INT4 QUANTIZATION FOR FLASHATTENTION

Yaofu Liu, Harry Yang

An Exploration for Large-Scale Video Diffusion Models

MOTIVATION: WHY INT4 FLASHATTENTION?

- >> Attention remains the dominant bottleneck in large video diffusion models due to extremely long sequences (10k–50k tokens) and high resolutions.
- >> While FlashAttention is efficient, another order-of-magnitude speedup is required.
- >> Low-bit quantization (INT4) is the most promising path, but attention is numerically fragile, especially post-softmax.
- >> **Goal:** Use INT4 for as much of FlashAttention as possible without degrading video quality.

BACKGROUND: FLASHATTENTION COMPUTATION

- >> FlashAttention computes attention block-wise for memory efficiency, avoiding materialization of the large attention matrix.
- >> The core operations involve matrix products for scores, an online softmax, and aggregation of value vectors.
- >> Key quantization targets are the input matrices (Q, K, V) and the numerically sensitive post-softmax attention matrix (P).

$$S_{ij} = Q_i K_j^T / \sqrt{d} \quad P_{ij} = \exp(S_{ij} - m_i) \quad O_i = \frac{P_{ij} V_j}{\sum_j \exp(S_{ij} - m_i)}$$

STATE OF THE ART: EXISTING METHODS

SageAttention Series

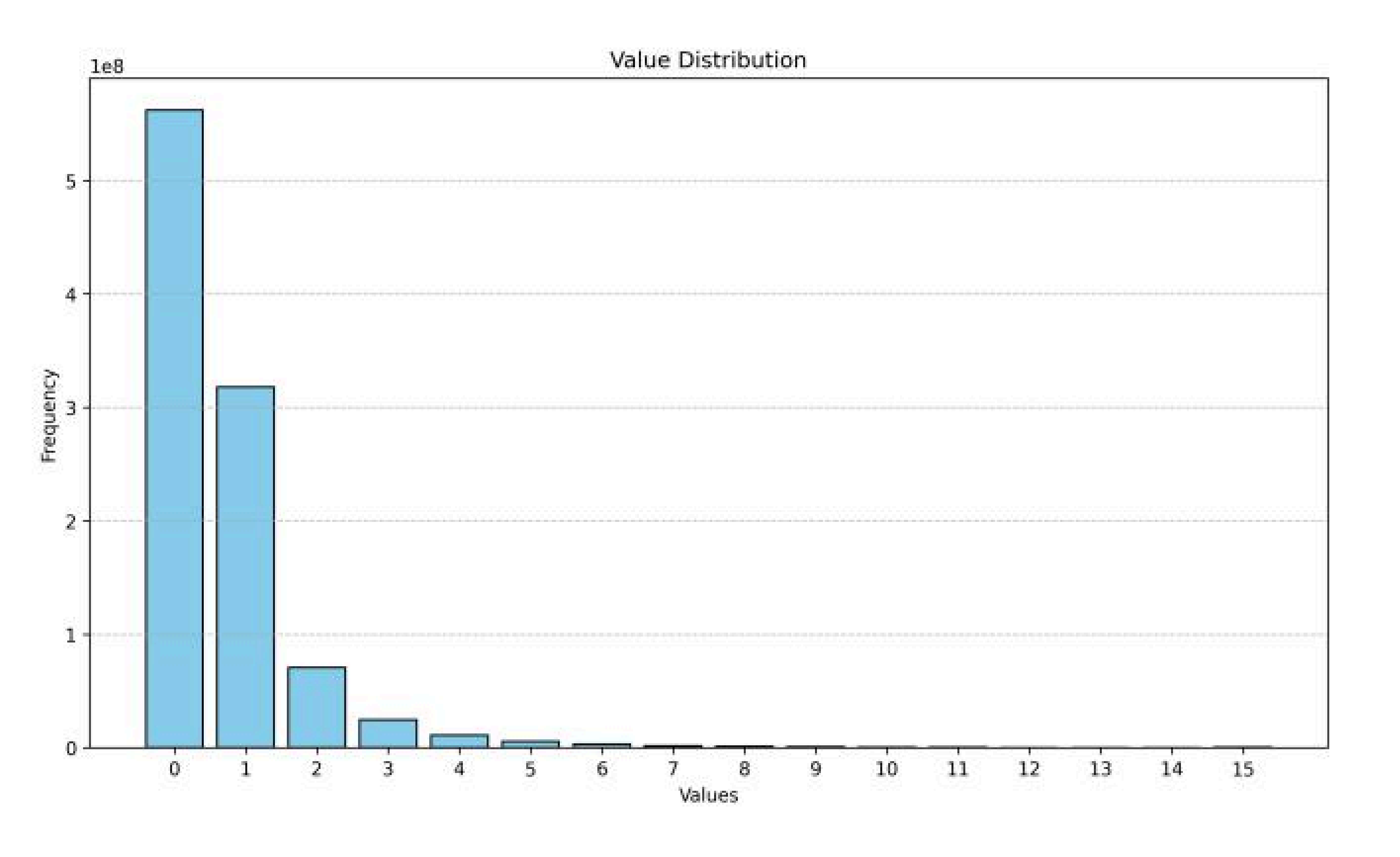
- >> Applies mean-centering (Smooth Quantization) to Q, K, V to suppress outliers.
- >> Quantizes the post-softmax matrix P using max-based scaling to FP4.
- >> **Limitations:** Relies on FP4 which requires specific hardware (Blackwell GPUs) and involves non-standard formats.

PAROAttention

- >> Focuses on the post-softmax matrix P, not QKV.
- >> Addresses the skewed distribution of P by reordering tokens based on spatial locality before softmax computation.
- >> **Limitations:** The reordering is heuristic and limited to fixed spatial permutations.

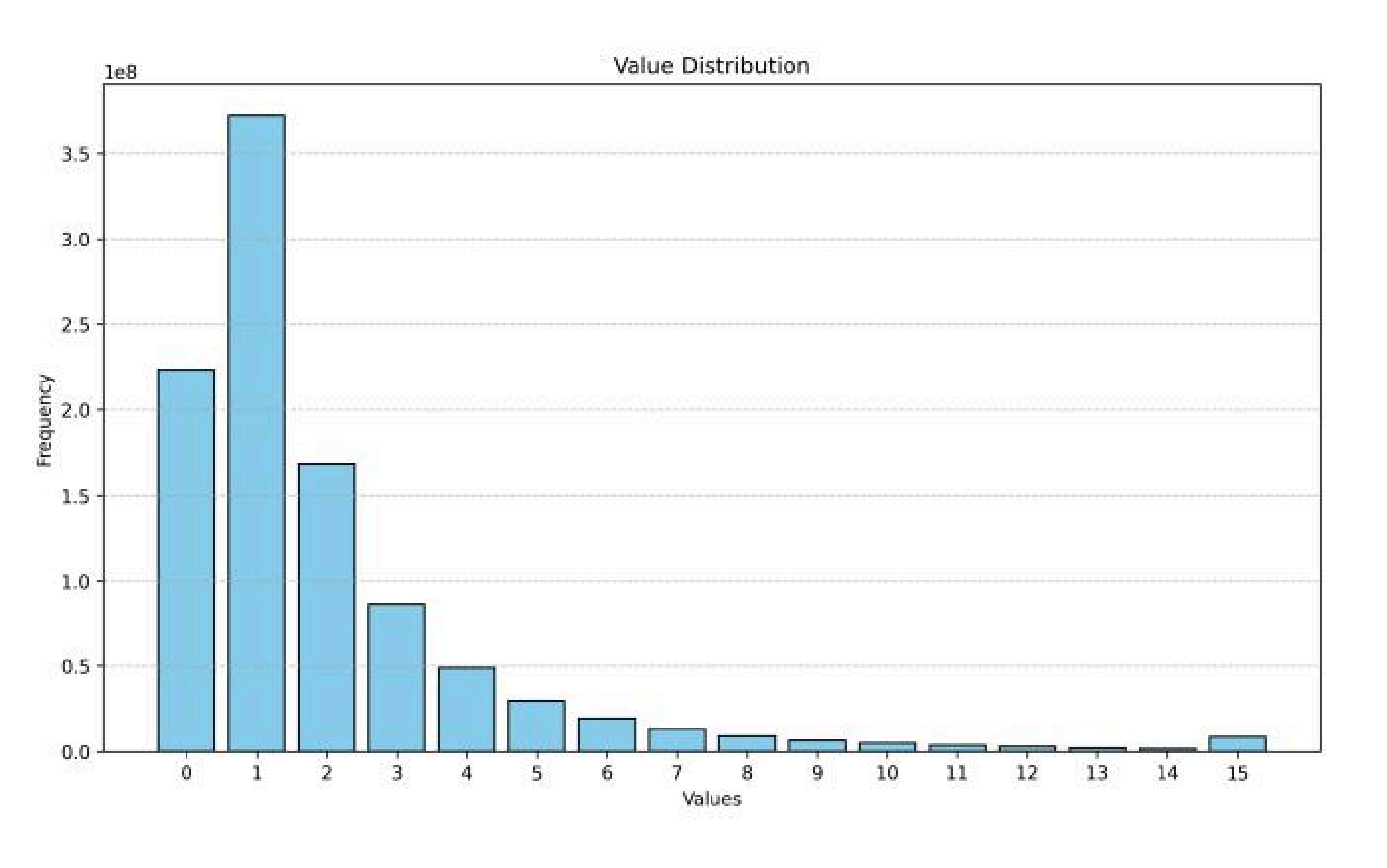
THE CORE CHALLENGE: WHY INT4 IS HARD

- >> **P (Post-Softmax Matrix):** The distribution is highly skewed and concentrated near zero. Standard INT4 quantization wastes over half its representational range (e.g., [-8, 7] for values in (0, 1]).
- >> **QK (Query/Key Matrices):** Outliers are a significant problem. Standard solutions like inserting a rotation matrix conflict with Rotary Positional Embeddings (RoPE) in diffusion models with long sequences, making fusion infeasible.



INNOVATION 1: QUANTIZING P WITH FIXED SCALE-ZERO

- **Fixed Scale-Zero INT4:** Instead of dynamic scaling, use a fixed affine transform $\hat{P} = P \times 15 - 8$ to map P from (0, 1] to the full INT4 range [-8, 7].
- **Local-Max Softmax:** Use the local max of each block ($m_i^j = \operatorname{rowmax}(S_{ij})$) instead of a global running max. This ensures $\max(P)=1$ for every block, guaranteeing full INT4 range utilization.
- **Fusion Trick:** The multiplication by 15 can be fused into the exp operation as a simple addition ($2^{S-\max+\log_2 15}$), making it highly efficient.



INNOVATION 2: ROPE-COMPATIBLE QK ROTATION

>> **Problem:** Generic rotation matrices used for outlier suppression cannot be efficiently fused with the existing RoPE matrices used in video models.

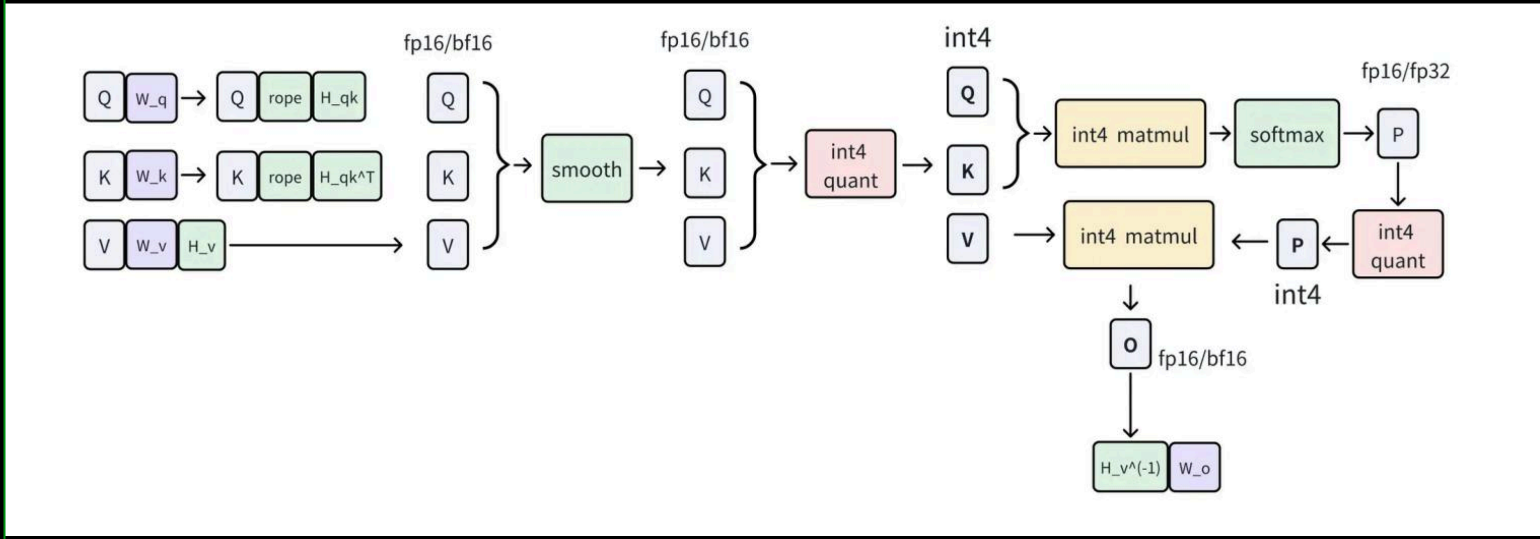
>> **Insight:** Design a rotation matrix H with the same sparse, block-diagonal structure as RoPE matrices.

>> **Solution:** Construct H from simple, sparse blocks like a Hadamard matrix. This allows the rotation to be fused with RoPE, incurring no extra runtime cost while preserving the attention output ($(QH)(KH)^T = QK$).

$$H_2 = \begin{bmatrix} 1 & 1\,1 & -1 \end{bmatrix} \quad H = \begin{bmatrix} H_2 & \mathbf{0} & \dots & \mathbf{0}\,\mathbf{0} & H_2 & \dots & \mathbf{0}\,\dots & \dots & \ddots & \dots\,\mathbf{0} & \mathbf{0} & \dots & H_2 \end{bmatrix}$$

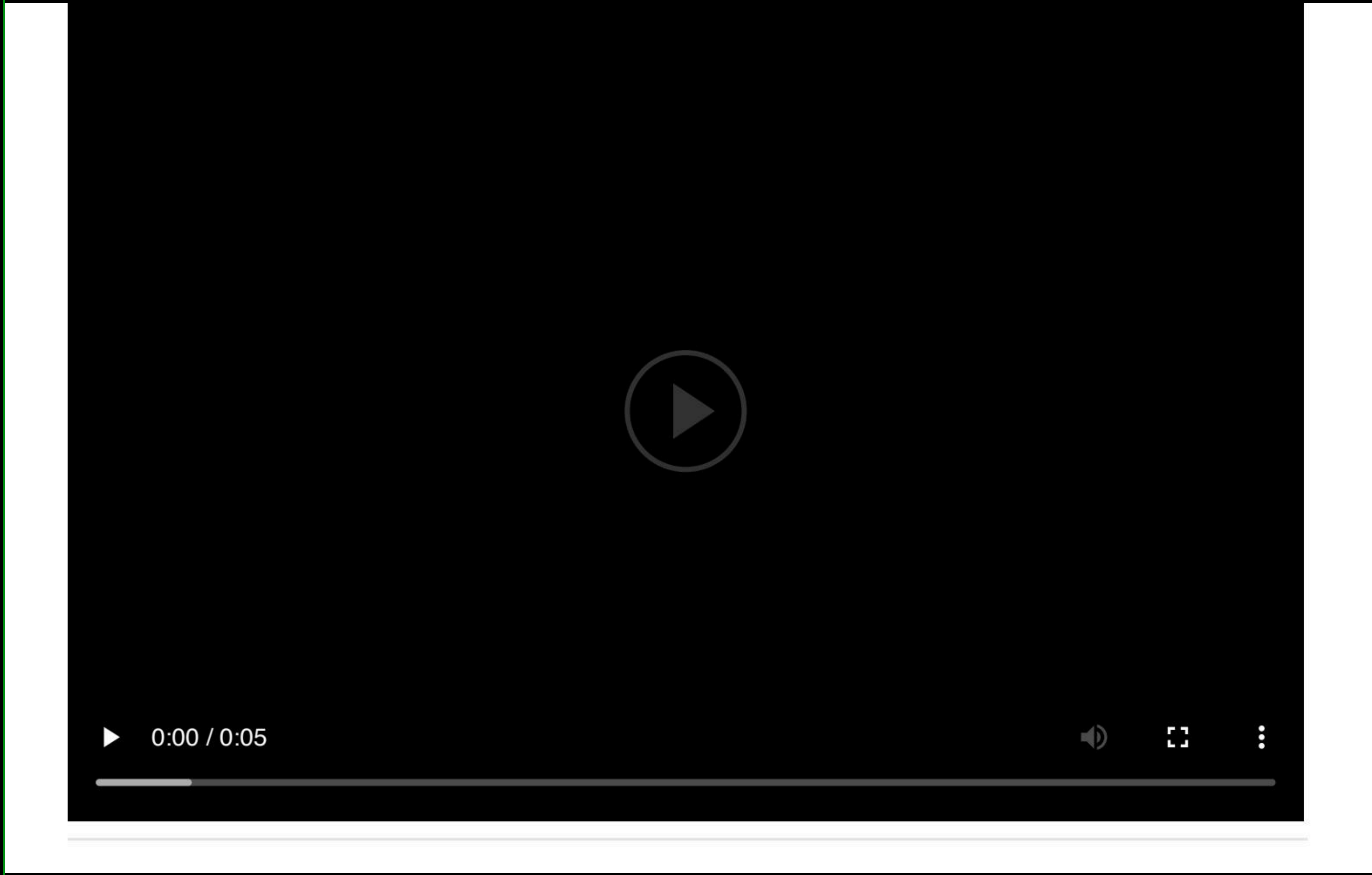
PROPOSED WORKFLOW OVERVIEW

- >> Input Q, K, V are first processed with SmoothQuant (mean-centering).
- >> A sparse, RoPE-compatible rotation (Hadamard) is applied to Q and K to manage outliers.
- >> Attention scores are computed, followed by a local-max softmax.
- >> The resulting P matrix is quantized using our efficient fixed scale-zero INT4 method.
- >> The final output is computed by aggregating V vectors.



CURRENT RESULTS & OPEN PROBLEMS

- **Configuration:** A hybrid approach is used, with 74.2% of attention computation successfully performed in INT4.
- **480p Videos:** Results are strong and visually comparable to FP16.
- **720p Videos (Open Problem):** Videos become noticeably blurry. This is likely due to accumulated quantization noise at higher spatial resolutions, which demands stronger QK outlier suppression.



CONCLUSION

- >> INT4 offers a portable, hardware-agnostic path to accelerate attention, but requires careful handling of numerical precision.
- >> Our innovations in P and QK quantization create a viable INT4 FlashAttention for video models.
- >> **Future Work:**
  - >> Integrate and study learnable sparse rotations (e.g., from FlatQuant).
  - >> Explore advanced token reordering via clustering to improve P's distribution.
  - >> Reduce the computational overhead of smoothing QK by exploiting temporal stability.