

RotateAttention: RoPE-Aware Rotation and Range Rectification for INT4 Quantized Attention in Video Generation

Yaofu Liu^{1*}, Wanli Lan^{2*}, Jinxi Li^{2*}, Binhang Yuan^{1†}, and Harry Yang^{1†}

¹ The Hong Kong University of Science and Technology, Hong Kong SAR, China
yliuls@connect.ust.hk, biyuan@ust.hk, yangharry@ust.hk

² Independent Researcher, China

Abstract. In DiT-based video generation models equipped with **3D Rotary Position Embeddings (3D RoPE)**, the attention mechanism remains a primary computational bottleneck due to its quadratic complexity with respect to sequence length. While quantized **FlashAttention** offers a promising path toward hardware acceleration, existing low-bit quantization methods overlook two critical challenges in this setting: **1)** applying online rotation matrices—a widely used technique for mitigating outliers in Queries (Q) and Keys (K)—is difficult to reconcile with **RoPE**; and **2)** the non-negative attention matrix $P = \exp(QK - \max(QK))$ makes symmetric quantization waste half of the 4-bit dynamic range. In this work, we observe that the outlier distributions of Q and K are strongly affected by the dimensional partitioning of **3D RoPE**. Based on this finding, we propose **RotateAttention**, an efficient **mixed-precision INT4 FlashAttention** framework tailored for **DiT-based video generation models with 3D RoPE**, using selective **FP16 fallback** for accuracy-sensitive attention blocks and denoising steps. RotateAttention introduces two core techniques: **1) RoPE-aware Rotation**, which employs either mergeable rotation matrices that can be fused into RoPE or negligible-overhead matrices to mitigate RoPE-induced outliers in Q and K ; and **2) Range-optimized P Quantization**, which uses fixed scales and zero-points to fully exploit the **INT4 numerical range** with minimal computational overhead. Experiments show that **RotateAttention** preserves video generation quality nearly identical to full-precision baselines while achieving up to $1.68\times$ end-to-end speedup and $2.2\times$ kernel-level acceleration.

Keywords: Video generation · Quantized attention · 3D RoPE

1 Introduction

Diffusion models [26] have achieved remarkable success in video generation; however, they face a severe computational bottleneck within the attention mechanism. Since video generative models often involve sequences exceeding 10K

* Equal contribution.

† Corresponding authors.

tokens, the attention cost scales quadratically with the token count. A promising approach to mitigate this is 4-bit quantization of FlashAttention, a method that is orthogonal to other strategies such as sparse attention [3, 4, 12, 23, 36, 40, 41, 43, 44, 46, 48, 51, 52]. We develop a **highly compatible mixed-precision INT4 attention framework** for DiT-based video generation models with **3D-RoPE**, substantially improving quantization accuracy by selectively falling back to **FP16** in accuracy-sensitive attention blocks and denoising steps.

Outliers and Rotation Transformation The accuracy of low-bit quantized attention is constrained by the presence of *outliers* in the tensors—elements whose magnitudes deviate significantly from the mean distribution. These extreme values compress the dynamic range available for the majority of the tensor, leading to severe precision loss. Rotation transformations are established techniques to mitigate outliers by redistributing tensor values, as discussed in detail in Sec. 2.

Limitations of Existing Work. Current 4-bit attention methods, when applied to DiT-based video generation models with 3D RoPE, face several limitations: 1) The Query (**Q**) and Key (**K**) matrices are usually not rotated, although rotation can effectively equalize outliers. This is because common rotation matrices are difficult to fuse with Rotary Position Embeddings (RoPE) [31], while online rotation introduces non-negligible overhead. 2) Existing methods employ symmetric quantization for the probability matrix $\mathbf{P} = \exp(\mathbf{S} - m)$. Since $\mathbf{P} \in (0, 1]$, this leaves half of the INT4 representation range unused. Furthermore, LLM-oriented quantized attention methods with **Q/K** rotations do not directly transfer to mixed-precision INT4 attention in 3D-RoPE video DiTs.

Motivation and Key Findings. In this work, we make the following observations: 1) We investigate the impact of 3D RoPE on the outlier distribution of **Q** and **K** in 3D-RoPE video DiTs. We identify specific patterns in the outlier distribution that can be exploited to implement a highly efficient and effective rotation matrix. 2) Fully utilizing the 4-bit numerical range for the quantization of **P** yields a significant improvement in attention accuracy.

Method. We introduce **RotateAttention**, which comprises two core innovations complemented by a standard offline technique: 1) **RoPE-aware Rotation:** A block-diagonal matrix composed of 2×2 orthogonal blocks that effectively mitigates outliers in **Q** and **K** with minimal overhead, designed to be compatible with or mergeable into RoPE. 2) **Range-Optimized P Quantization:** A fixed-scale and zero-point calibration for matrix **P** that fully exploits the INT4 dynamic range at low computational cost. Additionally, we apply a **Hadamard Transform for V**, a well-established technique [24, 33] that mitigates outliers in **V** by fusing Hadamard matrices into the existing linear projections at zero inference cost.

Experimental results demonstrate that our method is both efficient and effective. More importantly, our strategy can be integrated into other quantized attention frameworks to further enhance their performance.

2 Related Work

Recent advances in quantization [2, 7, 9, 10, 15–20, 22, 30, 32, 34, 35, 53, 55] have significantly improved model efficiency. For matrix multiplication $\mathbf{Y} = \mathbf{X}\mathbf{W}^\top$, where $\mathbf{X} \in \mathbb{R}^{k \times n}$ and $\mathbf{W} \in \mathbb{R}^{m \times n}$, common strategies for maintaining high accuracy include:

1) *Per-channel Scaling*. This method [42] mitigates outliers by applying a diagonal scaling matrix $\text{diag}(\mathbf{c})$ to transfer outlier energy from $\mathbf{X}$ to $\mathbf{W}$: $\mathbf{Y} = (\mathbf{X}\text{diag}(\mathbf{c})^{-1}) \cdot (\text{diag}(\mathbf{c})\mathbf{W}^\top)$.

2) *Rotation Transformation*. A more general approach uses a transformation matrix $\mathbf{H}$ and its inverse transpose $\mathbf{H}^{-\top}$ such that $\mathbf{Y} = (\mathbf{X}\mathbf{H})(\mathbf{W}\mathbf{H}^{-\top})^\top$. Orthogonal rotations (e.g., Hadamard) [1] redistribute outlier energy across channels to reduce maximum values within each matrix (*intra-matrix*). Non-orthogonal invertible transforms act as a composite of rotation and scaling.

Low-bit Attention in Video Generation. Current low-bit video attention methods focus on quantizing $\mathbf{Q}\mathbf{K}^\top$ and $\mathbf{P}\mathbf{V}$ using alternative techniques: (1) **QKV Smoothing**: The SageAttention series [47, 50] uses INT4/INT8 formats, employing channel smoothing (e.g., $\mathbf{Q} - \mathbb{E}[\mathbf{Q}]$) to suppress biased outliers. SageAttention3 [49] further extends this line with NVFP4 micro-scaling formats. (2) **Token Reordering**: PAROAttention [54] reorders tokens in $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ so that each block of $\mathbf{P}$ contains values with similar magnitudes, thereby improving per-block quantization precision. This token-level strategy is *complementary* to our channel-level rotation approach and can potentially be combined for further gains.

Low-bit Attention in LLMs. LLM quantization typically targets the memory-bound KV cache [21]. To mitigate quantization error, rotation-based methods like FlatQuant and SpinQuant [24, 33] utilize invertible transforms, while **QuaRot** [1] employs Hadamard matrices. Recent advances, such as FlashAttention-3 [29], further optimize efficiency by reducing rotation overhead within high-performance kernels.

Key Differences: LLM vs. 3D-RoPE Video DiTs. Fig. 1 illustrates the key differences between LLM decoding and 3D-RoPE video DiT attention, highlighting why rotation strategies designed for LLM KV-cache quantization do not directly transfer to DiT-based video generation models with 3D RoPE.

We identify two fundamental distinctions: (1) **Computational Intensity**: LLM decoding is memory-bound [25, 28, 57], allowing online rotations to overlap with RoPE or attention kernels. In contrast, 3D-RoPE video DiTs are compute-bound, making online rotation overhead significant and difficult to hide. (2) **Quantization Symmetry**: LLM quantization often keeps $\mathbf{Q}$ in FP16 and only quantizes $\mathbf{K}$ for the KV-cache. Conversely, 3D-RoPE video DiTs require simultaneous quantization of $\mathbf{Q}$ and $\mathbf{K}$ to leverage low-bit matmul hardware for actual speedup.

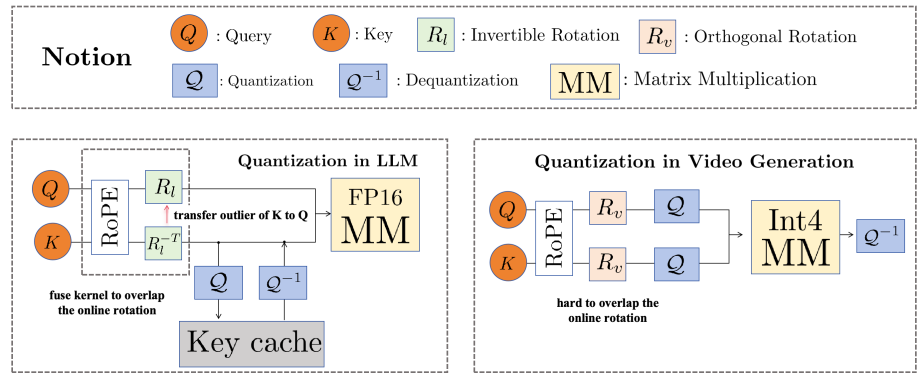


Fig. 1: Quantization differences between LLM decoding and 3D-RoPE video DiTs. In LLMs, rotations can prioritize **K** quantization and transfer part of the outlier burden to **Q**. In 3D-RoPE video DiTs, both **Q** and **K** must be quantized simultaneously, which requires orthogonal transforms and low-overhead execution.

Furthermore, we extend Hadamard transforms to the Value (**V**) matrix. While previously used in LLM KV-caches, we formally prove their compatibility with smoothing for **V**. By bridging these rotation strategies and enhancing **P** quantization, **RotateAttention** significantly improves mixed-precision INT4 attention quality for 3D-RoPE video DiTs.

3 RoPE Effect on Outlier Distribution

Let **Q**, **K**, and **V** denote the query, key, and value matrices, respectively, each belonging to $\mathbb{R}^{N \times D}$, where N represents the sequence length and D denotes the feature dimension.

In DiT-based video generation models, 3D RoPE is widely adopted to capture complex spatiotemporal dependencies. Theoretically, the feature dimension D of the **Q** and **K** matrices is partitioned into **three distinct segments**— D_f , D_h , and D_w —corresponding to the temporal (frame), height, and width dimensions, respectively, such that $D = D_f + D_h + D_w$.

The 3D RoPE rotary matrix $\mathbf{M} \in \mathbb{R}^{D \times D}$ is a block-diagonal matrix composed of $D/2$ fundamental rotation blocks $\mathbf{M}_i \in \text{SO}(2)$. This matrix is partitioned into three sub-matrices— $\mathbf{M}^f$, $\mathbf{M}^h$, and $\mathbf{M}^w$ —each responsible for encoding temporal, height, and width information within their respective dimensional segments.

Implementation. Due to the inherent sparsity of the rotary matrix $\mathbf{M}$, the computation of RoPE substitutes the standard matrix multiplication with a more efficient element-wise operation.

Outlier Metric. To analyze the distribution of outliers that predominantly affect quantization precision, we introduce the *incoherence* metric $\Psi(\mathbf{X})$, which

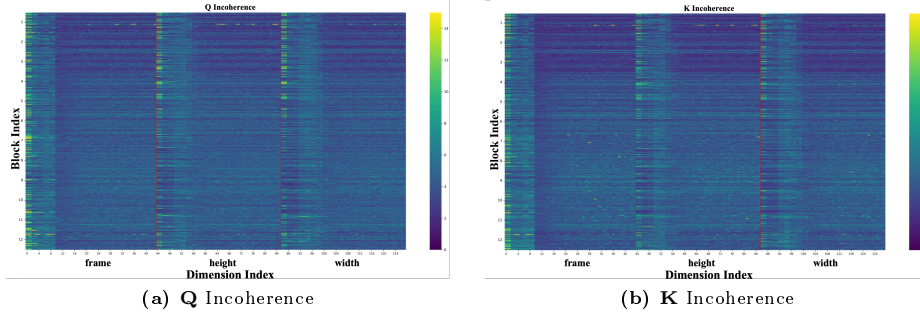


Fig. 2: Channel-wise incoherence profiles of **Q and **K** for Wan2.2.** Incoherence is partitioned into three regions corresponding to the 3D RoPE spatiotemporal segments. We observe strong distributional symmetry between **Q** and **K**, and structured sparsity within each segment.

captures the maximum deviation within each channel. For a given matrix $\mathbf{X} \in \{\mathbf{Q}, \mathbf{K}\}$, we first compute the sequence-wise mean vector $\mathbb{E}[\mathbf{X}] \in \mathbb{R}^{1 \times D}$:

$$\mathbb{E}[\mathbf{X}] = \frac{1}{N} \sum_{i=1}^N \mathbf{X}_{i,:}, \quad (1)$$

where $\mathbf{X}_{i,:}$ denotes the i -th row of $\mathbf{X}$. The incoherence $\Psi(\mathbf{X}) \in \mathbb{R}^{1 \times D}$ is then defined as:

$$\Psi(\mathbf{X}) = \max_{i \in \{1, \dots, N\}} |\mathbf{X}_{i,:} - \mathbb{E}[\mathbf{X}]|, \quad (2)$$

where the absolute value and maximum operations are applied element-wise. This metric provides a channel-level statistical profile, reflecting the outlier concentration within **Q** and **K** that poses challenges for fixed-point quantization.

3.1 Data Distribution of **Q** and **K**

By evaluating the incoherence $\Psi(\mathbf{X})$ for $\mathbf{X} \in \{\mathbf{Q}, \mathbf{K}\}$, as shown in Fig. 2, we identify two key patterns in the outlier distribution:

Spatiotemporal Segmental Incoherence. Outliers are strictly partitioned into three specific segments—frame, height, and width—matching the channel segments defined by the 3D RoPE. Within these segments, we observe that one half of the dimensions typically contains the vast majority of the incoherence, while the remaining half remains relatively smooth.

Cross-Matrix Symmetry. We find that the outlier distribution patterns of **Q** and **K** exhibit high symmetry. Specifically, channels at identical indices in **Q** and **K** generally possess nearly equivalent incoherence magnitudes.

Theoretical Interpretation. The segmented outlier structure is consistent with 3D RoPE frequency assignment: low-frequency channel pairs accumulate larger angular spread over long sequences and thus higher variance, while high-frequency pairs remain comparatively smooth. Because $\mathbf{Q}$ and $\mathbf{K}$ share RoPE parameters and are projected from the same hidden states, their channel-wise outlier profiles remain symmetric.

4 Method

We focus on attention modules in DiT-based video generation models equipped with 3D RoPE. We consider queries ($\mathbf{Q}$), keys ($\mathbf{K}$), and values ($\mathbf{V}$), which are equipped with biases shared across all tokens. To address this, we first perform smoothing on $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$.

Furthermore, we observe that $\mathbf{Q}$ and $\mathbf{K}$ exhibit specific distributional patterns of outliers. Based on these patterns, we apply a specially designed rotation for the Query and Key matrices to redistribute and equalize these outliers, thereby improving quantization robustness.

4.1 Rotation for Query and Key

Outlier Equalization via Orthogonal Rotation. Motivated by these observations, we introduce an orthogonal rotation matrix $\mathbf{R} \in \mathbb{R}^{D \times D}$ to equalize the outlier distribution. We define $\mathbf{R}$ as a block-diagonal matrix:

$$\mathbf{R} = \text{diag} \left(\{\mathbf{R}_i\}_{i=1}^{D/2} \right), \quad (3)$$

where each $\mathbf{R}_i \in \text{O}(2)$ is a 2×2 orthogonal matrix. When the rotation matrix $\mathbf{R}$ is non-trainable, we initialize the 2×2 orthogonal blocks $\mathbf{R}_i$ with a 2×2 Hadamard matrix, defined as:

$$\mathbf{H}_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}. \quad (4)$$

Mathematical Analysis on the Necessity of Orthogonality To maintain computational equivalence, when a transformation matrix $\mathbf{R}$ is applied to $\mathbf{Q}$, the matrix $\mathbf{K}$ must be transformed by $\mathbf{R}^{-\top}$:

$$\mathbf{QK}^\top = (\mathbf{QR})(\mathbf{KR}^{-\top})^\top. \quad (5)$$

This equality holds for any invertible $\mathbf{R}$ in full precision. However, under low-bit quantization, the transformation not only changes the coordinate system, but also directly affects the numerical ranges and outlier distributions of the transformed $\mathbf{Q}$ and $\mathbf{K}$. Since $\mathbf{Q}$ and $\mathbf{K}$ usually exhibit symmetric and comparable outlier patterns after RoPE, the transformation should preserve their balance rather than suppressing outliers in one matrix while amplifying them in the other.

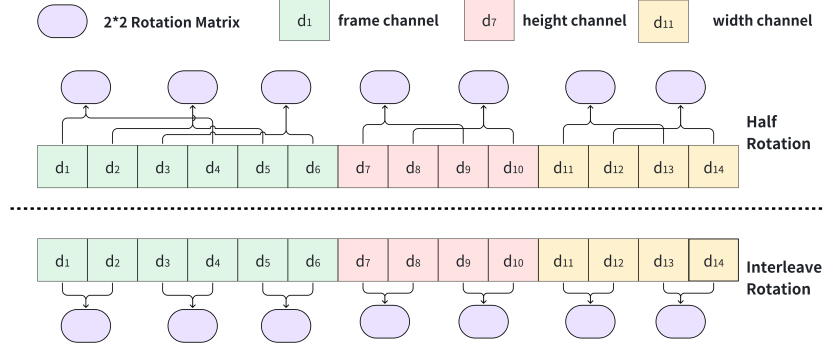


Fig. 3: Rotation strategies for $\mathbf{Q}$ and $\mathbf{K}$. Interleaved Rotation groups adjacent dimensions into 2×2 orthogonal blocks, while Half Rotation pairs each dimension with its counterpart in the other half-segment.

Let $\mathbf{R} = \mathbf{U}\mathbf{\Sigma}\mathbf{L}^\top$ be the SVD of $\mathbf{R}$, where $\mathbf{U}, \mathbf{L} \in \mathbb{O}(D)$ and $\mathbf{\Sigma} = \text{diag}(\sigma_1, \dots, \sigma_D)$. The corresponding transformation on $\mathbf{K}$ is

$$\mathbf{R}^{-\top} = ((\mathbf{U}\mathbf{\Sigma}\mathbf{L}^\top)^{-1})^\top = \mathbf{U}\mathbf{\Sigma}^{-1}\mathbf{L}^\top. \quad (6)$$

Therefore, the singular values applied to $\mathbf{K}$ are $\{1/\sigma_i\}_{i=1}^D$. If $\sigma_i < 1$, the i -th principal direction of $\mathbf{Q}$ is compressed, but the corresponding direction of $\mathbf{K}$ is amplified by $1/\sigma_i > 1$. Conversely, if $\sigma_i > 1$, the outliers in $\mathbf{Q}$ are amplified while those in $\mathbf{K}$ are suppressed. In both cases, anisotropic scaling introduces an asymmetric range change between $\mathbf{Q}$ and $\mathbf{K}$, which can enlarge the quantization error of either matrix.

To avoid this cross-matrix amplification and preserve balanced quantization difficulty, the transformation should impose the same singular scaling on both branches. This requires $\mathbf{\Sigma} = \mathbf{\Sigma}^{-1}$, i.e., $\sigma_i = 1$ for all i . Consequently, $\mathbf{\Sigma} = \mathbf{I}$ and $\mathbf{R} = \mathbf{U}\mathbf{L}^\top$, which is an orthogonal matrix. This analysis explains why orthogonal rotations are particularly suitable for the symmetric outlier patterns induced by RoPE, as shown in Fig. 2.

Interleaved Rotation (Merge-able) To integrate with RoPE, we design *Interleaved Rotation*. We define a block-diagonal orthogonal matrix $\mathbf{R}$, where each 2×2 block acts on adjacent channel pairs (d_{2i}, d_{2i+1}) for $i \in \{1, \dots, D/2\}$. This transformation redistributes incoherence, yielding a more uniform outlier profile across the feature dimensions of $\mathbf{Q}$ and $\mathbf{K}$.

The core advantage of this design is its **operator mergeability**. Since both the rotary matrix $\mathbf{M}$ and our rotation $\mathbf{R}$ consist of 2×2 orthogonal blocks, their product $\mathbf{M}' = \mathbf{R}\mathbf{M}$ (or $\mathbf{M}\mathbf{R}$) preserves the block-diagonal structure. Consequently, the operations can be fused:

$$\text{RoPE}(\mathbf{Q})\mathbf{R} = \mathbf{Q}(\mathbf{M}\mathbf{R}), \quad \text{RoPE}(\mathbf{K})\mathbf{R} = \mathbf{K}(\mathbf{M}\mathbf{R}). \quad (7)$$

Because $\mathbf{M}\mathbf{R}$ maintains the same sparsity pattern as $\mathbf{M}$, it remains fully compatible with standard element-wise RoPE kernels. By pre-fusing $\mathbf{R}$ into the RoPE

weights, we achieve outlier redistribution with **zero inference overhead**, providing a "free" improvement to quantization precision.

Half Rotation Since outliers predominantly occupy one half of each segment, we introduce **Half Rotation**—a low-cost strategy that couples dimensions d_i^j and $d_{i+D_j/2}^j$ for $i \in \{1, \dots, D_j/2\}$. Here, d_i^j represents the i -th dimension of segment $j \in \{f, h, w\}$.

We define a block-diagonal permutation matrix $\mathbf{\Pi} = \text{diag}(\mathbf{\Pi}^f, \mathbf{\Pi}^h, \mathbf{\Pi}^w)$, where each $\mathbf{\Pi}^j$ reorders the dimensions of segment j such that dimensions separated by $D_j/2$ are moved to adjacent positions. The rotation matrix $\mathbf{R}$ is similarly structured as $\mathbf{R} = \text{diag}(\mathbf{R}^f, \mathbf{R}^h, \mathbf{R}^w)$, where each $\mathbf{R}^j$ consists of $D_j/2$ orthogonal blocks $\mathbf{R}_k^j \in O(2)$.

The final rotated query matrix is obtained via segment-wise transformation and concatenation:

$$\mathbf{QIR} = [\mathbf{Q}_f \mathbf{\Pi}^f \mathbf{R}^f, \mathbf{Q}_h \mathbf{\Pi}^h \mathbf{R}^h, \mathbf{Q}_w \mathbf{\Pi}^w \mathbf{R}^w]. \quad (8)$$

For each segment, the operation $\mathbf{Q}_j(\mathbf{\Pi}^j \mathbf{R}^j)$ exploits the inherent sparsity of the combined permutation-rotation matrix. Consequently, this matrix multiplication can be transformed into a highly efficient element-wise operation, analogous to the optimized implementation of RoPE. Through Half Rotation, incoherence is effectively equalized across the entire segment while maintaining high computational efficiency.

Fig. 3 summarizes the two proposed RoPE-aware rotation strategies. Interleaved Rotation groups adjacent channel dimensions and can be merged into RoPE, while Half Rotation pairs dimensions across the two half-segments to equalize RoPE-induced outliers with negligible overhead.

Learnable Rotation (Preliminary) As a preliminary investigation, we also explore making the rotation **learnable**. Let $\mathbf{Y} = \mathbf{Q}^{(s)}(\mathbf{K}^{(s)})^\top$ denote the full-precision attention score matrix computed from the smoothed query and key matrices. We learn the rotation matrix $\mathbf{R}$ by minimizing the following calibration objective with gradient-based optimization:

$$\mathbf{R}^* = \arg \min_{\{\mathbf{R}_i\}} \left\| \mathbf{Y} - \mathcal{Q}(\mathbf{Q}^{(s)} \mathbf{R}) \mathcal{Q}((\mathbf{K}^{(s)} \mathbf{R})^\top) \right\|_F^2, \quad (9)$$

$$\text{s.t. } \mathbf{R}_i \in O(2), \quad i = 1, \dots, D/2. \quad (10)$$

Here, $\mathbf{Q}^{(s)}$ and $\mathbf{K}^{(s)}$ are the smoothed matrices, and $\mathcal{Q}(\cdot)$ denotes the quantization function. As shown in Sec. 4.1, this orthogonality constraint is essential—without it, the rotation degrades attention output quality, confirming that LLM-style unconstrained rotations [24, 33] do not transfer to mixed-precision INT4 attention in 3D-RoPE video DiTs. This experiment serves primarily to validate the orthogonality requirement; a thorough calibration study is left to future work (see Sec. 5).

4.2 Range-optimized P Quantization

In the FlashAttention framework [5, 6, 29], the attention scores are computed block-wise. Let $\mathbf{S}_{i,j} = \mathbf{Q}_i \mathbf{K}_j^\top / \sqrt{D}$ denote the pre-softmax attention scores for a given block. To maintain numerical stability during the online softmax process, a running row maximum $m_{i,j} = \max(m_{i,j-1}, \text{rowmax}(\mathbf{S}_{i,j}))$ is tracked, and the local attention probabilities are defined as $\mathbf{P}_{i,j} = \exp(\mathbf{S}_{i,j} - m_{i,j})$.

Affine Range Expansion Standard quantization practices typically normalize the attention probabilities by their maximum value to fit the target bit-width. For INT4 symmetric quantization, the mapping is:

$$\tilde{\mathbf{P}}_{i,j}^{\text{sym}} = \left\lfloor \frac{\mathbf{P}_{i,j}}{\max(\mathbf{P}_{i,j})} \times 7 \right\rfloor \in (0, 7]. \quad (11)$$

This approach, however, leaves the negative representative space of INT4 ($[-8, 0)$) entirely unused, effectively wasting half of the available quantization levels and leading to significant precision loss.

To maximize the utilization of the 4-bit dynamic range, we propose a mapping with a fixed scale of 15 and a zero-point of -8 . This affine transformation projects the normalized probabilities onto the full INT4 interval $[-8, 7]$:

$$\tilde{\mathbf{P}}_{i,j} = \left(\frac{\mathbf{P}_{i,j}}{\max(\mathbf{P}_{i,j})} \times 15 \right) - 8 \in (-8, 7]. \quad (12)$$

By expanding the representation from 8 levels (0 to 7) to 16 levels (-8 to 7), we effectively double the quantization resolution, significantly reducing the rounding error.

4.3 Hadamard Transformation for the Value Matrix

To mitigate the impact of outliers within the Value matrix $\mathbf{V}$, we employ an orthogonal rotation transformation using a $D \times D$ Hadamard matrix $\mathbf{H}$, where $D = 2^n$ ($n \in \mathbb{Z}^+$). The 2×2 Hadamard matrix is defined in Eq. (4). Hadamard matrices are constructed recursively via the Kronecker product:

$$\mathbf{H}_{2^n} = \mathbf{H}_2 \otimes \mathbf{H}_{2^{n-1}}. \quad (13)$$

The Hadamard rotation can be mathematically absorbed into the value projection $\mathbf{W}_v$ and the output projection $\mathbf{W}_o$. Given the value projection $\mathbf{V} = \mathbf{XW}_v$, the standard attention output is:

$$\begin{aligned} \mathbf{O} &= \text{Softmax} \left(\frac{\mathbf{QK}^\top}{\sqrt{d_k}} \right) \mathbf{VW}_o \\ &= \text{Softmax} \left(\frac{\mathbf{QK}^\top}{\sqrt{d_k}} \right) \mathbf{X}(\mathbf{W}_v \mathbf{H})(\mathbf{H}^\top \mathbf{W}_o). \end{aligned} \quad (14)$$

We can merge the Hadamard transform into the weights as $\widetilde{\mathbf{W}}_v = \mathbf{W}_v \mathbf{H}$ and $\widetilde{\mathbf{W}}_o = \mathbf{H}^\top \mathbf{W}_o$, respectively. Since these modified weights are computed offline during the preprocessing stage, the Hadamard transformation for $\mathbf{V}$ is achieved with **zero additional computational overhead** during inference.

5 Experiments

5.1 Experimental Setup

We evaluate our proposed method on the **Wan2.2** [37] T2V and I2V models at 480P resolution with 81 frames (40 sampling steps), and the **HunyuanVideo** [13] T2V model at 480P resolution with 129 frames (50 sampling steps). Our evaluation dataset comprises diverse prompts curated from the official Wan2.2, Wan-Lightning [8], and Open-Sora [56] example sets, following the evaluation protocol of SageAttention [50]. For the attention mechanism, we employ **per-token quantization** for the query ($\mathbf{Q}$) and key ($\mathbf{K}$) matrices, and **per-group quantization** for the value ($\mathbf{V}$) matrix.

For **Wan2.2-I2V**, we preserve FP16 precision for the first 6 and last 4 sampling steps, as well as the first and last 2 DiT blocks; all remaining FlashAttention operations are executed with our INT4 attention kernel, resulting in 68% of FlashAttention operations running in INT4. For **Wan2.2-T2V**, FP16 is maintained for the first and last 2 sampling steps and DiT blocks, enabling 81% INT4 attention execution. Similarly, for **HunyuanVideo-T2V**, we keep FP16 for the first 4 and last 2 sampling steps and the first and last 2 DiT blocks, with 76% of attention operations running in INT4.

Hardware Configuration. We benchmark INT4 attention kernel speed on **NVIDIA A10** GPUs using custom CUDA kernels, reflecting practical deployment hardware. For accuracy evaluation of the full video generation pipeline, we use **NVIDIA H20** GPUs with **TileLang** [38] operators, which provide precise numerical validation.

Model Scope. We evaluate on Wan2.2 and HunyuanVideo, two leading open-source video generation models, both employing DiT-based architectures with 3D RoPE. This choice is deliberate: DiT with 3D RoPE has become the *de facto* standard for state-of-the-art open-source video generation (also adopted by CogVideoX [45], Open-Sora 2.0 [27], and others), making it the most practically relevant setting.

5.2 Evaluation Metrics

We primarily employ **Relative Difference Metrics** to quantify the discrepancy between our quantized models and the FP16 baseline. Specifically, **PSNR** [14] and **Cosine Similarity** measure low-level pixel-space differences, while **SSIM** [39] evaluates structural similarity. These relative metrics are more

sensitive to the fine-grained artifacts introduced by quantization [54], providing a clearer signal for comparing quantization strategies.

We also evaluate absolute video quality using **VBench** [11], which assesses semantic coherence, temporal consistency, and motion quality. Since VBench scores are less discriminative across quantization variants (methods with visible pixel-level differences often receive similar VBench scores), we use relative metrics as the primary comparison in the main paper and report VBench results in the supplementary material.

5.3 Baselines and Ablation Study

The primary baseline is the standard attention mechanism in **FP16**. We also compare against **SageAttention** [50] with **QKV** smoothing but without our optimized **P** and rotation strategies. Furthermore, we evaluate a full rotation baseline using a $D \times D$ Hadamard matrix. Extensive **ablation studies** are performed to analyze the impact of different rotation and optimized **P** strategies. For the Half Rotation and Interleaved Rotation results in Table 1, the rotation matrix **R** defaults to the block-diagonal initialization using **H₂**.

When applying rotations to **Q** and **K**, we include a Hadamard transformation on **V** by default. Since this transformation is computationally mergeable—and follows standard practice in frameworks like FlatQuant and SpinQuant [24, 33]—we adopt it as a default component and omit separate ablation studies for this operation.

We note that PAROAttention [54] addresses the complementary problem of token reordering for **P** quantization, which operates at the token level rather than the channel level. Our rotation strategy and PAROAttention’s reordering are orthogonal techniques that can be combined. We provide a software-level comparison and combination experiment in the supplementary material, demonstrating additive accuracy gains when both approaches are applied.

Learned Rotation Setup. As a preliminary investigation into learned rotations (see Sec. 4.1), we train under two conditions: **(1) Unconstrained** and **(2) Orthogonal**. We use Half Rotation with **H₂** initialization on Wan2.2-I2V and 1–2 calibration samples.

5.4 Experimental Results

Table 1 summarizes the impact of different rotation strategies on the Wan2.2 (I2V/T2V) and HunyuanVideo models.

Quantitative Analysis. Difference metrics indicate that **Half Rotation** achieves the best performance for Wan2.2-I2V in terms of **Cosine Similarity** and **MSE**, while remaining highly competitive in **SSIM** and **PSNR**. On HunyuanVideo, Half Rotation yields the best Cosine Similarity, MSE, and SSIM. Notably, combining Half Rotation with **Optimized P Quantization** consistently outperforms SageAttention across all tasks.

Table 1: Difference Metrics for Wan2.2 I2V, T2V and HunyuanVideo T2V

Model	Method			Difference Metrics				Speedup
	Datatype	Optim-P	Rotation	Cosine $\uparrow$	MSE $\downarrow$	SSIM $\uparrow$	PSNR $\uparrow$	
Wan2.2 -I2V	FP16	–	–	1.0000	0.0000	1.0000	∞	–
	INT4	$\times$	W/O	0.9809	519.30	0.7752	22.828	
	INT4	$\checkmark$	W/O	0.9789	575.17	0.7524	22.069	
	INT4	$\checkmark$	Half	0.9824	493.35	0.7767	22.903	
	INT4	$\checkmark$	Interleave	0.9792	583.88	0.7502	22.057	
	INT4	$\checkmark$	Full	0.9807	515.34	0.7808	23.176	1.51 $\times$
	INT4	$\checkmark$	Trained	0.9809	519.94	0.7746	22.713	
Wan2.2 -T2V	FP16	–	–	1.0000	0.0000	1.0000	∞	–
	INT4	$\times$	W/O	0.9458	1574.5	0.6348	17.829	
	INT4	$\checkmark$	W/O	0.95178	1410.1	0.6582	18.427	
	INT4	$\checkmark$	Half	0.9492	1466.0	0.6504	18.462	
	INT4	$\checkmark$	Interleave	0.95622	1392.1	0.6628	18.544	1.68 $\times$
	INT4	$\checkmark$	Full	0.95270	1370.2	0.6737	19.211	
	INT4	$\checkmark$	Full	0.95270	1370.2	0.6737	19.211	
Hyvideo -T2V	FP16	–	–	1.0000	0.0000	1.0000	∞	–
	INT4	$\times$	W/O	0.9663	1095.3	0.6621	18.332	
	INT4	$\checkmark$	W/O	0.9712	1089.6	0.6620	18.945	
	INT4	$\checkmark$	Half	0.9751	1041.9	0.6677	18.971	1.61 $\times$
	INT4	$\checkmark$	Interleave	0.9750	1045.8	0.6588	19.037	
	INT4	$\checkmark$	Full	0.9744	1076.4	0.6646	18.854	
	INT4	$\checkmark$	Full	0.9744	1076.4	0.6646	18.854	

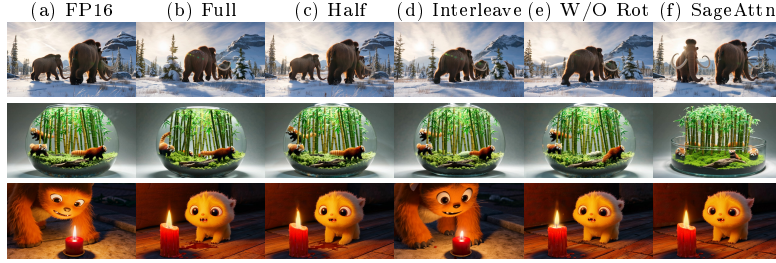
**Fig. 4:** Comparison of different rotation strategies in Wan2.2 T2V models.**Fig. 5:** Comparison of different rotation strategies in Wan2.2 I2V models (top row: FP16/Full/Half; bottom row: Interleave/W/O Rot/SageAttn).



Fig. 6: Comparison of effect of trained Rotation (left to right: FP16, trained with orthogonality, trained without orthogonality).

The **Interleaved Rotation** method significantly improves Cosine Similarity for Wan2.2-I2V, though gains in other metrics are marginal. However, on Wan2.2-T2V, it achieves the highest Cosine Similarity and delivers substantial improvements in MSE, SSIM, and PSNR—second only to Full Rotation. For HunyuanVideo, this method achieves the best PSNR. While its performance on Wan2.2-I2V and HunyuanVideo is comparable to SageAttention, it demonstrates clear superiority on the Wan2.2-T2V model.

Range-Optimized P Quantization yields clear improvements on Wan2.2-T2V and HunyuanVideo across all metrics, yet provides marginal gains on Wan2.2-I2V. We attribute this to the I2V model’s image conditioning, which constrains the attention distribution $\mathbf{P}$ to be sharply peaked (low entropy) with values concentrated near 0 and 1. In this regime, standard symmetric quantization effectively preserves the dominant structure, while range-optimized quantization can introduce subtle precision degradation. In contrast, T2V models produce more diffuse $\mathbf{P}$ distributions that benefit substantially from the full INT4 dynamic range. These findings suggest that range-optimized quantization is most beneficial for moderate-to-high entropy attention, and may be disabled for strongly conditioned, low-entropy patterns.

Analysis of Rotation Strategy Differences. Performance differences match RoPE-induced outlier structure: **Half Rotation** is strongest when outliers concentrate in one half-segment (I2V and HunyuanVideo), while **Interleaved Rotation** is better when local adjacent-channel correlation dominates (T2V). **Full Rotation** can over-smooth useful local structure despite stronger global redistribution.

Practical recommendation: For deployment on new models, we recommend **Half Rotation** as the default strategy due to its consistent performance across all tested configurations. Interleaved Rotation offers a compelling zero-cost alternative when it can be pre-fused into RoPE weights, particularly for T2V applications.

Visual inspection in Fig. 4 reveals that for T2V tasks, Interleaved and Half Rotation occasionally outperform Full Rotation in specific cases. Moreover, as shown in Figs. 4 and 5, both strategies exhibit clear visual superiority over

the SageAttention baseline, particularly in maintaining structural integrity and temporal consistency.

Learnable Rotation Result. Learned rotation is effective *only* under strict orthogonality constraints, as shown in Fig. 6; unconstrained training degrades quality (Sec. 4.1), confirming that LLM-style unconstrained rotations [24, 33] do not transfer directly to 3D-RoPE video DiTs. With 1–2 calibration samples, orthogonal learned rotation does not outperform fixed Half Rotation.

Speed and Cost Analysis. Our INT4 kernel yields $2.2\times$ speedup versus custom FP16 attention, and end-to-end speedups of $1.51\times$ (Wan2.2-I2V), $1.68\times$ (Wan2.2-T2V), and $1.61\times$ (HunyuanVideo). Full rotation scales linearly with sequence length, whereas attention scales quadratically. Consequently, for sequences exceeding 30k tokens, full rotation overhead remains under 1% of attention computation. However, when paired with sparse attention on shorter sequences (<10k tokens), its relative overhead increases to over 10% of the attention computation. Notably, Half Rotation reduces this overhead to approximately 3% of the full rotation cost.

6 Conclusion

In this work, we propose **RotateAttention**, an efficient mixed-precision INT4 attention framework designed for DiT-based video generation models with 3D RoPE. Our approach introduces two core innovations: (1) **RoPE-aware Rotation**, which utilizes mergeable or negligible-cost rotation matrices to effectively mitigate **Q** and **K** outliers while maintaining seamless integration with RoPE; (2) **Range-optimized P Quantization**, which leverages fixed scales and zero-points to fully exploit the **INT4 numerical range** with minimal computational overhead. Additionally, we apply a standard Hadamard transform to **V**, fused offline into the linear projections at zero inference cost.

Experimental results demonstrate that **RotateAttention** achieves video generation quality nearly identical to full-precision baselines and significantly outperforms SageAttention in visual fidelity. Our INT4 attention kernel achieves a $2.2\times$ speedup at the kernel level, translating to end-to-end speedups of $1.51\times$ – $1.68\times$ across the evaluated 3D-RoPE video DiTs. **RotateAttention** establishes a practical paradigm for high-performance, mixed-precision INT4 attention in 3D-RoPE video DiTs.

Limitations. Our rotation strategies are specifically designed for DiT-based video generation models employing 3D RoPE with segmented dimensional partitioning. Applying them to other positional encoding schemes requires adapting the rotation structure to the corresponding outlier patterns. Our evaluation covers two model families, Wan2.2 and HunyuanVideo; evaluating a broader range of architectures would further strengthen generalizability.

References

1. Ashkboos, S., Mohtashami, A., Croci, M.L., Li, B., Cameron, P., Jaggi, M., Alistarh, D., Hoefer, T., Hensman, J.: Quarot: Outlier-free 4-bit inference in rotated llms. *Advances in Neural Information Processing Systems* **37**, 100213–100240 (2024)
2. van Breugel, B., Bondarenko, Y., Whatmough, P., Nagel, M.: Fptquant: Function-preserving transforms for llm quantization. *arXiv preprint arXiv:2506.04985* (2025)
3. Chen, Y., Qian, S., Tang, H., Lai, X., Liu, Z., Han, S., Jia, J.: LongLoRA: Efficient fine-tuning of long-context large language models. In: *The Twelfth International Conference on Learning Representations* (2024)
4. Chu, X., Tian, Z., Wang, Y., Zhang, B., Ren, H., Wei, X., Xia, H., Shen, C.: Twins: Revisiting the design of spatial attention in vision transformers. *Advances in neural information processing systems* **34**, 9355–9366 (2021)
5. Dao, T.: Flashattention-2: Faster attention with better parallelism and work partitioning. In: *The Twelfth International Conference on Learning Representations* (2024)
6. Dao, T., Fu, D., Ermon, S., Rudra, A., Ré, C.: Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in neural information processing systems* **35**, 16344–16359 (2022)
7. Dettmers, T., Lewis, M., Belkada, Y., Zettlemoyer, L.: Gpt3. int8 (): 8-bit matrix multiplication for transformers at scale. *Advances in neural information processing systems* **35**, 30318–30332 (2022)
8. Fan, X., Qiu, Z., Wu, Z., Wang, F., Lin, Z., Ren, T., Lin, D., Gong, R., Yang, L.: Phased dmd: Few-step distribution matching distillation via score matching within subintervals. *arXiv preprint arXiv:2510.27684* (2025)
9. Frantar, E., Ashkboos, S., Hoefer, T., Alistarh, D.: Gptq: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323* (2022)
10. Huang, Y., Gong, R., Liu, J., Ding, Y., Lv, C., Qin, H., Zhang, J.: Qvgen: Pushing the limit of quantized video generative models. *arXiv preprint arXiv:2505.11497* (2025)
11. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., 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 (2024)
12. Jiang, H., LI, Y., Zhang, C., Wu, Q., Luo, X., Ahn, S., Han, Z., Abdi, A.H., Li, D., Lin, C.Y., Yang, Y., Qiu, L.: MInference 1.0: Accelerating pre-filling for long-context LLMs via dynamic sparse attention. In: *The Thirty-eighth Annual Conference on Neural Information Processing Systems* (2024)
13. Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., et al.: Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603* (2024)
14. Korhonen, J., You, J.: Peak signal-to-noise ratio revisited: Is simple beautiful? In: *2012 Fourth international workshop on quality of multimedia experience*. pp. 37–38. IEEE (2012)
15. Li, M., Lin, Y., Zhang, Z., Cai, T., Li, X., Guo, J., Xie, E., Meng, C., Zhu, J.Y., Han, S.: Svdquant: Absorbing outliers by low-rank components for 4-bit diffusion models. *arXiv preprint arXiv:2411.05007* (2024)

16. Li, S., Hu, Y., Ning, X., Liu, X., Hong, K., Jia, X., Li, X., Yan, Y., Ran, P., Dai, G., et al.: Mbq: Modality-balanced quantization for large vision-language models. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 4167–4177 (2025)
17. Li, S., Ning, X., Wang, L., Liu, T., Shi, X., Yan, S., Dai, G., Yang, H., Wang, Y.: Evaluating quantized large language models. In: *Forty-first International Conference on Machine Learning* (2024)
18. Li, X., Liu, Y., Lian, L., Yang, H., Dong, Z., Kang, D., Zhang, S., Keutzer, K.: Q-diffusion: Quantizing diffusion models. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 17535–17545 (2023)
19. Lin, H., Xu, H., Wu, Y., Cui, J., Zhang, Y., Mou, L., Song, L., Sun, Z., Wei, Y.: Duquant: Distributing outliers via dual transformation makes stronger quantized LLMs. In: *The Thirty-eighth Annual Conference on Neural Information Processing Systems* (2024)
20. Lin, J., Tang, J., Tang, H., Yang, S., Chen, W.M., Wang, W.C., Xiao, G., Dang, X., Gan, C., Han, S.: Awq: Activation-aware weight quantization for on-device llm compression and acceleration. *Proceedings of machine learning and systems* **6**, 87–100 (2024)
21. Liu, T., Li, S., Yang, J., Zhao, T., Zhou, F., Song, X., Dai, G., Yan, S., Yang, H., Wang, Y.: PM-KVQ: Progressive mixed-precision KV cache quantization for long-cot LLMs. In: *The Fourteenth International Conference on Learning Representations* (2026)
22. Liu, X., Li, Z., Xiao, J., Chen, M., Li, J., Gu, Q.: Eda-dm: Enhanced distribution alignment for post-training quantization of diffusion models. *arXiv preprint arXiv:2401.04585* (2024)
23. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 10012–10022 (2021)
24. Liu, Z., Zhao, C., Fedorov, I., Soran, B., Choudhary, D., Krishnamoorthi, R., Chandra, V., Tian, Y., Blankevoort, T.: Spinqant: LLM quantization with learned rotations. In: *The Thirteenth International Conference on Learning Representations* (2025)
25. Patel, P., Choukse, E., Zhang, C., Shah, A., Goiri, Í., Maleki, S., Bianchini, R.: Splitwise: Efficient generative llm inference using phase splitting. In: *2024 ACM/IEEE 51st Annual International Symposium on Computer Architecture (ISCA)*. pp. 118–132. IEEE (2024)
26. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 4195–4205 (2023)
27. Peng, X., Zheng, Z., Shen, C., Young, T., Guo, X., Wang, B., Xu, H., Liu, H., Jiang, M., Li, W., et al.: Open-sora 2.0: Training a commercial-level video generation model in \$200 k. *arXiv preprint arXiv:2503.09642* (2025)
28. Qin, R., Li, Z., He, W., Cui, J., Tang, H., Ren, F., Ma, T., Cai, S., Zhang, Y., Zhang, M., et al.: Mooncake: A kvcache-centric disaggregated architecture for llm serving. *ACM Transactions on Storage* (2024)
29. Shah, J., Bikshandi, G., Zhang, Y., Thakkar, V., Ramani, P., Dao, T.: Flashattention-3: Fast and accurate attention with asynchrony and low-precision. In: *The Thirty-eighth Annual Conference on Neural Information Processing Systems* (2024)

30. Shang, Y., Yuan, Z., Xie, B., Wu, B., Yan, Y.: Post-training quantization on diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1972–1981 (2023)
31. Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., Liu, Y.: Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing* **568**, 127063 (2024)
32. Sui, Y., Li, Y., Kag, A., Idelbayev, Y., Cao, J., Hu, J., Sagar, D., Yuan, B., Tulyakov, S., Ren, J.: Bitsfusion: 1.99 bits weight quantization of diffusion model. *Advances in Neural Information Processing Systems* **37**, 76775–76818 (2024)
33. Sun, Y., Liu, R., Bai, H., Bao, H., Zhao, K., Li, Y., JiaxinHu, Yu, X., Hou, L., Yuan, C., Jiang, X., Liu, W., Yao, J.: Flatquant: Flatness matters for LLM quantization. In: Forty-second International Conference on Machine Learning (2025)
34. Tang, S., Wang, X., Chen, H., Guan, C., Wu, Z., Tang, Y., Zhu, W.: Post-training quantization with progressive calibration and activation relaxing for text-to-image diffusion models. In: European Conference on Computer Vision. pp. 404–420. Springer (2024)
35. Tao, K., You, H., Sui, Y., Qin, C., Wang, H.: Plug-and-play 1. x-bit kv cache quantization for video large language models. arXiv preprint arXiv:2503.16257 (2025)
36. Venkataramanan, S., Ghodrati, A., Asano, Y.M., Porikli, F., Habibian, A.: Skip-attention: Improving vision transformers by paying less attention. In: The Twelfth International Conference on Learning Representations (2024)
37. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
38. Wang, L., Cheng, Y., Shi, Y., Tang, Z., Mo, Z., Xie, W., Ma, L., Xia, Y., Xue, J., Yang, F., et al.: Tilelang: A composable tiled programming model for ai systems. arXiv preprint arXiv:2504.17577 (2025)
39. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* **13**(4), 600–612 (2004)
40. Xi, H., Yang, S., Zhao, Y., Xu, C., Li, M., Li, X., Lin, Y., Cai, H., Zhang, J., Li, D., Chen, J., Stoica, I., Keutzer, K., Han, S.: Sparse video-gen: Accelerating video diffusion transformers with spatial-temporal sparsity. In: Forty-second International Conference on Machine Learning (2025)
41. Xiao, C., Zhang, P., Han, X., Xiao, G., Lin, Y., Zhang, Z., Liu, Z., Sun, M.: Infflm: Training-free long-context extrapolation for llms with an efficient context memory. *Advances in neural information processing systems* **37**, 119638–119661 (2024)
42. Xiao, G., Lin, J., Seznec, M., Wu, H., Demouth, J., Han, S.: Smoothquant: Accurate and efficient post-training quantization for large language models. In: International conference on machine learning. pp. 38087–38099. PMLR (2023)
43. Xiao, G., Tian, Y., Chen, B., Han, S., Lewis, M.: Efficient streaming language models with attention sinks. arXiv preprint arXiv:2309.17453 (2023)
44. Yang, S., Xi, H., Zhao, Y., Li, M., Zhang, J., Cai, H., Lin, Y., Li, X., Xu, C., Peng, K., Chen, J., Han, S., Keutzer, K., Stoica, I.: Sparse videogen2: Accelerate video generation with sparse attention via semantic-aware permutation. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
45. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer. arXiv preprint arXiv:2408.06072 (2024)
46. Yuan, Z., Zhang, H., Pu, L., Ning, X., Zhang, L., Zhao, T., Yan, S., Dai, G., Wang, Y.: Dtfastattn: Attention compression for diffusion transformer models. *Advances in Neural Information Processing Systems* **37**, 1196–1219 (2024)

47. Zhang, J., Huang, H., Zhang, P., wei, J., Zhu, J., Chen, J.: Sageattention2: Efficient attention with smoothing q and per-thread quantization. In: First Workshop on Scalable Optimization for Efficient and Adaptive Foundation Models (2025)
48. Zhang, J., Wang, H., Jiang, K., Yang, S., Zheng, K., Xi, H., Wang, Z., Zhu, H., Zhao, M., Stoica, I., et al.: Sla: Beyond sparsity in diffusion transformers via fine-tunable sparse-linear attention. arXiv preprint arXiv:2509.24006 (2025)
49. Zhang, J., wei, J., Wang, H., Zhang, P., Xu, X., Huang, H., Jiang, K., Zhu, J., Chen, J.: Sageattention3: Microscaling FP4 attention for inference and an exploration of 8-bit training. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
50. Zhang, J., wei, J., Zhang, P., Zhu, J., Chen, J.: Sageattention: Accurate 8-bit attention for plug-and-play inference acceleration. In: The Thirteenth International Conference on Learning Representations (2025)
51. Zhang, J., Xiang, C., Huang, H., wei, J., Xi, H., Zhu, J., Chen, J.: Spargeattention: Accurate and training-free sparse attention accelerating any model inference. In: Forty-second International Conference on Machine Learning (2025)
52. Zhang, P., Chen, Y., Su, R., Ding, H., Stoica, I., Liu, Z., Zhang, H.: Fast video generation with sliding tile attention. arXiv preprint arXiv:2502.04507 (2025)
53. Zhao, T., Fang, T., Huang, H., Wan, R., Soedarmadji, W., Liu, E., Li, S., Lin, Z., Dai, G., Yan, S., Yang, H., Ning, X., Wang, Y.: Vidit-q: Efficient and accurate quantization of diffusion transformers for image and video generation. In: The Thirteenth International Conference on Learning Representations (2025)
54. Zhao, T., Hong, K., Yang, X., Xiao, X., Li, H., Ling, F., Xie, R., Chen, S., Zhu, H., Yichong, Z., Wang, Y.: PAROAttention: Pattern-aware reordering for efficient sparse and quantized attention in visual generation models. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
55. Zhao, T., Ning, X., Fang, T., Liu, E., Huang, G., Lin, Z., Yan, S., Dai, G., Wang, Y.: 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)
56. Zheng, Z., Peng, X., Yang, T., Shen, C., Li, S., Liu, H., Zhou, Y., Li, T., You, Y.: Open-sora: Democratizing efficient video production for all. arXiv preprint arXiv:2412.20404 (2024)
57. Zhong, Y., Liu, S., Chen, J., Hu, J., Zhu, Y., Liu, X., Jin, X., Zhang, H.: Dist-serve: Disaggregating prefill and decoding for goodput-optimized large language model serving. In: 18th USENIX Symposium on Operating Systems Design and Implementation (OSDI 24). pp. 193–210 (2024)