

# WF-VAE: Enhancing Video VAE by Wavelet-Driven Energy Flow for Latent Video Diffusion Model

## Supplementary Material

The supplementary materials include further details as follows:

- We present additional notations in Sec. 1.
- We analyze subband energy and entropy of wavelet transform in Sec. 2, which further validates our motivation.
- We present our training parameters in Sec. 3
- We present the derivation of the Causal Cache formulation in Sec. 4.
- We provide additional experimental results in Sec. 5.

### 1. Notations

The notations and their descriptions in the paper are shown in Tab. 1.

Notations	Descriptions
$WT(\cdot)$	Wavelet transform
$IWT(\cdot)$	Inverse wavelet transform
$S_{\square\square\square}^{(l)}$	Wavelet subband within layer $l$ , where $\square\square\square$ specifies the type of filtering (high or low pass) applied in three dimensions.
$\mathcal{W}^{(l)}$	The set of all subbands within layer $l$

Table 1. Notations symbols and their descriptions.

### 2. Wavelet Subband Analysis

We analyze the energy and entropy distributions across the subbands obtained after wavelet transform. As illustrated in Fig. 1b, the energy and entropy of the video are primarily concentrated in the  $hhh$  low-frequency subband. This concentration suggests that low-frequency components carry more significant information and necessitate lower compression rates to ensure superior reconstruction performance. This observation further validates the rationale behind our proposed approach.

### 3. Training Details

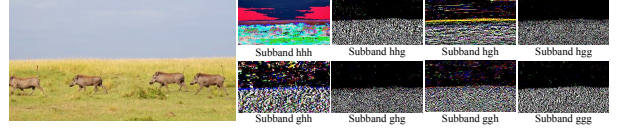
The training hyperparameters are shown in Tab. 2.

### 4. Derivation of Causal Cache

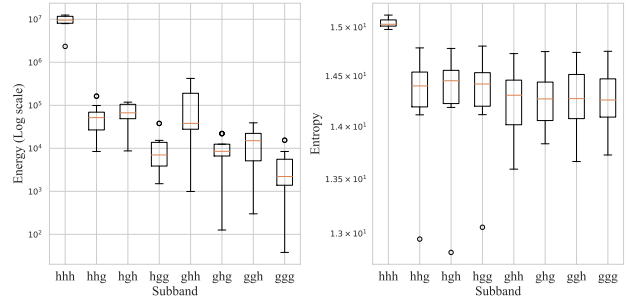
Let us define a convolution with sliding window index  $n \in \mathbb{N}_0$  and chunk index  $m \in \mathbb{N}_0$ . Given a convolutional stride  $s$  and kernel size  $k$ , as shown in Fig. 3, the starting and ending frame indices for each sliding window are:

$$t_{window,start}(n) = ns, \quad (1)$$

$$t_{window,end}(n) = t_{window,start}(n) + k - 1. \quad (2)$$



(a) Visualization of the eight subbands obtained after wavelet transform of the video.



(b) Energy and entropy of each subband.

Figure 1. Visualization of the subbands and their respective energy and entropy.

Parameter	Setting
<i>Stage I - 800k step</i>	
Learning Rate	1e-5
Total Batch Size	8
Peceptual(LPIPS) Weight	1.0
WL Loss Weight ( $\lambda_{WL}$ )	0.1
KL Weight ( $\lambda_{KL}$ )	1e-6
Resolution	256×256
Num Frames	25
EMA Decay	0.999
<i>Stage II - 200k step</i>	
Num Frames	49
<i>Stage III - 200k step</i>	
Peceptual(LPIPS) Weight	0.1

Table 2. Training hyperparameters across three stages.

For chunk boundaries, we define:

$$t_{chunk,end}(m) = k - 1 + mT_{chunk} \quad (3)$$

where  $T_{chunk}$  denotes the chunk size. For a given chunk index  $m$ , the maximum sliding index  $n_{max}(m)$  is determined by the constraint  $t_{window,end}(n) \geq t_{chunk,end}(m)$ :

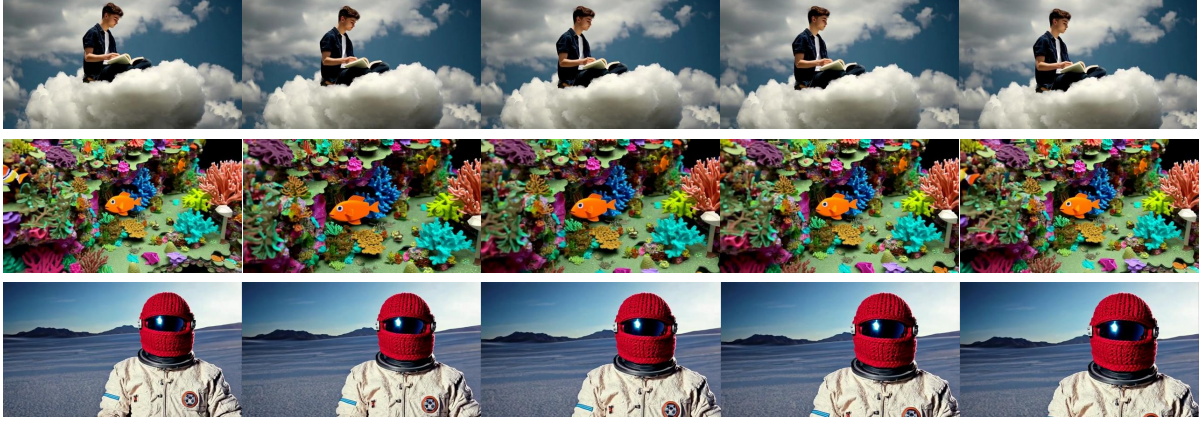


Figure 2. **Qualitative experiments in video generation model pretraining.** It demonstrates that WF-VAE can be effectively applied to the training of downstream diffusion models.

Method	480P				720P			
	PSNR↑	LPIPS↓	rFVD↓	SSIM↑	PSNR↑	LPIPS↓	rFVD↓	SSIM↑
<i>4 latent channels</i>								
WF-VAE-L	<b>30.56</b>	<b>0.0595</b>	<b>55.65</b>	<b>0.8713</b>	<b>31.12</b>	<b>0.0617</b>	<b>49.93</b>	<b>0.8799</b>
Allegro	30.06	0.0689	105.70	0.8673	30.78	0.0668	86.85	0.8795
<i>16 latent channels</i>								
WF-VAE-L	<b>34.28</b>	<b>0.0275</b>	<b>20.43</b>	<b>0.9347</b>	<b>34.82</b>	<b>0.0294</b>	<b>19.27</b>	<b>0.9384</b>
CogVideoX	33.85	0.0317	32.85	0.9319	34.24	0.0331	24.82	0.9364

Table 3. Quantitative evaluation on Inter4K dataset, using **65 frames**.

$$n_{\max}(m) = \left\lfloor \frac{mT_{\text{chunk}}}{s} + 1 \right\rfloor. \quad (4)$$

Consequently, the required cache size  $T_{\text{cache}}(m)$  for chunk  $m$  is:

$$T_{\text{cache}}(m) = t_{\text{chunk},\text{end}}(m) - t_{\text{window},\text{start}}(n_{\max}(m)) + 1 = mT_{\text{chunk}} + k - \left\lfloor \frac{mT_{\text{chunk}}}{s} + 1 \right\rfloor s \quad (5)$$

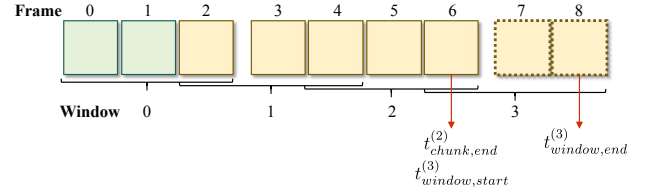


Figure 3. Illustration of *Causal Cache* with parameters  $k=3$ ,  $s=2$ , and chunk size  $T_{\text{chunk}}=4$ .

## 5. Additional Experiments

**Evaluation Across Different Resolutions.** To validate the robustness of WF-VAE across different resolutions, we conduct metric evaluations on the Inter4K dataset at 480P and 720P resolutions. As shown in Tab. 3, WF-VAE demonstrates competitive performance in reconstruction tasks across varying resolutions.

**Validation in Diffusion Model Pretraining.** To verify the applicability of WF-VAE to LVDM, we select Open-Sora Plan for pretraining on large-scale datasets with a resolution

of  $512 \times 288$  pixels. Qualitative experiments, as illustrated in 2, demonstrate promising generative performance.

**More Qualitative Evaluations.** To further demonstrate the capability of our model in achieving state-of-the-art reconstruction performance with low computational cost, we conduct additional qualitative evaluations against the representative VAE, CogVideoX. Refer to Fig. 4 and supplementary material for more video examples.



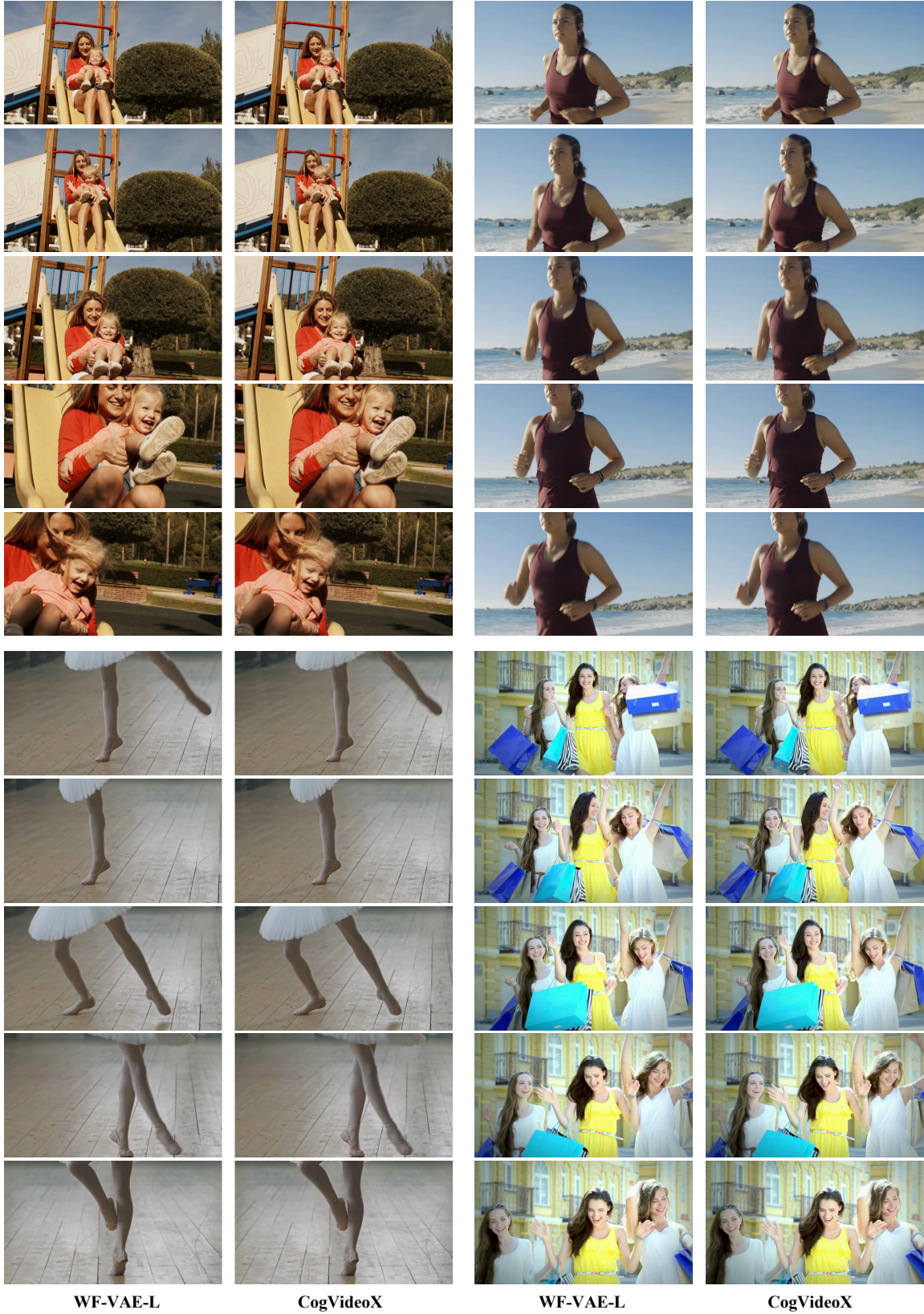


Figure 4. **Qualitative experiments in high motion videos.** We include more 480P comparison videos in the supplementary materials.