: Distribution-aware group quantization for text-to-image diffusion models. *arXiv preprint arXiv:2501.04304*, 2025.

Yuzhang Shang, Zhihang Yuan, Bin Xie, Bingzhe Wu, and Yan Yan. Post-training quantization on diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1972–1981, 2023.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Zachary Teed and Jia Deng. RAFT: Recurrent all-pairs field transforms for optical flow. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 402–419. Springer, 2020.

Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. `https://github.com/huggingface/diffusers`, 2022.

Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*, 2023.

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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pp. 38–45, 2020.

Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives, 2023. URL https://arxiv.org/abs/2211.04894.

Junyi Wu, Haoxuan Wang, Yuzhang Shang, Mubarak Shah, and Yan Yan. Ptq4dit: Post-training quantization for diffusion transformers. *Advances in neural information processing systems*, 37: 62732–62755, 2024.

Junyi Wu, Zhiteng Li, Zheng Hui, Yulun Zhang, Linghe Kong, and Xiaokang Yang. Quantcache: Adaptive importance-guided quantization with hierarchical latent and layer caching for video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15035–15044, 2025.

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.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

Zhewei Yao, Zhen Dong, Zhangcheng Zheng, Amir Gholami, Jiali Yu, Eric Tan, Leyuan Wang, Qijing Huang, Yida Wang, Michael Mahoney, et al. Hawq-v3: Dyadic neural network quantization. In *International Conference on Machine Learning*, pp. 11875–11886. PMLR, 2021.

Zhewei Yao, Reza Yazdani Aminabadi, Minjia Zhang, Xiaoxia Wu, Conglong Li, and Yuxiong He. Zeroquant: Efficient and affordable post-training quantization for large-scale transformers. *Advances in neural information processing systems*, 35:27168–27183, 2022.

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

Tianchen Zhao, Xuefei Ning, Tongcheng Fang, Enshu Liu, Guyue Huang, Zinan Lin, Shengen Yan, Guohao Dai, and Yu Wang. Mixdq: Memory-efficient few-step text-to-image diffusion models with metric-decoupled mixed precision quantization. In *European Conference on Computer Vision*, pp. 285–302. Springer, 2024.

Tianchen Zhao, Tongcheng Fang, Haofeng Huang, Rui Wan, Widyadewi Soedarmadji, Enshu Liu, 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. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 65811–65841, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/a4a1ee071ce0fe63b83bce507c9dc4d7-Paper-Conference.pdf.

Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*, 2024.

A Related Work

A.1 Post-Training Quantization

Post-training quantization (PTQ) reduces the memory footprint and computational cost of pretrained networks without retraining. Early PTQ methods calibrate clipping ranges, rounding policies, and reconstruction objectives to reduce the discrepancy between full-precision and quantized outputs (Banner et al., 2019; Krishnamoorthi, 2018). Later methods improve low-bit PTQ through block-wise reconstruction, second-order sensitivity, Hessian-aware mixed precision, and activation-aware scaling (Li et al., 2021; Frantar et al., 2023; Dong et al., 2019; 2020; Yao et al., 2021; Xiao et al., 2023; Lin et al., 2024).

Our evaluation follows a calibration-only PTQ protocol. We determine quantization parameters through scale calibration, activation-statistics collection, and forward-pass profiling on a small calibration set. No gradient updates are applied to the pretrained model parameters. The main numerical comparison excludes methods that require quantization-aware training, distillation, or backpropagation during quantization. We evaluated PTQ on pretrained distilled models in Tabs. 8 and 9. These experiments applied calibration-only PTQ to already distilled models, without additional training or distillation during quantization.

Most PTQ objectives optimize scalar reconstruction or sensitivity scores, such as squared error, trace-Hessian sensitivity, or Fisher-weighted criteria. Although block-wise reconstruction (Li et al., 2021) and normalized Hessian variants (Dong et al., 2020; Yao et al., 2021) incorporate input-distribution or second-order structure, they do not explicitly distinguish trajectory-parallel from trajectory-orthogonal perturbations. This distinction is crucial for deterministic flow-matching video generation: parallel perturbations mainly alter traversal speed, whereas orthogonal perturbations rotate the velocity direction and can move the sample away from the full-precision integration path. ***TrajeQ*** therefore complements reconstruction-based PTQ with trajectory-geometric criteria for temporal bit allocation and layer preservation.

A.2 Quantization for Diffusion and Video Generation

Diffusion and video generation models are expensive to sample because they repeatedly evaluate large denoising or velocity networks. Diffusion-specific PTQ methods therefore account for timestep-dependent activation distributions and denoising-feature variation. Q-Diffusion shows that naively applying conventional PTQ can significantly degrade generation quality (Li et al., 2023). Reconstruction-focused approaches such as PTQD and PTQ4DM improve calibration and reconstruction for iterative sampling (He et al., 2023b; Shang et al., 2023), whereas EfficientDM uses quantization-aware fine-tuning (He et al., 2023a). Distribution- and outlier-aware methods such as TFMQ-DM, DGQ, and DMQ address timestep-dependent activation shifts, temporal feature variation, and channel-level outliers (Huang et al., 2024a; Ryu et al., 2025; Lee et al., 2025).

Recent diffusion-transformer quantization methods further address transformer-specific activation outliers and dynamic activation ranges. PTQ4DiT and Q-DiT consider spatial variance, temporal variance, and sample-wise dynamic activation quantization (Wu et al., 2024; Chen et al., 2025). Additionally, QuaRot reduces transformer outliers through rotation-based quantization (Ashkboos et al., 2024), while SmoothQuant balances activation and weight scales (Xiao et al., 2023). These methods are strong general-purpose or diffusion-transformer PTQ baselines, and we include representative methods from this line in our comparisons. However, their objectives mainly focus on reconstruction fidelity, outlier suppression, timestep-wise distribution shift, or scalar sensitivity, rather than the angular component of velocity-field error.

Video generation adds temporal consistency, motion coherence, and larger activation memory to the quantization problem. ViDiT-Q combines static-dynamic channel balancing with metric-decoupled mixed precision for image and video diffusion transformers (Zhao et al., 2025). While effective at suppressing activation outliers, ViDiT-Q does not separate trajectory-orthogonal from trajectory-parallel quantization error. Q-DiT similarly mitigates outlier-driven degradation through automatic quantization-granularity allocation and sample-wise dynamic activation quantization, but operates on per-step activation and weight variation rather than inter-step trajectory geometry (Chen et al., 2025). Other methods for image and video generation exploit temporal feature maintenance, mixed precision, or caching for efficiency (Huang et al., 2024a; Zhao et al., 2024; Wu et al., 2025).

Q-VDiT is closely related because it targets video-generation diffusion transformers with quantization and distillation (Feng et al., 2025). It uses a token-aware quantization estimator to compensate token- and feature-wise quantization errors, together with temporal-maintenance distillation to preserve spatiotemporal correlations. We discuss Q-VDiT as related work rather than include it as a numerical baseline because our evaluation protocol is calibration-only PTQ without gradient updates or distillation. Its distillation-based setting is complementary to our trajectory-geometric criterion and could be combined with trajectory-preserving quantization in future work.

SVDQuant (Li et al., 2024) absorbs outliers into a high-precision low-rank branch to enable 4-bit weight and activation quantization. DeltaQuant (Li et al., 2026) exploits local spatiotemporal activation similarity by quantizing delta tokens while retaining core tokens at higher precision. Our empirical comparison focuses on quantization within existing backbone layers to examine the effects of directional precision allocation. Therefore, we do not benchmark against methods employing auxiliary low-rank or core–delta branches, which fall outside the scope of this evaluation.

TrajeQ focuses on quantization for flow-matching-based video generation, with an emphasis on directional distortion during deterministic ODE sampling.

A.3 Flow Matching and Rectified Flow

Flow matching and rectified flow formulate generative modeling through continuous-time vector fields that transport a source distribution to a data distribution (Lipman et al., 2022; Liu et al., 2022). Unlike stochastic diffusion sampling, many FM and rectified-flow models use deterministic ODE integration at inference time. This enables efficient sampling but also makes the generation path sensitive to systematic perturbations in the velocity field. Recent large-scale models, including rectified-flow transformers and flow-matching video models, demonstrate the effectiveness of this paradigm for high-quality image and video generation (Esser et al., 2024; Jin et al., 2025; Black Forest Labs et al., 2025).

Our work studies PTQ specifically in this deterministic FM regime. While prior quantization methods often focus on denoising-feature reconstruction or activation outliers, FM quantization also requires preserving the geometry of the learned velocity field. In particular, perturbations parallel to the full-precision velocity mainly affect traversal speed, whereas orthogonal perturbations rotate the local update direction and can move the sample toward a different region of the state space. Despite the growing prominence of FM and rectified-flow models, PTQ for FM-based generative models remains relatively underexplored. ***TrajeQ*** addresses this gap by analyzing quantization through the geometric stability of the trajectory.

A.4 Numerical Error Propagation in Deterministic ODE Sampling

The stability of deterministic ODE sampling depends on how local perturbations propagate through the vector field. In numerical ODE analysis, local errors can be amplified by the local Lipschitz behavior of the dynamics, and Grönwall-type inequalities are a standard tool for bounding such growth (Hairer et al., 1993). Generative ODE samplers, including probability-flow ODEs, DDIM, DPM-Solver, rectified-flow samplers, and flow-matching ODE solvers, inherit this sensitivity because each perturbed state becomes the input to subsequent velocity or score evaluations (Song et al., 2020; Lu et al., 2022; Liu et al., 2022; Lipman et al., 2022). EDM further characterizes truncation and discretization error in deterministic ODE samplers, linking local solver error to accumulated sampling error (Karras et al., 2022).

This perspective motivates ***TrajeQ***. Rather than treating quantization error as an isolated layer-wise reconstruction discrepancy, we view it as a trajectory-level perturbation to deterministic ODE integration. *Curvature-Adaptive Quantization* identifies timesteps with large activation-direction changes. Our analysis relates velocity turning to the local Jacobian norm under a temporal-variation condition. We allocate additional precision at high-S_{dir} segments and evaluate this allocation through selection-agreement and generation-quality experiments. *Directional Bottleneck Preservation* complements this temporal adaptation by preserving layers vulnerable to angular distortion, using only forward-pass profiling on a small calibration set and thus preserving the calibration-only, no-retraining nature of our PTQ protocol.

B GEOMETRIC MOTIVATION

In this section, we provide a geometric analysis connecting trajectory turning to the Lipschitz stability (Hairer et al., 1993) of the probability flow ODE. We further analyze how local state changes are transferred to intermediate activations, while the correspondence between the activation-based S_{dir} and trajectory turning is validated empirically.

B.1 GEOMETRIC INTERPRETATION OF THE TURNING ANGLE VIA TAYLOR EXPANSION

Let θ_t denote the turning angle between consecutive velocity vectors $\mathbf{v}_t$ and $\mathbf{v}_{t-\Delta t}$. This angle directly characterizes the local directional change of the sampled trajectory:

$$\theta_t \triangleq \angle(\mathbf{v}_t, \mathbf{v}_{t-\Delta t}), \qquad 1-\cos\theta_t = 1 - \frac{\mathbf{v}_t \cdot \mathbf{v}_{t-\Delta t}}{\|\mathbf{v}_t\|\|\mathbf{v}_{t-\Delta t}\|}. \tag{8}$$

To rigorously justify why this cosine distance represents curvature, we analyze the local behavior of the trajectory using the Taylor series expansion. Assuming the step size Δt is sufficiently small, we expand the cosine function around $\theta_t \approx 0$:

$$1-\cos\theta_t = 1 - \left(1 - \frac{1}{2!}\theta_t^2 + \mathcal{O}(\theta_t^4)\right) = \frac{1}{2}\theta_t^2 + \mathcal{O}(\theta_t^4). \tag{9}$$

Crucially, Eq. 9 reveals that the first-order term vanishes (since $\frac{d}{d\theta}\cos(0) = -\sin(0) = 0$). Thus, the cosine-distance turning measure vanishes on locally straight segments ($\theta_t = 0$) and grows quadratically with small directional turns.

Under the assumptions of Lemma 1, define $\mathbf{u}(t) = \mathbf{v}_t/\|\mathbf{v}_t\|_2$. Since $\mathbf{u}(t)$ is a unit vector, we have:

$$1-\cos\theta_t = \frac{1}{2}\|\mathbf{u}(t) - \mathbf{u}(t-\Delta t)\|_2^2.$$

Taylor expansion gives $\mathbf{u}(t) - \mathbf{u}(t-\Delta t) = \Delta t\, \dot{\mathbf{u}}(t) + \mathcal{O}((\Delta t)^2)$. Using $\|\dot{\mathbf{u}}(t)\|_2 = \kappa(t)\|\mathbf{v}_t\|_2$, we obtain:

$$1-\cos\theta_t = \frac{1}{2}\kappa(t)^2\|\mathbf{v}_t\|_2^2(\Delta t)^2 + \mathcal{O}((\Delta t)^3). \tag{10}$$

Eq. (10) relates velocity turning to curvature, speed, and step size. The activation-based S_{dir} is used for timestep selection, with selection agreement evaluated in Appendix B.7.

B.2 RELATIONSHIP BETWEEN CURVATURE AND LIPSCHITZ STABILITY

We now analyze the conditional relationship between trajectory turning and local spatial variation of the velocity field. To this end, we seek a lower-bound certificate on the local Jacobian norm of the velocity field in terms of the trajectory curvature. Let $\mathbf{v}(t) \coloneqq v(\mathbf{x}(t), t)$, and consider non-stationary trajectory segments where $\|\mathbf{v}(t)\|_2 \geq v_{\min} > 0$, since curvature and directional stability become ill-conditioned near zero-velocity points.

In flow matching dynamics defined by $\dot{\mathbf{x}} = v(\mathbf{x}, t)$, the acceleration along the trajectory is governed by the material derivative:

$$\ddot{\mathbf{x}}(t) = \frac{\partial v}{\partial t}(\mathbf{x}(t), t) + J_v(\mathbf{x}(t), t)\,\mathbf{v}(t), \tag{11}$$

where $J_v(\mathbf{x}(t), t)$ denotes the spatial Jacobian of the velocity field with respect to $\mathbf{x}$. The geometric curvature of the trajectory is defined by the normal component of the acceleration:

$$\kappa(t) = \frac{\|P_{\mathbf{v}^\perp}\ddot{\mathbf{x}}(t)\|_2}{\|\mathbf{v}(t)\|_2^2}, \tag{12}$$

where $P_{\mathbf{v}^\perp}$ denotes the projection onto the orthogonal complement of $\mathbf{v}(t)$. Projecting Eq. 11 onto this orthogonal subspace gives:

$$P_{\mathbf{v}^\perp}\ddot{\mathbf{x}}(t) = P_{\mathbf{v}^\perp}\partial_t v(\mathbf{x}(t), t) + P_{\mathbf{v}^\perp} J_v(\mathbf{x}(t), t)\mathbf{v}(t). \tag{13}$$

This projection preserves the component that directly determines trajectory bending, rather than bounding the full acceleration norm before isolating the curvature-relevant direction.

We use the following operator-norm contraction:

$$\|P_{\mathbf{v}^\perp} J_v \mathbf{v}(t)\|_2 \le \|J_v \mathbf{v}(t)\|_2 \le \|J_v\|_2 \|\mathbf{v}(t)\|_2. \tag{14}$$

Applying the triangle inequality to Eq. 13 and using this contraction, we obtain:

$$\kappa(t)\|\mathbf{v}(t)\|_2^2 = \|P_{\mathbf{v}^\perp} \ddot{\mathbf{x}}(t)\|_2 \le \|P_{\mathbf{v}^\perp} \partial_t v(\mathbf{x}(t), t)\|_2 + \|J_v(\mathbf{x}(t), t)\|_2 \|\mathbf{v}(t)\|_2. \tag{15}$$

Rearranging Eq. 15 yields the curvature-induced lower-bound certificate on the local spatial Jacobian norm $L(t) \coloneqq \|J_v(\mathbf{x}(t), t)\|_2$:

$$L(t) \ge \kappa(t)\|\mathbf{v}(t)\|_2 - \frac{\|P_{\mathbf{v}^\perp} \partial_t v(\mathbf{x}(t), t)\|_2}{\|\mathbf{v}(t)\|_2}. \tag{16}$$

Eq. (16) establishes the direction relevant to our framework: high trajectory curvature provides a lower-bound certificate for a large local Lipschitz coefficient, up to the orthogonal temporal-variation correction $\|P_{\mathbf{v}^\perp} \partial_t v\|_2 / \|\mathbf{v}(t)\|_2$. When the orthogonal temporal variation satisfies the relative bound below, velocity turning provides a lower bound on the spatial Jacobian norm. More explicitly, suppose that the following condition holds:

$$\|P_{\mathbf{v}^\perp} \partial_t v(\mathbf{x}(t), t)\|_2 \le c\, \kappa(t)\|\mathbf{v}(t)\|_2^2 \qquad \text{for some } c < 1, \tag{17}$$

then Eq. (16) gives the following lower bound:

$$L(t) \ge (1 - c)\kappa(t)\|\mathbf{v}(t)\|_2. \tag{18}$$

Combining Lemma 1 with Eq. (18), under Eq. (17), gives:

$$\lim_{\Delta t \to 0} \frac{2(1 - \cos\theta_t)}{(\Delta t)^2} = \kappa(t)^2 \|\mathbf{v}(t)\|_2^2 \le \frac{L(t)^2}{(1 - c)^2}. \tag{19}$$

A large local Jacobian norm does not by itself establish perturbation growth. To state an accumulated-error bound, parameterize a fixed sampling stage by increasing integration time $t \in [0, T]$ and consider $\dot{\mathbf{x}}_{\mathrm{FP}}(t) = v(\mathbf{x}_{\mathrm{FP}}(t), t)$ and $\dot{\mathbf{x}}_Q(t) = v(\mathbf{x}_Q(t), t) + \boldsymbol{\epsilon}(t)$. Assume that $v(\cdot, t)$ is continuously differentiable on a neighborhood of the line segment joining the states, and that the following function is integrable:

$$\overline{L}(t) = \sup_{\alpha \in [0,1]} \|J_v(\mathbf{x}_{\mathrm{FP}}(t) + \alpha[\mathbf{x}_Q(t) - \mathbf{x}_{\mathrm{FP}}(t)], t)\|_2$$

For $e(t) = \|\mathbf{x}_Q(t) - \mathbf{x}_{\mathrm{FP}}(t)\|_2$, Grönwall's inequality (Hairer et al., 1993) gives:

$$e(T) \le e(0) \exp\left(\int_0^T \overline{L}(s)\, ds\right) + \int_0^T \|\boldsymbol{\epsilon}(t)\|_2 \exp\left(\int_t^T \overline{L}(s)\, ds\right) dt. \tag{20}$$

Here $\boldsymbol{\epsilon}(t)$ is the velocity perturbation evaluated along the quantized trajectory. The initial-error term vanishes for identical initial states. The exponential weight depends on the subsequent integration interval, so this bound does not determine an optimal allocation from the current local Jacobian norm alone.

The bound in Eq. (16) is non-vacuous whenever the following condition holds:

$$\kappa(t)\|\mathbf{v}(t)\|_2 > \frac{\|P_{\mathbf{v}^\perp} \partial_t v(\mathbf{x}(t), t)\|_2}{\|\mathbf{v}(t)\|_2}, \tag{21}$$

which makes the lower bound in Eq. (16) strictly positive. This result relates velocity turning to local spatial variation under the stated temporal-variation condition. The selection behavior of CAQ and the observed rollout deviations are reported in Appendix B.7 and Tab. 5, respectively.

B.3 From ODE State Changes to Layer-Activation Changes

While Lemma 1 characterizes the turning geometry of the ODE trajectory, CAQ measures S_{dir} on flattened intermediate activations at each monitored projection layer. For the local analysis, we consider a smooth reference mapping $\mathbf{h}_l(t) = f_l(\mathbf{x}(t), t)$ within a fixed sampling stage, with network parameters and prompt conditioning held fixed. Here, f_l denotes the composite mapping from the ODE state to the observed activation, including the preceding network operations and any explicit timestep dependence.

Intuition. The cosine score measures changes in the normalized activation direction. Therefore, its local behavior depends on the component of activation variation perpendicular to the current activation, as well as the activation norm. The following lemma makes this dependence explicit.

Lemma 3 (Local Expansion of Activation Directional Change). Suppose that $\mathbf{x}(t)$ and $f_l(\mathbf{x}, t)$ are twice continuously differentiable locally, and that $\|\mathbf{h}_l(t)\| > 0$ throughout a neighborhood of the timestep under consideration. Define the following quantities:

$$\mathbf{u}_l(t) = \frac{\mathbf{h}_l(t)}{\|\mathbf{h}_l(t)\|}, \qquad \mathbf{P}_l^{\perp}(t) = \mathbf{I} - \mathbf{u}_l(t)\mathbf{u}_l(t)^{\top}, \qquad \mathbf{J}_l(t) = \partial_{\mathbf{x}} f_l(\mathbf{x}(t), t). \tag{22}$$

The layer-wise directional sensitivity is defined as follows:

$$S_{dir}^{(l)}(t) \triangleq 1 - \langle \mathbf{u}_l(t), \mathbf{u}_l(t - \Delta t) \rangle. \tag{23}$$

Writing $\mathbf{v}(t) = \dot{\mathbf{x}}(t)$, we obtain the following leading-order expansion:

$$S_{dir}^{(l)}(t) = \frac{(\Delta t)^2}{2\|\mathbf{h}_l(t)\|^2} \left\| \mathbf{P}_l^{\perp}(t) \left[\mathbf{J}_l(t)\mathbf{v}(t) + \partial_t f_l(\mathbf{x}(t), t) \right] \right\|^2 + \mathcal{O}((\Delta t)^3). \tag{24}$$

Here, $\partial_t f_l$ denotes the partial derivative with respect to the explicit timestep input, holding the state fixed; it vanishes when the mapping has no explicit timestep dependence.

Proof. By the chain rule, we obtain:

$$\dot{\mathbf{h}}_l(t) = \mathbf{J}_l(t)\mathbf{v}(t) + \partial_t f_l(\mathbf{x}(t), t). \tag{25}$$

Differentiating the normalized activation gives:

$$\dot{\mathbf{u}}_l(t) = \frac{\mathbf{P}_l^{\perp}(t)\dot{\mathbf{h}}_l(t)}{\|\mathbf{h}_l(t)\|}. \tag{26}$$

Since both normalized activations have unit norm, we have:

$$S_{dir}^{(l)}(t) = \frac{1}{2}\|\mathbf{u}_l(t) - \mathbf{u}_l(t - \Delta t)\|^2. \tag{27}$$

Taylor expansion yields $\mathbf{u}_l(t) - \mathbf{u}_l(t - \Delta t) = \Delta t\, \dot{\mathbf{u}}_l(t) + \mathcal{O}((\Delta t)^2)$. Substitution gives $S_{dir}^{(l)}(t) = \frac{1}{2}(\Delta t)^2\|\dot{\mathbf{u}}_l(t)\|^2 + \mathcal{O}((\Delta t)^3)$, which establishes Eq. 24. □

Interpretation. Eq. 24 characterizes how local state evolution and explicit timestep conditioning contribute to changes in activation direction. The score depends on the composite mapping, the current activation direction, and its norm. This expansion does not establish a fixed multiplicative relationship with the cosine distance between state vectors, or an equivalence to velocity turning or trajectory curvature.

Implication for CAQ. CAQ (Sec. 3.1) uses S_{dir} to select timesteps for additional precision through $\mathbb{I}(S_{dir}(t) > \tau)$. Agreement between the selected timesteps and high FP velocity turning is evaluated empirically for representative layers in Appendix B.7. Complementary results in Tab. 5 show lower velocity angular error and trajectory deviation for ***TrajeQ*** than for MinMax. These measurements provide empirical evidence for the selection behavior and trajectory preservation, alongside the local characterization above.

B.4 Comparison with Hessian-based Metrics

We distinguish our proposed metric S_{dir} from Hessian-based sensitivity metrics (Hassibi & Stork, 1992) along two dimensions: trajectory-geometric alignment and computational tractability. We do not claim that all Hessian-aware PTQ methods ignore activation statistics. Block-wise reconstruction and normalized Hessian variants can incorporate input-distribution or second-order information. Our distinction is instead that such methods typically summarize quantization sensitivity as a scalar reconstruction or curvature score, without explicitly separating trajectory-parallel and trajectory-orthogonal perturbations.

Scalar Sensitivity vs. Directional Deformation. Hessian-based criteria estimate how perturbations affect a local objective, often through second-order approximations of reconstruction or task loss. Such metrics are effective for identifying parameters or layers whose perturbations cause large output changes. However, in deterministic flow-matching sampling, the effect of a perturbation depends not only on its magnitude but also on its orientation relative to the velocity field. A perturbation aligned with the velocity mainly changes the local update magnitude, whereas an orthogonal perturbation rotates the local update direction and can move the sample away from the full-precision integration path.

This distinction is closely related to geometric curvature. The curvature of the trajectory depends on the normal component of the acceleration relative to the squared velocity magnitude:

$$\kappa(t) = \frac{\|P_{\mathbf{v}^\perp}\ddot{\mathbf{x}}(t)\|_2}{\|\mathbf{v}(t)\|_2^2}. \tag{28}$$

Therefore, a large local second-order or reconstruction sensitivity does not necessarily imply a large trajectory-direction error. It may reflect high-energy variations that change output magnitude, while the orthogonal component directly changes the local update direction. Conversely, low-magnitude activation regimes can appear less important under scalar reconstruction metrics, even though small orthogonal perturbations may cause large angular deviations.

For layer preservation, DS_l uses the stabilized ratio of mean angular error to mean squared error (Eq. 7). Figs. 3d and 3c compare the resulting layer ranking with that of Hessian trace.

For timestep selection, S_{dir} directly measures normalized directional change between consecutive activations:

$$S_{dir}(t) = 1 - \frac{\mathbf{h}_l(t) \cdot \mathbf{h}_l(t-\Delta t)}{\|\mathbf{h}_l(t)\|_2 \|\mathbf{h}_l(t-\Delta t)\|_2}. \tag{29}$$

By using cosine distance, S_{dir} suppresses pure scale effects and emphasizes changes in the activation direction. Thus, it is better aligned with our objective of identifying timesteps where angular distortion can accumulate under deterministic ODE integration.

Computational Tractability and Complexity. Computing Hessian-based sensitivity at every sampling step is also expensive for high-dimensional video generation models. Exact Hessian computation is quadratic in the number of variables, while trace or Hessian-vector approximations require additional iterative evaluations or backward passes. Such overhead is undesirable for online timestep-wise precision adaptation, where the quantization policy must be updated during inference.

By contrast, S_{dir} requires only vector normalization and dot products between consecutive activations, giving linear complexity in the activation dimension. This makes it practical to integrate into the inference loop while preserving the calibration-only, no-retraining nature of our PTQ protocol.

B.5 Justification for Precision Boosting via Variance Reduction

We formally analyze the impact of quantization precision on noise variance to substantiate the proposed bit-boosting strategy.

Lemma 4 (Variance Scaling Law). For a uniform quantizer with fixed range R, step size $\Delta_b = R2^{-b}$, and granular error modeled as uniform on $[-\Delta_b/2, \Delta_b/2]$, an increment of 1 bit reduces the modeled error variance by a factor of 4.

Proof. Consider a uniform scalar quantizer with a dynamic range R and bit-width b. The quantization step size (or bin width) is defined as $\Delta_b = R \cdot 2^{-b}$.

For a quantizer employing a nearest-neighbor rounding scheme, the quantization error ϵ is bounded within the interval $[-\frac{\Delta_b}{2}, \frac{\Delta_b}{2}]$. Adopting the standard high-resolution hypothesis, we model ϵ as a random variable following a uniform distribution:

$$\epsilon \sim \mathcal{U}\left(-\frac{\Delta_b}{2}, \frac{\Delta_b}{2}\right). \tag{30}$$

The variance of this error is derived by integrating the probability density function over the bin width:

$$\sigma_b^2 = \mathrm{Var}(\epsilon) = \int_{-\frac{\Delta_b}{2}}^{\frac{\Delta_b}{2}} \frac{1}{\Delta_b}\epsilon^2 \, d\epsilon = \frac{\Delta_b^2}{12}. \tag{31}$$

When the precision is increased to $b+1$ bits, the step size changes according to the definition. By explicitly factoring out the exponent, we observe:

$$\Delta_{b+1} = \frac{R}{2^{b+1}} = \frac{R}{2 \cdot 2^b} = \frac{1}{2}\left(\frac{R}{2^b}\right) = \frac{1}{2}\Delta_b. \tag{32}$$

Substituting this halved step size into Eq. 31, the variance scales quadratically:

$$\sigma_{b+1}^2 = \frac{(\Delta_{b+1})^2}{12} = \frac{(\frac{1}{2}\Delta_b)^2}{12} = \frac{1}{4}\left(\frac{\Delta_b^2}{12}\right) = \frac{1}{4}\sigma_b^2. \tag{33}$$

Consequently, this establishes that $\sigma_{b+1}^2 = 0.25\sigma_b^2$, confirming a fourfold reduction in noise variance. □

For the signed symmetric quantizer in Alg. 1, the step size at fixed clipping bound c is $s_b = c/(2^{b-1}-1)$, ignoring stabilization. Under the same uniform-error model, the variance ratio is approximately:

$$\frac{\sigma_{b+1}^2}{\sigma_b^2} \approx \left(\frac{2^{b-1}-1}{2^b-1}\right)^2.$$

Thus, the quarter-variance rule is a high-resolution approximation for this quantizer, rather than an exact statement about total error on the observed activations.

B.6 Derivation and Geometric Interpretation of Directional Sensitivity

In this section, we provide a rigorous step-by-step derivation of the *Directional Sensitivity* (DS_l) metric proposed in Sec. 3.2. We explicitly derive the first-order approximation used to model the geometric instability of normalization operations.

We begin by analyzing the local sensitivity of the normalization function $f(\mathbf{x}) = \mathbf{x}/\|\mathbf{x}\|$. To quantify how input perturbations translate to output changes, we compute the Jacobian matrix $\mathbf{J}_f(\mathbf{x})$. Applying vector calculus, the derivative yields a scaled projection matrix:

$$\mathbf{J}_f(\mathbf{x}) = \frac{1}{\|\mathbf{x}\|}\left(\mathbf{I} - \frac{\mathbf{x}\mathbf{x}^\top}{\|\mathbf{x}\|^2}\right). \tag{34}$$

Here, the matrix term acts as an **orthogonal projection operator**. Intuitively, this operator removes any change in the radial direction (magnitude) and preserves only the changes tangential to the hypersphere surface.

Next, we examine the effect of a perturbation δ on this structure. By decomposing the noise into parallel and orthogonal components ($\delta = \delta^{\parallel} + \delta^{\perp}$) and applying the first-order Taylor expansion ($f(\mathbf{x}+\delta) - f(\mathbf{x}) \approx \mathbf{J}_f(\mathbf{x})\delta$), we derive the magnitude of the displacement as follows:

$$\begin{aligned} \|f(\mathbf{x}+\delta) - f(\mathbf{x})\| &\approx \|\mathbf{J}_f(\mathbf{x})\delta\| \\ &= \left\| \frac{1}{\|\mathbf{x}\|}\left(\mathbf{I} - \frac{\mathbf{x}\mathbf{x}^\top}{\|\mathbf{x}\|^2}\right)(\delta^{\parallel} + \delta^{\perp}) \right\| = \frac{\|\delta^{\perp}\|}{\|\mathbf{x}\|}. \end{aligned} \tag{35}$$

Since the parallel component ($\delta^{\parallel}$) vanishes, the magnitude of the displacement is determined solely by the orthogonal noise scaled by the inverse activation magnitude.

To bridge this linear displacement to the angular error, we utilize the geometric relationship for small angles. The displacement in Eq. 35 represents the chord length on the unit hypersphere. For small perturbations, this chord length is approximately equal to the arc length (angle θ in radians), implying $\theta \approx \|\delta^{\perp}\|/\|\mathbf{x}\|$. Combining this with the second-order Taylor expansion of the cosine distance ($\mathcal{E}_{ang} \approx \frac{1}{2}\theta^2$), we arrive at:

$$\mathcal{E}_{ang} \approx \frac{1}{2}\left(\frac{\|\delta^{\perp}\|}{\|\mathbf{x}\|}\right)^2 = \frac{1}{2}\frac{\|\delta^{\perp}\|^2}{\|\mathbf{x}\|^2}. \tag{36}$$

For fixed nonzero $\mathbf{x}_l$ and a nonzero perturbation with $\|\delta_l\|_2/\|\mathbf{x}_l\|_2 \to 0$, division by the squared perturbation norm gives:

$$\frac{\mathcal{E}_{ang}}{\|\delta_l\|_2^2} = \frac{\|\delta_l^{\perp}\|_2^2}{2\|\mathbf{x}_l\|_2^2\|\delta_l\|_2^2} + \mathcal{O}\left(\frac{\|\delta_l\|_2}{\|\mathbf{x}_l\|_2^3}\right). \tag{37}$$

The leading term combines inverse squared activation magnitude with the orthogonal fraction of perturbation energy. Let $\delta_l^{t,c} = \hat{\mathbf{x}}_l^{t,c} - \mathbf{x}_l^{t,c}$. For a fixed finite profiling set with nonzero reference activations and sufficiently small relative perturbations, averaging the angular-error expansion gives:

$$DS_l \approx \frac{1}{2} \frac{\mathbb{E}_{c,t}\left[\|(\delta_l^{t,c})^{\perp}\|_2^2 / \|\mathbf{x}_l^{t,c}\|_2^2\right]}{\mathbb{E}_{c,t}\left[\|\delta_l^{t,c}\|_2^2\right] + \epsilon}.$$

The activation norm remains inside the expectation because it can vary across prompts and timesteps. Alg. 2 computes Eq. (7) directly from the observed angular and squared errors.

B.7 Empirical Selection Agreement with FP Velocity Turning

CAQ uses S_{dir} only through the thresholding decision $\mathbb{I}(S_{dir} > \tau)$. Therefore, the relevant requirement for bit allocation is whether activation-based S_{dir} identifies high-turning timesteps, rather than whether it numerically matches the FP turning angle θ_t. We evaluate this selection agreement directly against FP velocity turning.

Tab. 6 illustrates that S_{dir} achieves substantially higher precision and IoU than random selection at the same budget, selecting only 3.5% of steps. Across all prompts, every selected step lies within the top-5% FP turning set, while the selected steps exhibit substantially larger median turning angles than the remaining steps. These results provide empirical evidence that activation-based S_{dir} effectively identifies high-turning trajectory intervals for CAQ.

Table 6: Agreement between steps selected by $S_{dir} > 0.25$ and an equal number of highest-turning FP steps (FLUX/Pyramid, OpenSora prompts, 462 turning comparisons $\times$ 10 prompts, per-prompt mean $\pm$ std, 1,000 random permutations; $\mathrm{IoU} = p/(2-p)$).

Step Selection	Selected Steps	Precision at equal budget	IoU	Top-5% inclusion rate	Median θ_t (selected / rest)
Random at equal budget	164 / 4,620 (3.5%)	$.035 \pm .003$	$.018 \pm .001$	$.049 \pm .002$	$3.5° / 3.4°$
S_{dir} at `proj_out`	164 / 4,620 (3.5%)	$.839 \pm .104$	$.735 \pm .162$	$1.000 \pm .000$	$33.5° / 3.3°$
S_{dir} at `attn.to_q`	161 / 4,620 (3.5%)	$.832 \pm .089$	$.721 \pm .132$	$1.000 \pm .000$	$33.5° / 3.3°$

C ALGORITHMS AND IMPLEMENTATION

We provide the pseudo-code for our proposed methods: Curvature-Adaptive Quantization (CAQ), Directional Bottleneck Preservation (DBP), and On-the-Fly Calibration.

C.1 ALGORITHM: CURVATURE-ADAPTIVE QUANTIZATION (CAQ)

Algorithm 1 Curvature-Adaptive Quantization (CAQ)

Input: Current layer activation $\mathbf{x}_t$, cached previous activation $\mathbf{x}_{t-\Delta t}$, base bits b_{base}, threshold τ
Output: Quantized activation $\hat{\mathbf{x}}_t$
Note: Run at each timestep t; the cached activation is updated after each forward step.
$B_t \leftarrow b_{\text{base}}$
if $\mathbf{x}_{t-\Delta t} \neq \texttt{null}$ **then**
  $S_{\text{dir}} \leftarrow 1 - \frac{\langle \text{vec}(\mathbf{x}_t),\, \text{vec}(\mathbf{x}_{t-\Delta t}) \rangle}{\|\text{vec}(\mathbf{x}_t)\|_2 \cdot \|\text{vec}(\mathbf{x}_{t-\Delta t})\|_2 + \epsilon}$
  if $S_{\text{dir}} > \tau$ **then**
    $B_t \leftarrow \min(B_t + 1,\ 16)$
  end if
end if
if $B_t \geq 16$ **then**
  $\hat{\mathbf{x}}_t \leftarrow \mathbf{x}_t$
else
  $q_{\max} \leftarrow 2^{B_t - 1} - 1$
  $\mathbf{s} \leftarrow \text{ComputeScale}(\mathbf{x}_t, q_{\max})$ ▷ Alg. 3
  $\hat{\mathbf{x}}_t \leftarrow \mathbf{s} \cdot \text{clip}\left(\lfloor \mathbf{x}_t / \mathbf{s} \rceil,\ -q_{\max},\ q_{\max}\right)$
end if
Cache $\mathbf{x}_{t-\Delta t} \leftarrow \mathbf{x}_t$
return $\hat{\mathbf{x}}_t$

C.2 ALGORITHM: DIRECTIONAL BOTTLENECK PRESERVATION (DBP)

Algorithm 2 Directional Bottleneck Preservation (DBP)

Input: Model $\mathcal{M}$, profiling prompts $\mathcal{C}$, timesteps $\mathcal{T}$, base bits (b_w, b_a), FP16 budget ratio ρ, stabilization constant $\epsilon > 0$
Output: Bit-width map $\mathcal{B}$
Note: Run once at initialization; profiling is used only for layer ranking, while quantization parameters are computed on-the-fly.
for each quantizable layer $l_i \in \mathcal{M}$ **do**
  $A_i \leftarrow 0,\ E_i \leftarrow 0$
  for each $(c, t) \in \mathcal{C} \times \mathcal{T}$ **do**
    Obtain FP16 $\mathbf{x}_i^{t,c}$ and layer-wise quantized $\hat{\mathbf{x}}_i^{t,c}$
    $A_i \leftarrow A_i + 1 - \cos(\mathbf{x}_i^{t,c}, \hat{\mathbf{x}}_i^{t,c})$
    $E_i \leftarrow E_i + \|\hat{\mathbf{x}}_i^{t,c} - \mathbf{x}_i^{t,c}\|_2^2$
  end for
  $DS_i \leftarrow A_i / (E_i + |\mathcal{C}|\,|\mathcal{T}|\epsilon)$
end for
Sort layers by DS_i in descending order
$\mathcal{S} \leftarrow \emptyset,\quad m_{\text{fp}} \leftarrow 0,\quad m_{\text{tot}} \leftarrow \sum_{l_j \in \mathcal{M}} |\Theta_j|$
for each sorted layer l_i **do** ▷ Budgeted top-DS_i selection
  if $m_{\text{fp}} + |\Theta_i| \leq \rho \cdot m_{\text{tot}}$ **then**
    $\mathcal{S} \leftarrow \mathcal{S} \cup \{l_i\};\quad m_{\text{fp}} \leftarrow m_{\text{fp}} + |\Theta_i|$
  end if
end for
$\mathcal{B}(l_i) \leftarrow \begin{cases} (16, 16), & l_i \in \mathcal{S}, \\ (b_w, b_a), & \text{otherwise}, \end{cases} \quad \forall l_i \in \mathcal{M}$ ▷ Construct final mixed-precision map
return $\mathcal{B}$

C.3 Algorithm: On-the-Fly Calibration

Algorithm 3 Static-Dynamic Calibration for Quantization

Input: Tensor $\mathbf{x}$ or weight $\mathbf{W}$, quantization range $q_{\max}$, $\epsilon > 0$
Output: Quantization scale $\mathbf{s}$
Note: Weight scales are computed once at initialization, while activation scales are updated on-the-fly at each forward step.
function CalibrateWeight($\mathbf{W}, q_{\max}$)
    $\mathbf{s}_w \leftarrow \mathrm{amax}(|\mathbf{W}|, \mathrm{dim} = 1)/(q_{\max} + \epsilon)$ ▷ Per-channel weight scale
    return $\max(\mathbf{s}_w, \epsilon)$
end function
function ComputeScale($\mathbf{x}, q_{\max}$)
    $c \leftarrow \mathrm{clip}(\|\mathbf{x}\|_\infty,\ 10^{-5},\ 10^{4})$ ▷ Dynamic activation range
    $\tilde{\mathbf{x}} \leftarrow \mathrm{clip}(\mathbf{x}, -c, c)$
    if $\dim(\mathbf{x}) \geq 3$ **then**
        $\mathbf{s} \leftarrow \mathrm{amax}(|\tilde{\mathbf{x}}|, \mathrm{dim} = -1)/(q_{\max} + \epsilon)$ ▷ Per-token activation scale
    else
        $\mathbf{s} \leftarrow \|\tilde{\mathbf{x}}\|_\infty/(q_{\max} + \epsilon)$ ▷ Per-tensor activation scale
    end if
    return $\max(\mathbf{s}, \epsilon)$
end function

D Experimental Setup

D.1 Hardware and Software Environment

All experiments are conducted on a high-performance computing node equipped with an NVIDIA A100 (80GB) GPU. The software environment uses Python 3.8.10, PyTorch 2.1.2 (Paszke et al., 2019), CUDA 12.1, Transformers v4.39.3 (Wolf et al., 2020), and Diffusers v0.30.1 (von Platen et al., 2022). Transformers is used for transformer-based backbone and text-encoding modules, while Diffusers is used for scheduler and auxiliary diffusion-pipeline utilities. Due to the high cost of video generation, each configuration is evaluated once, except for Fig. 7, which averages three runs with different random seeds. Methods within the same table share identical prompt–seed pairs for controlled comparison, whereas the same configuration may differ across tables because tables were produced in separate evaluation batches with their own seeds and run-to-run variation.

D.2 Generation and Quantization Hyperparameters

Videos are generated at a resolution of 1280×768 with temporal length parameter $\texttt{temp} = 16$ and saved at 24 FPS. The Pyramid Flow inference pipeline uses three stages. We set the image-stage inference steps to $[20, 20, 20]$ and the video-stage inference steps to $[10, 10, 10]$ across the three stages, using the FlowMatchEuler scheduler (Liu et al., 2022). Guidance scales are set to 7.0 for the first frame and 5.0 for subsequent video frames, using classifier-free guidance (Ho & Salimans, 2022). Due to the large number of videos required by the VBench evaluation, each prompt is generated once with a seed fixed within each evaluation batch, so that all methods compared in the same table receive identical prompt–seed pairs.

For ***TrajeQ***, we set the *Curvature-Adaptive Quantization (CAQ)* threshold to $\tau = 0.25$. For Figs. 4 and 7, we measure the unnormalized L_2 distance between consecutive decoded RGB frames. For each threshold, a 121-frame clip generated from the same prompt at 1280×768 yields 120 frame-pair distances:

$$D_k = \|\mathrm{vec}(F_{k+1}) - \mathrm{vec}(F_k)\|_2, \qquad \bar{D} = \frac{1}{120}\sum_{k=1}^{120} D_k.$$

Here, F_k denotes the k-th decoded RGB frame. All quantization methods use the same on-the-fly calibration protocol unless otherwise specified.

D.3 Model Backbones and Evaluation Protocol

We evaluate two flow-based backbone settings. Our primary backbone is Pyramid Flow mini-FLUX (Jin et al., 2025), a flow-matching video model with 8 joint-attention and 16 single-stream layers, with a hidden dimension of 1920 and a text-conditioning input dimension of 4096. To assess generality across rectified-flow transformer designs, we also evaluate an MMDiT-based architecture (Esser et al., 2024). Beyond FM, we evaluate CogVideoX (Yang et al., 2024) under deterministic sampling to test whether the proposed trajectory-aware quantization extends to diffusion-based video generation.

For quantitative evaluation, we primarily use VBench (Huang et al., 2024b). We additionally evaluate on the standard set of 10 prompts from the OpenSora benchmark (Zheng et al., 2024), enabling comparison across text-video alignment, perceptual quality, and temporal stability metrics. For fairness, all methods compared within a table are evaluated with the same prompt–seed pairings, following the standard VBench protocol and the settings used in ViDiT-Q (Zhao et al., 2025) and Q-DiT (Chen et al., 2025).

D.3.1 Evaluation Metrics: VBench Protocol

To conduct a multifaceted performance validation of the quantized models, we adopt VBench, which rigorously analyzes video quality across three core dimensions: *Visual Fidelity*, *Temporal Coherence*, and *Contextual Alignment*. Each dimension is quantified through the following metrics:

- **Visual Fidelity (Intra-frame Quality)**: Assesses the static visual quality of individual frames, incorporating MUSIQ-based **Imaging Quality** (noise and distortion levels) (Ke et al., 2021) and LAION-based **Aesthetic Quality** (artistic and human visual preference) (LAION-AI, 2022).
- **Temporal Coherence (Dynamic Stability)**: Evaluates the continuity between frames and the naturalness of motion. This includes **Dynamic Degree** (magnitude of object motion), **Motion Smoothness** (physical plausibility), **Subject Consistency** (preservation of object identity), and **Background Consistency** (temporal stability of the environment via CLIP (Radford et al., 2021)).
- **Contextual Alignment (Semantic Accuracy)**: Verifies the visual fidelity relative to text prompts. It consists of **Scene Consistency** (alignment with the described environment) and **Overall Consistency** (comprehensive style and semantic coherence measured via viCLIP (Wang et al., 2023)).

Prompt-to-Metric Mapping Strategy: Following the official VBench protocol and established literature such as ViDiT-Q, we conducted experiments using a total of 251 prompts categorized into three independent evaluation sets. Each set is engineered to maximize the discriminative power of specific targeted metrics:

- **overall_consistency** (93 prompts): Concentrates on the model's general generative capability and aesthetic excellence by measuring **Overall Consistency**, **Aesthetic Quality**, and **Imaging Quality**.
- **subject_consistency** (72 prompts): Assesses dynamic subjects to derive **Subject Consistency**, **Dynamic Degree**, and **Motion Smoothness**.
- **scene** (86 prompts): Used to assess environmental stability and textual adherence by deriving **Scene Consistency** and **Background Consistency**.

D.3.2 Evaluation Metrics: OpenSora Protocol

To assess the generation quality from multiple perspectives, we generate 10 video samples corresponding to the standard OpenSora prompt set (Zheng et al., 2024). We employ six quantitative metrics covering text-video alignment, temporal consistency, and visual fidelity.

- **CLIPSIM (Text-Video Alignment):** This metric quantifies the semantic alignment between the generated video frames and the input text prompt. We compute the CLIP score for each

frame using a pre-trained CLIP model (Radford et al., 2021) and report the average score, measuring how accurately the visual content reflects the textual description.

- **CLIP-Temp (Semantic Stability):** To evaluate the stability of semantic content over time, CLIP-Temp calculates the average cosine similarity of CLIP embeddings between consecutive frames (Radford et al., 2021). A higher score indicates that the semantic context remains consistent throughout the video duration.
- **VQA-Aesthetic:** Evaluates high-level perceptual and artistic appeal (e.g., lighting, composition, color). We employ the aesthetic branch of DOVER (Wu et al., 2023) to capture these aesthetic sensibilities independently of technical distortions.
- **VQA-Technical:** Measures low-level signal integrity, isolating technical fidelity from aesthetic appeal. We employ DOVER (Wu et al., 2023) to detect specific degradations (e.g., noise, blur, artifacts), ensuring a robust evaluation of generation quality.
- **Flow Score (Motion Dynamics):** We employ RAFT (Teed & Deng, 2020) to extract dense optical flow between adjacent frames. The Flow Score is the average optical-flow magnitude. Δ Flow Score measures the deviation of this score from the FP16 reference ($\downarrow$).
- **Warping Error (Temporal Consistency):** Evaluates temporal coherence by computing the photometric error between frame $t+1$ and the optical flow-warped frame t (Teed & Deng, 2020). Lower values ($\downarrow$) indicate smoother motion.

D.3.3 Evaluation Metrics: Computational Cost

We evaluate the quality-efficiency trade-off using resource metrics that capture model storage, inference-time memory usage, and end-to-end generation throughput.

- **Performance Metrics (Quality):** We use **LPIPS** (Zhang et al., 2018) and **SSIM** (Wang et al., 2004) to assess video fidelity. LPIPS measures perceptual distance in deep feature space ($\downarrow$), while SSIM measures structural and photometric similarity ($\uparrow$).
- **Efficiency Metrics (Resources):** We report **Model Size**(GB), **Peak Memory** (MB), and **Total Generation Time**(h) to measure storage cost, inference-time VRAM usage, and end-to-end evaluation throughput, respectively.

Table 7: Effect of τ on generation quality at W3A6 on OpenSora prompts. $\tau = 0.25$ gives the best overall quality trade-off, attaining the highest CLIPSIM, VQA-Aes, and VQA-Tech scores while maintaining competitive CLIP-Temp.

τ	CLIPSIM	CLIP-Temp	VQA-Aes	VQA-Tech
0.15	0.3135	**0.9966**	57.94	53.54
0.20	0.3154	0.9959	58.38	53.80
0.25	**0.3228**	0.9961	**60.60**	**54.94**
0.30	0.3199	0.9960	58.62	53.94
0.35	0.3226	**0.9966**	58.63	53.38
0.50	0.3202	0.9957	58.69	53.46

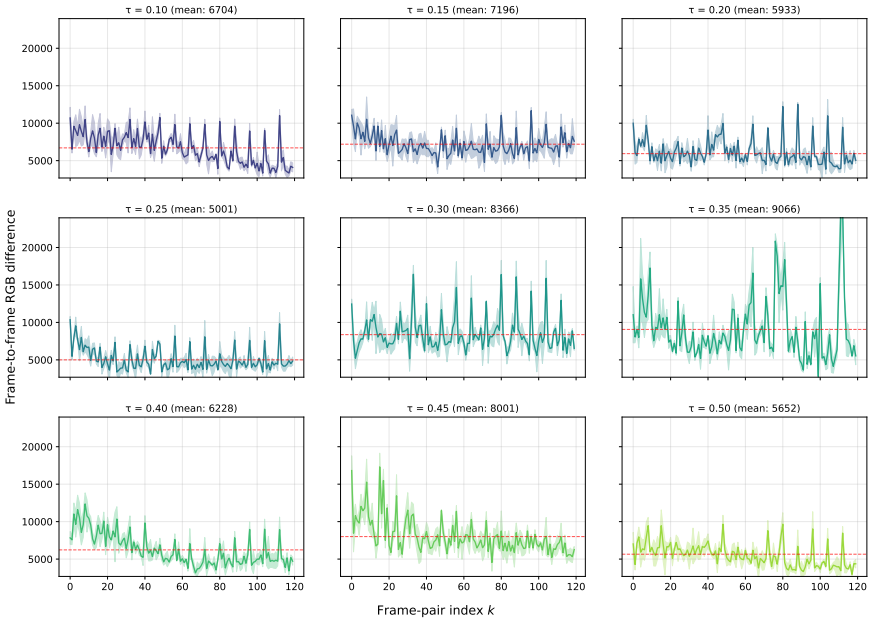


Figure 7: **Frame-to-frame RGB differences across CAQ thresholds.** Each panel illustrates the temporal differences between consecutive decoded frames, where the red dashed line denotes the mean value. Among all evaluated thresholds, $\tau = 0.25$ achieves the lowest mean difference (5001). Results averaged over three runs with different random seeds.

E Component Analysis

E.1 Hyperparameter Analysis: Selection of Threshold τ for CAQ

We examine the CAQ threshold τ using frame-to-frame RGB differences and generation-quality metrics. Fig. 7 shows differences between consecutive decoded frames for $\tau \in [0.10, 0.50]$. Among the tested values, $\tau = 0.25$ yields the lowest mean difference (5001), although this quantity alone does not establish generation quality or ODE trajectory stability. We use $\tau = 0.25$ as a robust operating point across the evaluated models. Consistent gains on both the MMDiT architecture (Tab. 13) and CogVideoX (Tab. 14) indicate that the same threshold remains effective across different deterministic generation backbones.

To further validate its robustness, we subsequently conduct a quality sweep over τ under the W3A6 setting. Tab. 7 shows that $\tau = 0.25$ provides the strongest overall trade-off, achieving the best CLIPSIM, VQA-Aes, and VQA-Tech scores while maintaining competitive temporal consistency. These generation-quality results support $\tau = 0.25$ as a robust operating point in the evaluated settings.

We additionally evaluate CAQ threshold sensitivity on the 4-step distilled Self-Forcing/Wan2.1 model (Huang et al., 2026) (Tab. 8). We also examine this sensitivity at two additional output resolutions while keeping all other settings fixed (Tab. 9).

Table 8: Sensitivity to τ with Self-Forcing/Wan2.1 using 4-step sampling at 480×832 on OpenSora prompts. Shaded rows denote the default threshold $\tau = 0.25$. **Bold** indicates the best score within each quantized bit-width setting. Spread denotes the maximum minus minimum score across the three thresholds. VQA scores are reported on a 0–1 scale.

Bit-width	τ	CLIPSIM	CLIP-Temp	VQA-Aes	VQA-Tech
FP16	–	0.3479	0.9973	0.6507	0.6957
W4A4	0.15	0.3402	**0.9960**	0.6321	0.6456
W4A4	**0.25**	**0.3420**	0.9959	**0.6372**	**0.6540**
W4A4	0.35	0.3390	0.9956	0.6327	0.6424
Spread		0.0030	0.0004	0.0051	0.0116
W3A6	0.15	0.3299	0.9952	**0.5012**	0.6289
W3A6	**0.25**	**0.3307**	**0.9961**	0.5009	**0.6294**
W3A6	0.35	0.3302	0.9960	0.4997	0.6286
Spread		0.0008	0.0009	0.0015	0.0008

Table 9: Sensitivity to τ across output resolutions with Self-Forcing/Wan2.1 using 4-step sampling on OpenSora prompts. All settings other than output resolution are unchanged from Tab. 8. Shaded rows denote the default threshold $\tau = 0.25$. **Bold** indicates the best score within each resolution and quantized bit-width setting. Spread denotes the maximum minus minimum score across the three thresholds. VQA scores are reported on a 0–1 scale.

Bit-width	τ	CLIPSIM	CLIP-Temp	VQA-Aes	VQA-Tech
Resolution: 480×480					
FP16	–	0.3505	0.9964	0.6353	0.7021
W4A4	0.15	0.3398	0.9958	0.5948	0.6400
W4A4	**0.25**	**0.3443**	**0.9962**	**0.5977**	0.6427
W4A4	0.35	0.3418	0.9959	0.5934	**0.6432**
Spread		0.0045	0.0004	0.0043	0.0032
W3A6	0.15	**0.3198**	**0.9976**	0.4857	0.6209
W3A6	**0.25**	0.3192	**0.9976**	**0.4861**	**0.6215**
W3A6	0.35	0.3186	0.9974	0.4858	0.6207
Spread		0.0012	0.0002	0.0004	0.0008
Resolution: 624×624					
FP16	–	0.3515	0.9949	0.6561	0.7027
W4A4	0.15	0.3446	**0.9955**	0.6477	0.6534
W4A4	**0.25**	**0.3459**	0.9953	**0.6519**	**0.6560**
W4A4	0.35	0.3441	0.9953	0.6359	0.6536
Spread		0.0018	0.0002	0.0160	0.0026
W3A6	0.15	0.3392	0.9970	0.5363	0.6247
W3A6	**0.25**	0.3398	**0.9971**	**0.5367**	0.6268
W3A6	0.35	**0.3412**	0.9970	0.5341	**0.6279**
Spread		0.0020	0.0001	0.0026	0.0032

Tab. 8 demonstrates that generation quality remains stable across the tested τ values under 4-step sampling. Under W3A6, the score spread across the tested thresholds is at most 0.0015 across all metrics. Under W4A4, the largest spread is 0.0116 in VQA-Tech. The default $\tau = 0.25$ achieves the highest CLIPSIM, VQA-Aes, and VQA-Tech scores under W4A4, retaining 97.9% of the FP16 VQA-Aes score (0.6372 vs. 0.6507). These results support retaining the default threshold in this few-step setting without additional tuning.

Furthermore, Tab. 9 illustrates how threshold sensitivity changes with output resolution. Across the two additional resolutions, all score spreads are at most 0.0045, except for W4A4 VQA-Aes at 624×624, where the spread reaches 0.0160. In this configuration, $\tau = 0.25$ yields a VQA-Aes score of 0.6519, close to the FP16 score of 0.6561. Although sensitivity varies across metrics and configurations, these results support using the same default threshold across the tested resolutions without resolution-specific tuning.

E.2 DBP Analysis: Geometric Bottleneck Layers

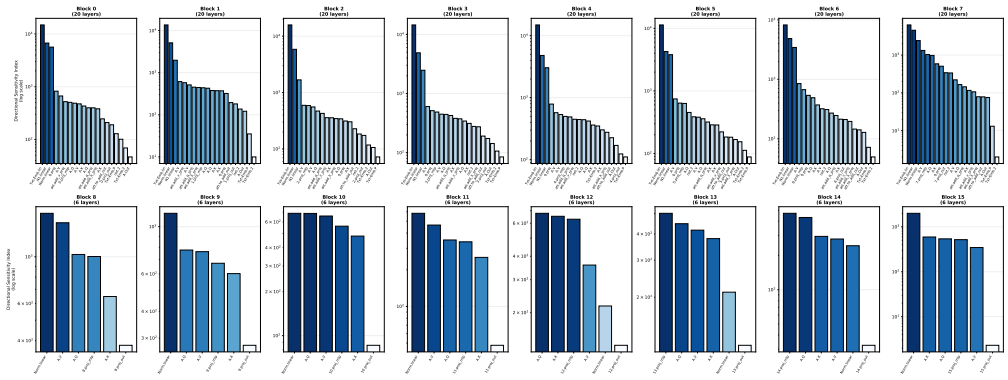


Figure 8: Layer-wise Directional Sensitivity (DS_l) in all eight joint-attention blocks (0–7) and the first eight single-stream blocks (8–15). Each panel shows one block on a logarithmic scale with its own y-axis range. Layers are sorted by DS_l in descending order.

Fig. 8 illustrates the distribution of the layer-wise directional sensitivity scores (DS_l) defined in Eq. 7. In the dual-stream Blocks 0–7, these scores exhibit substantial within-block variance, with peak values heavily concentrated in a small subset of layers. Across all eight blocks, `Txt-Emb.linear` consistently emerges as the most sensitive, while `N1.linear` and `Norm.linear` systematically occupy the remaining top-three positions. This recurring pattern reflects consistently high stabilized ratios of mean angular error to mean squared error for these layers under the evaluated quantization setting.

Guided by these observations, DBP performs a global ranking of all quantizable layers rather than relying on block level statistics (Alg. 2). Although Fig. 8 individually sorts layers within each block for visualization, DBP directly allocates its FP16 budget to the layers with the highest DS_l globally. In our setting, this selectively preserves less than 1% of the total parameters, consisting primarily of embedding and normalization modules. Tab. 11 provides an ablation study isolating the contribution of preserving each component.

E.3 DBP Analysis: Layer-Selection Ablation

To isolate the effect of the layer-selection rule, we compare DS_l with MSE and Angular-only under the same FP16 parameter budget. MSE ranks layers by mean elementwise squared error, whereas Angular-only ranks them by mean cosine error. Tab. 10 shows that DS_l achieves the highest scores across all three metrics. Compared with Angular-only, DS_l improves CLIPSIM by 0.0758, VQA-Aes by 6.99, and VQA-Tech by 5.32. These results indicate that evaluating angular distortion relative to squared reconstruction error improves layer selection for FP16 preservation in this setting.

Table 10: Comparison of layer-selection rules at W3A6 on OpenSora prompts with CAQ disabled. All rules use the same FP16 parameter budget of 17,337,600 parameters (0.88% of the model) and identical prompt–seed pairs.

Selection Rule	CLIPSIM	VQA-Aes	VQA-Tech
MSE	0.2425	52.86	46.78
Angular-only	0.2331	50.24	48.19
DS_l **(Ours)**	**0.3089**	**57.23**	**53.51**

E.4 DBP Analysis: Why Bottleneck Layers Should Skip Quantization

To justify the necessity of our *Directional Bottleneck Preservation (DBP)* strategy under aggressive quantization regimes, we compare it against two alternative outlier mitigation techniques: *Mean Imputation* and *Outlier Clamping*. As shown in Fig. 9, the handling mechanism for sensitive layers determines the fidelity of the generated video.

First, *Mean Imputation* replaces outlier values with channel-wise means and leads to severe structural degradation in Fig. 9b. This operation changes both activation magnitudes and directions. The resulting example shows lower visual fidelity than FP16 preservation.

Second, *Outlier Clamping* clips activation values to a maximum threshold in Fig. 9c. While this preserves feature signs and offers improved coherence over mean imputation, it inherently distorts the magnitude of activation vectors. The resulting example differs from the FP16 reference in pose and viewing angle. Coordinate-wise clipping changes both vector magnitude and direction.

In contrast, *DBP* avoids these distortions by skipping quantization for the selected bottleneck layers and retaining them in FP16. With these layers retained in FP16, DBP closely matches the FP16 baseline in visual fidelity, viewpoint, and structural layout in the example shown in Fig. 9d.



(a) FP16 (Original)



(b) Mean Imputation



(c) Clamping



(d) Quant Skip (*DBP*, Ours)

Figure 9: Visual comparison of handling strategies for bottleneck layers under W4A6 quantization. **(a) FP16 (Original)** serves as the full-precision reference. **(b) Mean Imputation** causes severe semantic drift and structural degradation. **(c) Clamping** partially preserves coarse structure but alters the subject's pose and viewing angle. **(d) Quantization Skipping (*DBP*, Ours)** skips quantization only for DBP-selected bottleneck layers, retaining them in FP16 and closely matching the original visual quality.

Table 11: DBP component ablation at W3A6 on OpenSora prompts with CAQ disabled. We vary FP16 preservation of the DBP-selected embedding and normalization layers. All other quantization settings and prompt–seed pairs are fixed across rows. "✓" denotes FP16 preservation, and "✗" denotes quantization.

Strategy	Emb (FP16)	Norm (FP16)	CLIPSIM	VQA-Aes	VQA-Tech
No_DBP	✗	✗	0.2933	49.01	51.21
Norm_Only	✗	✓	0.2839	50.17	50.15
Emb_Only	✓	✗	0.3070	53.55	52.82
Full_DBP (Ours)	✓	✓	**0.3086**	**57.26**	**53.48**

Furthermore, Tab. 11 evaluates the individual contributions of the bottleneck layers preserved by DBP. Since the selected layers primarily consist of embedding and normalization modules, we ablate their effects by independently excluding each component from quantization. While *Emb_Only* improves all three metrics compared to *No_DBP*, *Norm_Only* enhances VQA-Aes but degrades CLIPSIM and VQA-Tech. *Full_DBP* consistently achieves the highest performance across all metrics, demonstrating the necessity of jointly preserving these components.

Notably, DBP adopts a sparse preservation strategy rather than broad mixed-precision expansion. By focusing FP16 preservation on the highest-sensitivity bottlenecks (DS_l), we effectively mitigate directional distortion while maintaining optimal compression efficiency.

E.5 Comparison with Higher-Precision Baselines

To examine the contribution of selective precision allocation, we compare ***TrajeQ*** with ViDiT-Q at increased bit-widths. At nominal W3A6, CAQ increases precision by one bit for 3.16% of activation-quantization calls, giving a call-weighted average of 6.03 bits over these calls. DBP retains less than 1% of parameters in FP16, keeping the parameter-weighted weight precision below 3.13 bits. Tab. 12 compares generation quality across these settings without exactly matching computation or memory usage.

Despite using two additional activation bits, ViDiT-Q at W3A8 remains below ***TrajeQ*** at nominal W3A6 on all four metrics. At W4A8, ViDiT-Q achieves higher CLIPSIM and VQA-Tech scores, while ***TrajeQ*** maintains higher CLIP-Temp and VQA-Aes. Increasing ViDiT-Q's activation precision alone is therefore insufficient to match ***TrajeQ***'s performance in this setting.

Table 12: Comparison with higher-precision ViDiT-Q settings on FLUX/Pyramid using OpenSora prompts. ***TrajeQ***'s W3A6 label denotes its nominal bit-width. Bold and underlined values indicate the best and second-best results among quantized methods, respectively. Higher values are better for all metrics.

Method	Bit-width (W/A)	CLIPSIM	CLIP-Temp	VQA-Aes	VQA-Tech
FP16	16/16	0.3394	0.9983	64.96	64.08
ViDiT-Q	3/6	0.2706	0.9921	49.47	48.80
ViDiT-Q	3/8	0.2888	<u>0.9950</u>	49.88	50.85
ViDiT-Q	4/8	**0.3249**	0.9949	<u>60.08</u>	**59.82**
TrajeQ	3/6	<u>0.3186</u>	**0.9971**	**61.40**	<u>59.29</u>

F GENERALIZATION EXPERIMENTS

F.1 MMDIT ARCHITECTURE (ALTERNATIVE FM BACKBONE)

To assess architectural generality, we extend our evaluation to the MMDiT architecture (Esser et al., 2024). As shown in Tab. 13, ***TrajeQ*** improves over standard PTQ baselines (MinMax, SmoothQuant) and other quantization methods (QuaRot, Q-DiT, ViDiT-Q) across most metrics. The gains are most pronounced under W3A6, where competing methods show substantial quality degradation, while ***TrajeQ*** better preserves generation fidelity and consistency. For example, it improves *Imaging Quality* by $+20.24$ over the strongest baseline. These results suggest that the trajectory-preserving design of ***TrajeQ*** extends beyond a single backbone and remains effective across different flow-matching architectures.

Table 13: Evaluation of text-to-video generation on VBench with MMDiT. The "16" bit-width represents full precision (FP16). **Bold** is the best result, and underline is second-best result. "N/A" denotes instances where valid measurement was impossible due to severe noise.

Method	Bit-width (W/A)	Imaging Quality	Aesthetic Quality	Motion Smooth.	Dynamic Degree	BG. Consist.	Subject Consist.	Scene Consist.	Overall Consist.
-	16/16	64.93	65.14	99.10	66.66	97.16	94.38	43.53	26.27
MinMax	4/6	57.63	53.89	97.21	43.20	94.38	87.52	42.88	23.70
SmoothQuant	4/6	55.72	56.34	98.01	43.93	95.11	88.87	43.03	23.17
QuaRot	4/6	54.46	53.87	97.13	42.11	90.27	86.81	42.36	25.73
Q-DiT	4/6	57.18	57.97	98.43	45.23	88.66	88.69	42.89	25.97
ViDiT-Q	4/6	57.21	58.14	98.34	45.79	95.24	88.97	43.07	**26.24**
TrajeQ	4/6	**60.51**	**60.34**	**98.99**	**45.83**	**96.38**	**90.30**	**43.31**	26.22
MinMax	4/4	51.08	22.12	N/A	N/A	N/A	N/A	N/A	01.97
SmoothQuant	4/4	51.21	23.61	N/A	N/A	N/A	N/A	N/A	01.99
QuaRot	4/4	48.24	22.23	93.72	09.86	82.36	64.88	20.46	11.21
Q-DiT	4/4	53.27	43.06	96.11	16.41	89.33	74.96	25.77	18.29
ViDiT-Q	4/4	54.49	43.12	97.97	17.44	88.25	75.35	26.43	19.95
TrajeQ	4/4	**56.06**	**47.24**	**98.34**	**18.05**	**90.14**	**75.48**	**28.34**	**22.40**
MinMax	3/6	20.70	26.77	95.41	22.22	92.01	73.06	01.38	05.94
SmoothQuant	3/6	21.06	26.66	96.42	15.27	92.26	72.60	00.65	05.85
QuaRot	3/6	15.89	21.72	96.23	17.31	92.73	72.63	08.86	04.93
Q-DiT	3/6	21.36	28.13	96.51	24.10	91.96	72.79	11.36	06.24
ViDiT-Q	3/6	22.06	26.87	96.53	17.50	92.67	74.07	01.30	07.56
TrajeQ	3/6	**42.30**	**32.46**	**98.71**	**28.36**	**93.04**	**74.30**	**18.86**	**13.85**

F.2 DETERMINISTIC DIFFUSION EVALUATION WITH COGVIDEOX-2B

Beyond FM backbones, we further examine whether the trajectory-level degradation discussed in Sec. 2.1 also appears in deterministic video diffusion sampling. To this end, we evaluate CogVideoX-2B (Yang et al., 2024), a diffusion-transformer-based text-to-video model that provides a complementary testbed to our FM-based evaluation. As shown in Tab. 14, ***TrajeQ*** consistently improves low-bit generation quality under both W4A4 and W3A6 quantization. In particular, ***TrajeQ*** achieves stronger text-video alignment and temporal stability, as reflected by improved CLIPSIM, CLIP-Temp, Δ Flow Score, and Warping Error. These results suggest that orientation-sensitive quantization errors are not restricted to FM backbones, but can also arise in deterministic video diffusion sampling.

F.3 LONG-VIDEO GENERATION UNDER LOW-BIT QUANTIZATION

To evaluate the robustness of ***TrajeQ*** over extended temporal horizons, we conduct long-video experiments under the aggressive W3A6 configuration. This setting increases the opportunity for quantization errors to accumulate across autoregressive generation steps. As shown in Tab. 15, ***TrajeQ*** maintains stronger generation quality than competing baselines under this compounded challenge. While baseline methods exhibit pronounced degradation over longer sequences, ***TrajeQ*** better preserves visual fidelity and temporal coherence. These results suggest that the trajectory-preserving design of ***TrajeQ*** extends beyond short clips and remains effective in longer generation horizons.

Table 14: Evaluation on OpenSora prompts with CogVideoX-2B. This deterministic diffusion-sampler setting complements our FM evaluation. "16" denotes full-precision. **Bold** and underline indicate the best and second-best results, respectively. ↓ indicates lower is better.

Method	Bit-width (W/A)	CLIPSIM	CLIP-Temp	VQA-Aesthetic	VQA-Technical	Δ Flow Score (↓)	Warping Error (↓)
-	16/16	0.3460	0.9898	62.76	62.03	-	0.026
MinMax	4/4	0.2384	0.9340	32.33	40.98	2.899	0.053
SmoothQuant	4/4	0.2169	0.9554	33.34	44.51	0.644	0.204
QuaRot	4/4	0.3153	0.9695	43.48	47.68	0.838	0.090
Q-DiT	4/4	0.2065	0.9422	33.32	46.02	0.870	0.246
ViDiT-Q	4/4	0.3127	0.9736	42.83	43.50	1.062	0.056
TrajeQ	4/4	**0.3175**	**0.9741**	**44.82**	**48.81**	**0.322**	**0.036**
MinMax	3/6	0.2194	0.9694	26.32	35.53	2.260	0.115
SmoothQuant	3/6	0.2325	0.9641	31.57	39.61	1.501	0.141
QuaRot	3/6	0.2076	0.9676	32.78	34.30	1.132	0.184
Q-DiT	3/6	0.2050	0.9529	32.80	35.72	0.485	0.205
ViDiT-Q	3/6	0.2280	0.9876	26.63	34.42	0.844	0.036
TrajeQ	3/6	**0.2504**	**0.9886**	**36.59**	**43.34**	**0.390**	**0.033**

Table 15: Evaluation of long-video text-to-video generation on OpenSora prompts with FLUX under W3A6 quantization. We evaluate 8-second clips using 25 sampled frames. The "16" bit-width represents full-precision (FP16). **Bold** indicates the best result, and underline indicates the second-best result. The symbol ↓ indicates that lower values are better.

Method	Bit-width (W/A)	CLIPSIM	CLIP-Temp	VQA-Aesthetic	VQA-Technical	Δ Flow Score (↓)	Warping Error (↓)
-	16/16	0.3183	0.9976	59.42	59.64	-	0.021
MinMax	3/6	0.2321	0.9826	42.11	50.16	0.197	0.060
SmoothQuant	3/6	0.2313	0.9828	42.52	49.90	0.174	0.066
QuaRot	3/6	0.2393	0.9698	37.73	41.34	0.706	0.088
Q-DiT	3/6	0.2456	0.9827	43.84	41.20	0.304	0.049
ViDiT-Q	3/6	0.2686	0.9829	46.29	48.13	0.450	0.062
TrajeQ	3/6	**0.2988**	**0.9892**	**49.81**	**52.12**	**0.137**	**0.039**

G Efficiency Evaluation and Additional Quality Comparisons

G.1 Computational Cost Evaluation

Under W3A6, competing methods suffer severe structural degradation (Fig. 1), whereas ***TrajeQ*** maintains generation fidelity, achieving 0.4231 LPIPS and 0.5516 SSIM (Tab. 16). This comes with a ∼5% generation-time overhead (+1.7h vs. ViDiT-Q). CAQ computes S_{dir} from current and cached activations and adjusts bit-widths across timesteps. The associated activation storage and temporary quantization buffers mainly account for the higher peak memory. DBP retains less than 1% of parameters in FP16, adding little to the estimated quantized model size. Thus, the runtime and memory overhead of simulated quantization accompanies substantial fidelity preservation even as competing methods severely degrade.

G.2 Qualitative Analysis under Extreme W4A4 Quantization

In Tab. 1, MinMax and SmoothQuant are reported as "N/A" under W4A4 because their generated videos collapse into severe noise, preventing VBench feature extractors from computing valid embeddings. As shown in Fig. 10, these methods fail to preserve recognizable structure. This failure is also reflected in Tab. 1, where MinMax and SmoothQuant yield "N/A" for several W4A4 VBench metrics, near-zero Scene Consistency (0.00/0.21), and very low Overall Consistency (3.29/3.28), indicating unreliable video representations.

QuaRot and Q-DiT partially recover global structure but still exhibit visible artifacts and fidelity degradation. ViDiT-Q better preserves coarse structure, yet suffers from luminance shifts and loss

Table 16: Efficiency-quality trade-off on VBench under W3A6. LPIPS and SSIM measure perceptual fidelity. Model size (GB) is estimated from bit-widths. Peak memory (MB) and generation time (h) are measured using floating-point kernels with simulated quantization.

Method	Performance		Efficiency		
	LPIPS (↓)	SSIM (↑)	Size	Mem.	Time
MinMax	0.8469	0.2266	0.69	33608	33.4
SmoothQuant	0.8241	0.2802	0.69	34028	33.5
QuaRot	0.8672	0.2213	0.69	39927	36.2
Q-DiT	0.7318	0.2673	0.81	38690	33.5
ViDiT-Q	0.6837	0.2697	0.70	37499	33.5
TrajeQ	**0.4231**	**0.5516**	0.71	42670	35.2

of detail. In contrast, ***TrajeQ*** maintains more stable visual structure, color fidelity, and temporal coherence, remaining closest to the FP16 baseline under W4A4.

G.3 Additional Qualitative Results

We provide additional qualitative comparisons between ***TrajeQ*** and PTQ baselines, including MinMax, SmoothQuant, QuaRot, Q-DiT, and ViDiT-Q, under the W4A6 configuration. As shown in Figs. 11–16, the examples cover diverse scenarios, including photorealistic scenes, natural landscapes, and artistic styles.

Across these samples, baseline methods exhibit distinct degradation patterns under aggressive weight and activation quantization. **MinMax** often suffers from color and contrast collapse, producing darkened backgrounds, oversaturated regions, or unnatural color shifts in high-dynamic-range scenes such as *Volcano*, *Skyscraper*, and *Mountains*. **SmoothQuant** tends to lose structural detail and texture, especially in stylized or high-frequency scenes such as *Van Gogh*, *Train*, and *Volcano*. **QuaRot** frequently produces blurred outputs with weakened subject identity, as observed in the *Train* and *Astronaut* samples. **Q-DiT** shows strong chromatic distortion and color bleeding, particularly in *Beach*, *Skyscraper*, and *Astronaut*. **ViDiT-Q** generally preserves coarse structure better than other baselines, but still exhibits object distortion and detail loss, such as degraded equipment in *Astronaut* and blurred architectural silhouettes in *Skyscraper*.

In contrast, ***TrajeQ*** better preserves visual fidelity, texture, and semantic structure across these diverse cases. It maintains fine details such as the horse musculature, train surface reflections, and stylized brushwork, while remaining closely aligned with the FP16 reference. These qualitative results support the effectiveness of trajectory-preserving quantization for maintaining perceptual quality under low-bit FM video generation (Figs. 11–16).

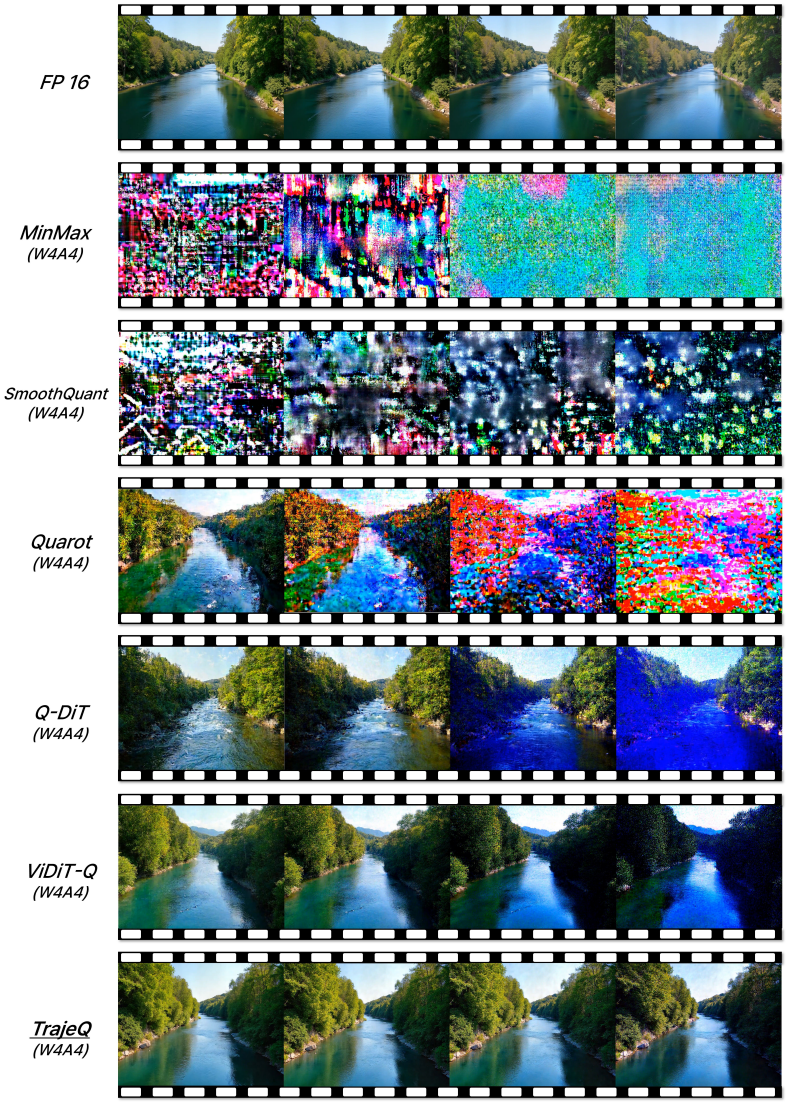


Figure 10: Qualitative comparison under extreme W4A4 quantization. FP16 produces high-fidelity frames, whereas MinMax and SmoothQuant collapse into severe noise. QuaRot and Q-DiT partially recover structure but still suffer from noticeable artifacts and fidelity loss. ViDiT-Q better preserves global structure but shows luminance and detail degradation. In contrast, ***TrajeQ*** maintains visual structure, color fidelity, and temporal coherence closest to the FP16 baseline.



Figure 11: Qualitative comparison at W4A6 using the prompt: "`An astronaut is riding a horse in the space in a photorealistic style.`" Baselines (MinMax, SmoothQuant, QuaRot, Q-DiT) suffer from severe color degradation and compositional deviation, while ViDiT-Q exhibits structural distortion. Conversely, ***TrajeQ*** preserves the intricate textures of the space suit and the muscular definition of the horse, maintaining the original photorealistic composition.

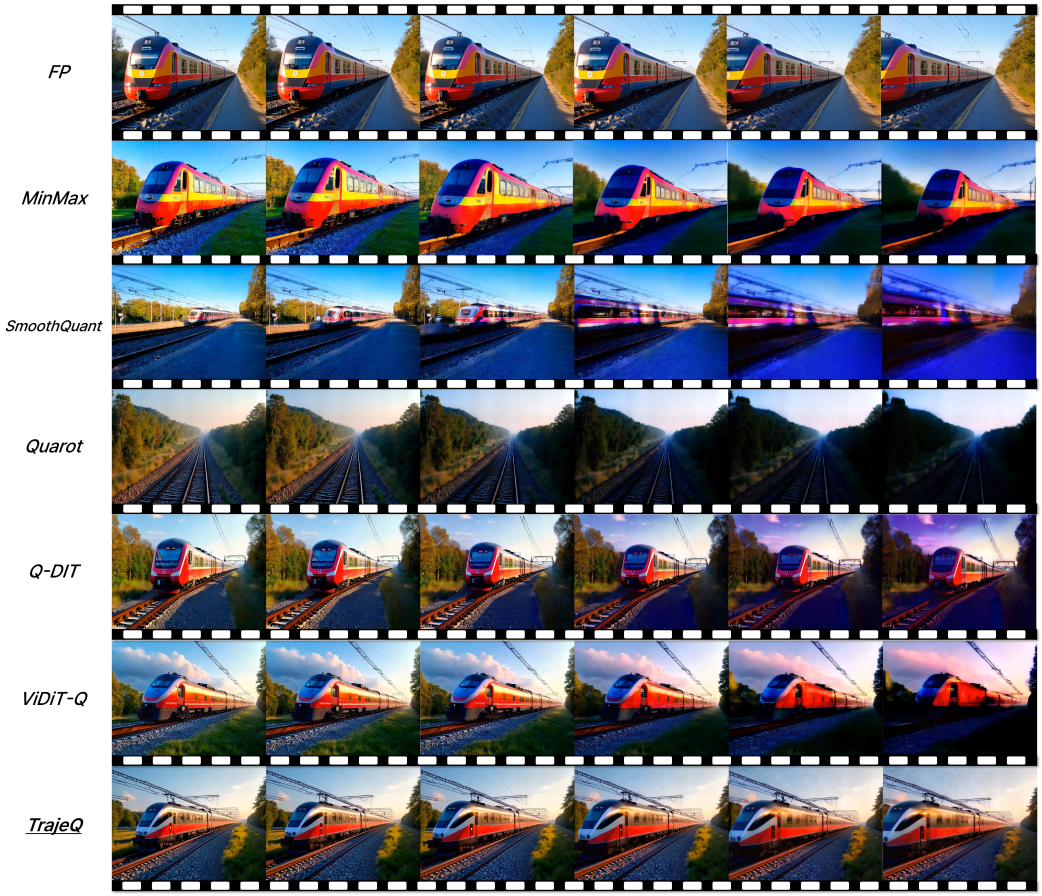


Figure 12: Qualitative comparison at W4A6 using the prompt: "`train railway`". Competing methods display marked artifacts, specifically color oversaturation (MinMax), structural collapse (SmoothQuant), **extreme blurring rendering the subject entirely unidentifiable (QuaRot)**, and detail degradation (Q-DiT, ViDiT-Q). Conversely, ***TrajeQ*** preserves the sharp geometry of the railway tracks and the realistic metallic sheen of the train carriage, keeping the scene crisp and coherent.

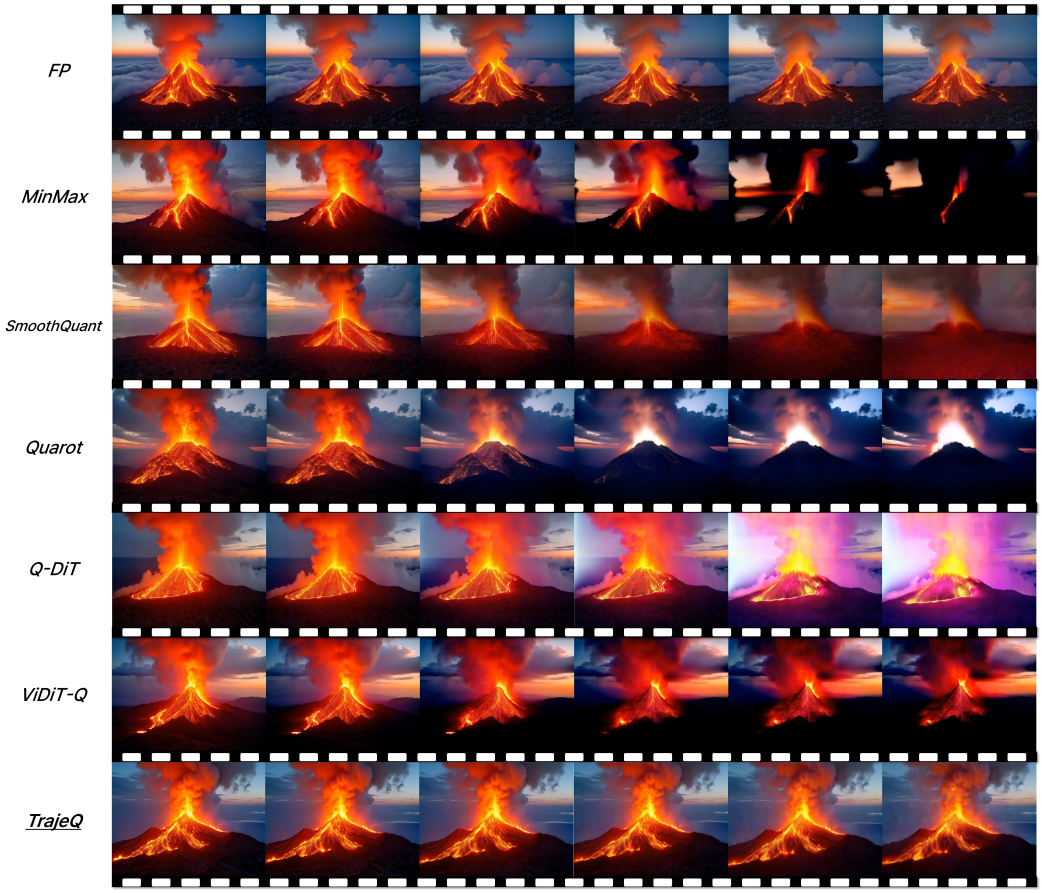


Figure 13: Qualitative comparison at W4A6 using the prompt: "`Volcano`". Baselines reveal substantial quality degradation, particularly severe luminance collapse (MinMax, QuaRot) and texture washout (SmoothQuant, Q-DiT). In comparison, ***TrajeQ*** vividly captures the incandescent glow of the flowing lava and the volumetric density of the ash plumes, maintaining the intense contrast of the eruption.

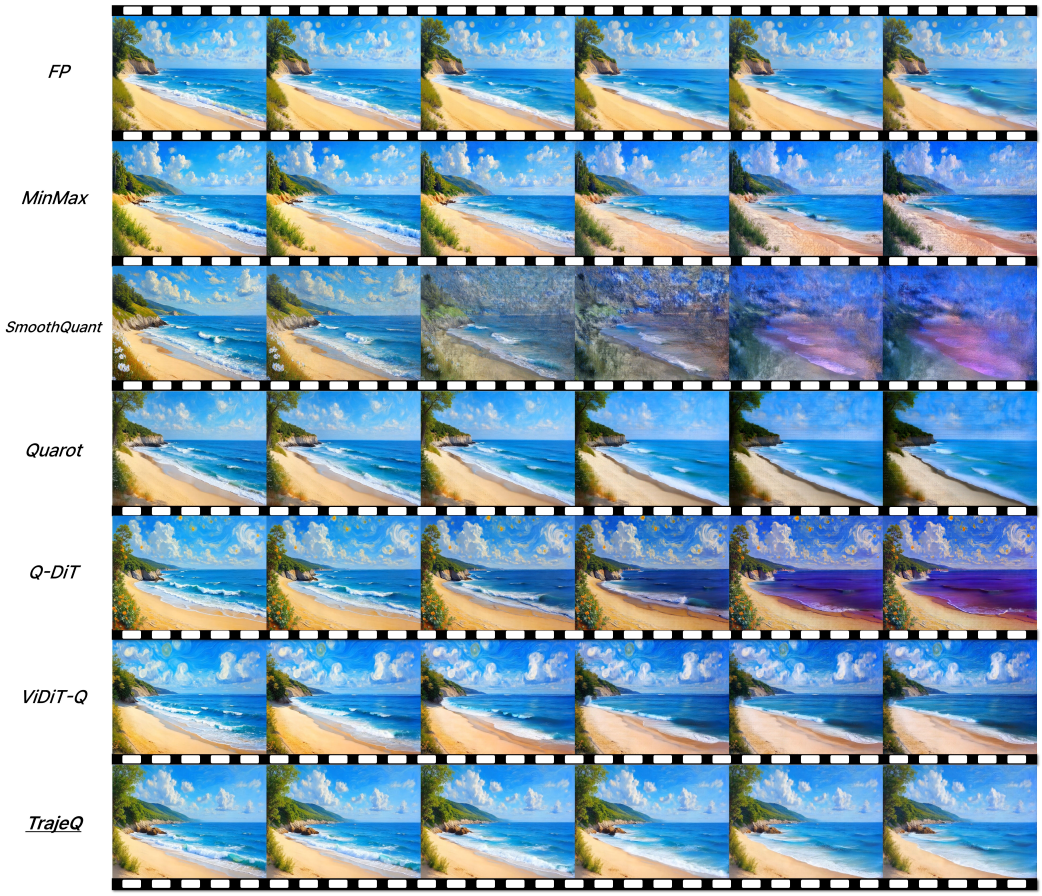


Figure 14: Qualitative comparison at W4A6 using the prompt: "`A beautiful coastal beach in spring, waves lapping on sand by Vincent van Gogh`". QuaRot and SmoothQuant manifest catastrophic coherence loss, whereas MinMax, Q-DiT and ViDiT-Q suffer from texture degradation. In contrast, ***TrajeQ*** successfully reproduces the signature swirling sky patterns and the rhythmic impasto brushwork, capturing the vibrant vitality of the post-impressionist scene.



Figure 15: Qualitative comparison at W4A6 using the prompt: "`Snow rocky mountains peaks canyon. snow blanketed rocky mountains surround and shadow deep canyons. the canyons twist`". Baselines exhibit noticeable visual impairments, such as chromatic distortion (MinMax, Q-DiT, ViDiT-Q) and severe detail smoothing (SmoothQuant, QuaRot). Meanwhile, ***TrajeQ*** faithfully renders the jagged mountain silhouettes and the crisp snow textures, preserving the atmospheric depth of the twisting canyon.



Figure 16: Qualitative comparison at W4A6 using the prompt: "`skyscraper`". Competing methods demonstrate pronounced visual defects, specifically contrast collapse (MinMax, Q-DiT), extreme chromatic deviation (SmoothQuant), and structural blurring (QuaRot, ViDiT-Q). Conversely, ***TrajeQ*** faithfully reconstructs the intricate window textures and the subtle dusk gradients, preserving the sharp architectural silhouette.