

# Supplementary Materials: QVD: Post-training Quantization for Video Diffusion Models

Anonymous Authors

## A QUANTIZED INFERENCE OF HIDI-TQ

Here, we provide a detailed formulation and derivation of the quantized inference for the HiDi-TQ. As defined in Equation 1,

$$T_{emb}^z = clip\left(\left\lfloor -\log_2 \frac{|T_{emb} - \beta|}{s_T} \right\rfloor, 0, 2^b - 1\right), \quad (1)$$

where  $s_T$  is a scalar. We denote the quantized temporal feature as  $T_{emb}^z$ , thus the dequantized temporal feature can be defined as Equation 2:

$$\widehat{T}_{emb} = sign(T_{emb}) \cdot s_T \cdot 2^{-T_{emb}^z} + \beta, \quad (2)$$

then, the computation of the quantized time projection layer is depicted as the Equation 3:

$$\begin{aligned} W^T T_{emb} &\simeq \widehat{W}^T \widehat{T}_{emb} \\ &= (s_w^T \cdot W_q^T) \left[ sign(T_{emb}) \cdot s_T \cdot 2^{-T_{emb}^z} + \beta \right] \\ &= (s_w^T \cdot s_T) \cdot W_q^T \left[ sign(T_{emb}) \cdot 2^{-T_{emb}^z} \right] + (s_w^T \cdot \beta \cdot W_q^T) \\ &= s \cdot W_q^T \left[ sign(T_{emb}) \cdot 2^{-T_{emb}^z} \right] + (s_w^T \cdot \beta \cdot W_q^T) \\ &= s \cdot W_1 2^{-T_{emb}^z} + W_2 \\ &= s \cdot \text{BitShift}(W_1, -T_{emb}^z) + W_2. \end{aligned} \quad (3)$$

For brevity, we utilize symmetric weight quantization. The term  $W^T$  signifies the transposed weight matrix, while  $W_q^T$  represents its quantized counterpart. The symbols  $s_w^T$  and  $s_T$  correspond to the quantizer parameters for weights and activations, respectively. The expression  $W_2$  is defined as  $s_w^T \cdot \beta \cdot W_q^T$ . Ultimately, the matrix multiplication operation is approximated by BitShift and addition, thereby enhancing the speed of inference and reducing the demand for memory consumption.

The temporal feature  $T_{emb}$  is actually derived from the repetition of the same vector, hence, the  $sign(T_{emb})$  can be simplified to a single vector, which is then integrated into  $W_q^T$  as Equation 4:

$$\begin{aligned} W_1 &= \begin{bmatrix} w_{00} & w_{01} & \dots & w_{0w} \\ w_{10} & w_{11} & \dots & w_{1w} \\ \dots & & & \\ w_{h0} & w_{h1} & \dots & w_{hw} \end{bmatrix} \cdot sign(T_{emb}) \\ &= \begin{bmatrix} w_{00}s_0 & w_{01}s_1 & \dots & w_{0w}s_w \\ w_{10}s_0 & w_{11}s_1 & \dots & w_{1w}s_w \\ \dots & & & \\ w_{h0}s_0 & w_{h1}s_1 & \dots & w_{hw}s_w \end{bmatrix} \\ &= \begin{bmatrix} w'_{00} & w'_{01} & \dots & w'_{0w} \\ w'_{10} & w'_{11} & \dots & w'_{1w} \\ \dots & & & \\ w'_{h0} & w'_{h1} & \dots & w'_{hw} \end{bmatrix}. \end{aligned} \quad (4)$$

Table 1: Results of discarding varying amounts of outliers

Percentile	FID-VID	FVD
0.85	306.20	2791.01
0.9	306.57	2791.53
0.95	305.65	2782.77

Equation 5 defines the detailed calculation process of the BitShift operation,

$$\text{BitShift}(W_1, -T_{emb}^z)_{i,j} = \sum_{j=0}^{j=w} w'_{i,j} \gg T_{emb_{i,j}}^z. \quad (5)$$

## B MORE EXPLORATION ON TEMPORAL FEATURES

**Outliers of temporal features.** The temporal features exhibit pronounced skewness, with a handful of outliers exceeding the magnitude of regular values by several-fold. A naive approach might involve discarding these outliers to compress the activation range. Motivated by this rationale, we execute an experiment, the specifics of which are delineated in Table 1. The results reveal that truncating even a scant 5% of the largest data values induces a drastic degradation in model performance. In conjunction with prior experiments, this outcome suggests that despite the robustness of outliers to quantization noise, their retention is critical for the integrity of temporal features.

**Skewness in temporal features of different models.** To explore whether the extreme distribution of temporal characteristics is a common occurrence, we conduct further analysis on the currently popular video diffusion model, SVD [1]. As shown in Figure 1, we present histograms of additional temporal features. Besides AnimateDiff and MagicAnimate, extreme distributions of temporal features are similarly observed in SVD.

**Combination with TFMQ.** TFMQ [2] introduces the Finite set calibration (FSC) for temporal features, assigning quantization parameters to each temporal feature. We attempt to transfer this approach to the HiDi-TQ and conduct experiment on the TED-talks dataset with a setting of W8A8, the results shows that FID-VID decrease to 58.38 while FVD decrease to 420.79. Figure 2 indicates that using FSC for quantization improves the TDScore and compromises the distinctiveness of the temporal features. To analyze the reasons, we plot the temporal features obtained from the two quantization methods. As shown in Figure 3, employing the FSC method in the HiDi-TQ results in the quantized temporal features covering fewer quantization levels.

**The potential of HiDi-TQ.** To explore the potential of HiDi-TQ, we investigate the performance of the model on the TED-talks dataset with varying bits. As shown in Table 3, HiDi-TQ still exhibits no significant performance degradation even at a 5-bit setting.

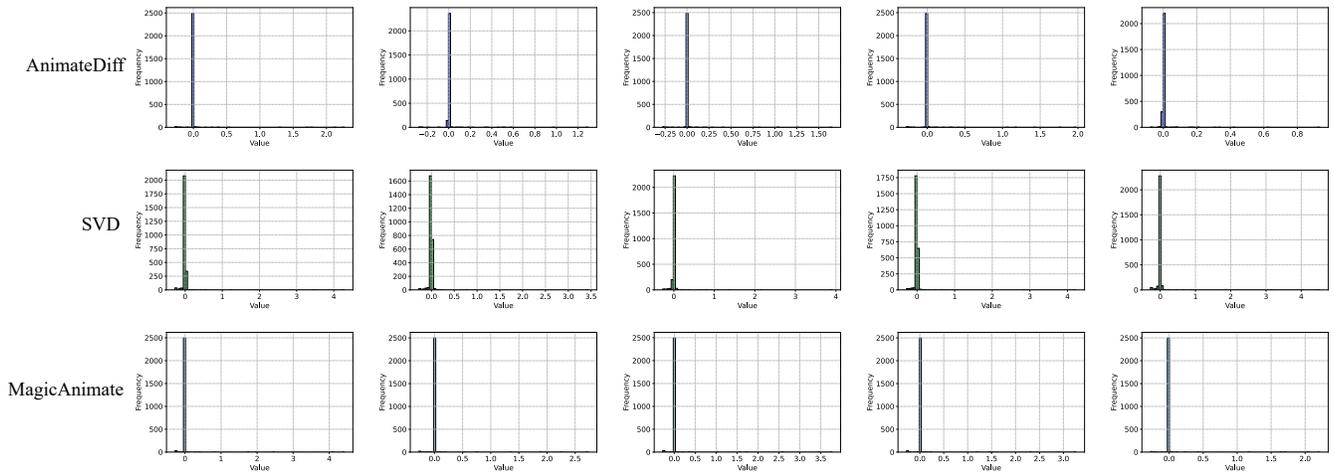


Figure 1: Histograms of AnimateDiff, SVD and MagicAnimate.

Table 2: Comparison between the original prompts and the enhanced prompts.

Original Prompts	Enhanced Prompts
some elephants and one is by some water	Some elephants, with one drinking by the water
A zebra all by itself in the green forest	A zebra walking alone in the green forest
A group of people who are standing together	A group of people standing together, talking and moving around
there are two large boats that are in the water	Two large boats sailing in the water

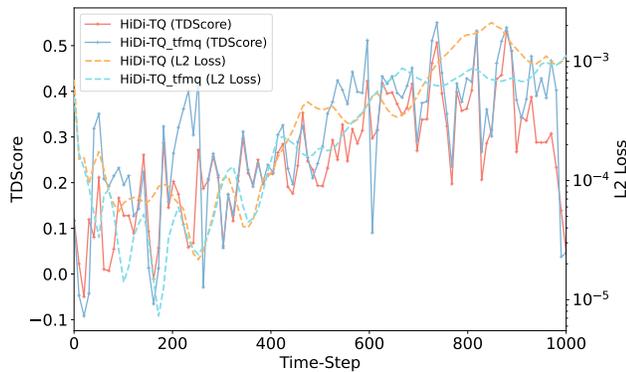


Figure 2: TDScore and L2 Loss of HiDi-TQ and HiDi-TQ with FSC.

Table 3: Results of discarding varying amounts of outliers.

bits	FID-VID	FVD
5	51.06	387.57
6	50.93	384.52
7	50.72	384.56
8	49.38	385.77

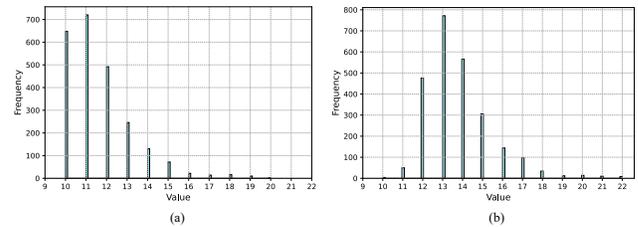


Figure 3: (a) and (b) show the histogram of the quantized temporal feature using HiDi-TQ with FSC and without FSC respectively.

## C MORE IMPLEMENTATION DETAILS

COCOCaption and FSCOCO are image-text pair datasets where the text describes the contents of the images but lacks descriptions of potential actions within the images. To better utilize these datasets for testing video generation models, we employ large language models to modify the original text descriptions. Specifically, we use ChatGPT-3.5 and the following prompt to transform the original textual descriptions: **"Please add several motion elements to the following prompt, keeping the new prompt within 77 tokens. Note the following: 1. Do not introduce elements not present in the original prompt. 2. The subject of the motion must be restricted to elements already in the prompt or logically added background elements. 3. The additions must adhere to real-world logic and physical laws."** Table 2 shows

233 the comparison between the original prompts and the enhanced  
234 prompts.

## 235 REFERENCES

237 [1] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej  
238 Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al.

2023. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127* (2023). 291  
292  
[2] Yushi Huang, Ruihao Gong, Jing Liu, Tianlong Chen, and Xianglong Liu. 2023. Tfmq-dm: Temporal feature maintenance quantization for diffusion models. *arXiv preprint arXiv:2311.16503* (2023). 293  
294

239 295  
240 296  
241 297  
242 298  
243 299  
244 300  
245 301  
246 302  
247 303  
248 304  
249 305  
250 306  
251 307  
252 308  
253 309  
254 310  
255 311  
256 312  
257 313  
258 314  
259 315  
260 316  
261 317  
262 318  
263 319  
264 320  
265 321  
266 322  
267 323  
268 324  
269 325  
270 326  
271 327  
272 328  
273 329  
274 330  
275 331  
276 332  
277 333  
278 334  
279 335  
280 336  
281 337  
282 338  
283 339  
284 340  
285 341  
286 342  
287 343  
288 344  
289 345  
290 346  
347  
348