Supplementary Material for:

Pyramidal Spectrum: Frequency-based Hierarchically Vector Quantized VAE for Videos

Tushar Prakash[1†], Onkar Susladkar[2†], Inderjit S. Dhillon[3], Sparsh Mittal[4]
[1]Independent Researcher, India [2]UIUC [3]UT Austin and Google Research, USA [4]IIT Roorkee, India
(tushar121prakash, onkarsus13)@gmail.com, inderjit@utexas.edu, sparsh.mittal@mfs.iitr.ac.in
[†]Co-First Authors

1. Implementation Details

Reconstruction quality is assessed using standard metrics, including Structural Similarity Index Measure (SSIM) [6], Peak Signal-to-Noise Ratio (PSNR) [1], and Learned Perceptual Image Patch Similarity (LPIPS) [9].

1.1. Spatio–Temporal Resolution Sampling

- **Spatial scales.** We uniformly sample one of ten resolutions[1] at every step:

  256×256 512×512 768×512 1024×512
  1024×1024 1024×2048 2048×2048
  1280×720 (720p) 1920×1080 (1080p) 3840×2160 (4K)

  Uniform sampling prevents the codebook from collapsing to square crops.

- **Temporal scales.** Clips are truncated (or zero-padded) to $\{16, 32, 48, 96\}$ consecutive frames.

1.2. Mixed-Precision Optimisation

We employ PyTorch AMP with `bfloat16` for activations/weights and `float32` master gradients; all loss reductions occur in `float32` to avoid numerical drift.

1.3. Optimiser and LR Schedule

- **Adam**: β_1=0.9, β_2=0.95, ϵ=1×10^{-8}.
- Initial LR: 1×10^{-5}; 10k warm-up steps, then linear decay to 0 over the remaining 790k steps.
- **EMA**: exponential moving average of all parameters with decay 0.999; EMA weights are used for validation.

1.4. Batching and Gradient Accumulation

Each of the 8×8=64 A100 80 GB GPUs processes a local batch of 4 clips; gradients are accumulated for 4 forward passes before an optimiser step. Hence, EffectiveBatch $= 64 \times 4 \times 4 = 1024$.

We train for 800k optimiser steps, totalling $\sim 8.2\times10^8$ clip presentations.

1.5. Distributed Training Stack

- **Hardware**: 8 HGX-A100 nodes (dual AMD EPYC 7763 + 8×A100 80GB), NVSwitch + InfiniBand NDR.
- **Software**: DeepSpeed 0.14.0, ZeRO-3 with 5 GiB reduce buckets, full parameter/optimizer sharding, and activation checkpointing.

1.6. Reproducibility

- RNG seeds (`PyTorch`, CUDA, `NumPy`) fixed at 42.
- Gradients clipped to L2-norm 1.0.
- We will release the DeepSpeed config, data-sampling script, and commit hash with the camera-ready code.

2. Block Diagram

2.1. PVQ-VAE Architecture

Figure 3 illustrates the internal architecture of the proposed PVQ-VAE. The encoder consists of a sequence of five convolutional blocks, each progressively downsampling the input video volume along the temporal and spatial dimensions. Each block is followed by a pyramidal vector quantization module, which discretizes the latent features at the bottleneck stage. The input of size $T \times H \times W \times 3$ is mapped through intermediate representations to a compact

[1] $W\times H$ in pixels.

Table 1. Ablation Study Based on Quantization Techniques

VQ Technique / Codebook Size / Dim	COCO-Val			WebVid-Val		
	PSNR(↑)	SSIM(↑)	LPIPS(↓)	PSNR(↑)	SSIM(↑)	LPIPS(↓)
VQ / 4096 / 256	31.45	0.825	0.093	32.91	0.838	0.092
GVQ / 4096 / 256	32.25	0.836	0.089	33.34	0.842	0.089
LFQ / 32800 / 16	34.22	0.842	0.084	33.92	0.855	0.085
RVQ / 8000 / 512	33.92	0.849	0.078	34.22	0.865	0.079
PSVQ / 8000 / 512	34.45	0.855	0.073	34.98	0.871	0.076
RVQ + LFQ / 32800 / 16	34.78	0.869	0.076	35.27	0.879	0.074
PVQ (Default) / 32800 / 16	**35.12**	**0.877**	**0.067**	**36.01**	**0.881**	**0.072**

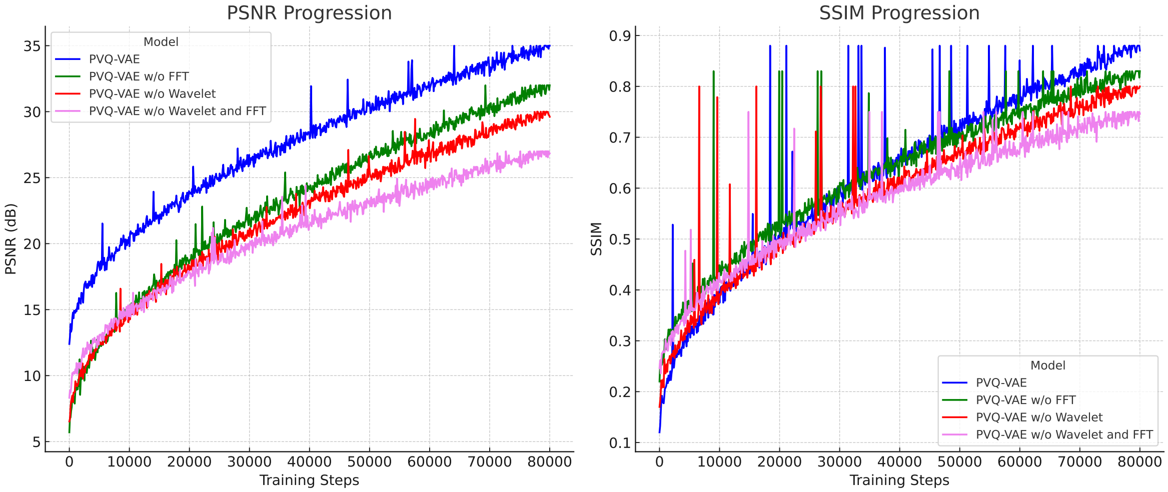

Figure 1. PSNR and SSIM progression during training for different ablations of the PVQ-VAE model.

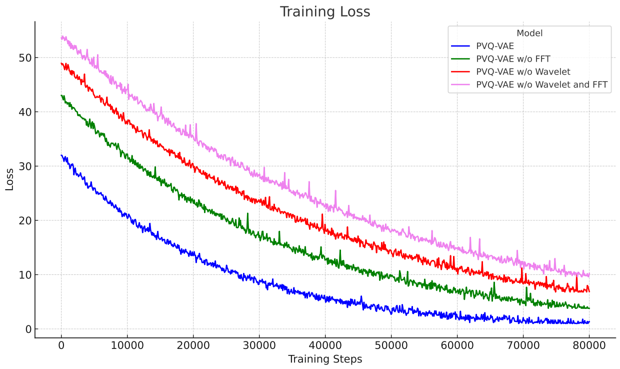

Figure 2. Training loss convergence for PVQ-VAE and its ablated variants.

latent tensor of shape $T/16 \times H/32 \times W/32 \times d$. This quantized representation is then upsampled through three decoder blocks, aided by projection layers to match the original resolution and channel dimensions. The architecture ensures temporal consistency and multi-scale feature learning via hierarchical compression.

2.2. Discriminator Architecture

Figure 4 presents the internal architecture of the spatiotemporal discriminator used to evaluate the realism of generated video samples. It comprises a stack of 3D convolutional layers with SiLU activations and BatchNorm3D, interleaved with max-pooling layers for hierarchical temporal-spatial feature extraction. Each convolutional stage increases the representational capacity while reducing the resolution, enabling effective discrimination across both motion and appearance. The architecture ends in a compact latent embedding, which is processed by a linear head (not shown in the block) to produce the final real/fake classification score.

3. Additional Experimental Results

In this section we will provide the extra ablation study and additional visual results.

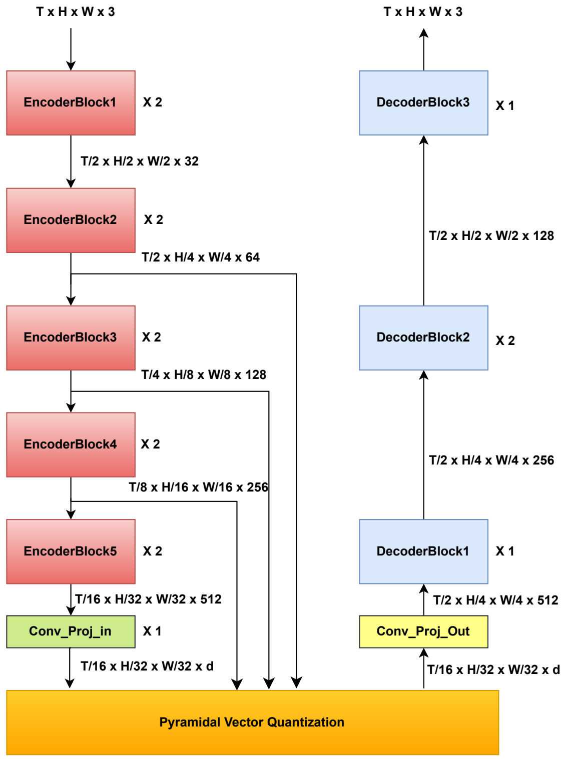

Figure 3. Internal block diagram of the PVQ-VAE architecture.

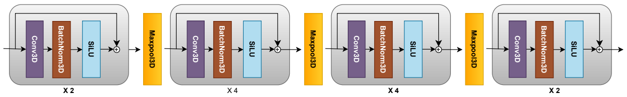

Figure 4. Block diagram of the 3D CNN-based discriminator.

3.1. Effect of various VQ-Techniques

Table 1 presents an ablation comparing various quantization strategies. Standard VQ [5] and GVQ [2] underperform due to limited codebook capacity, restricting their ability to capture diverse patterns. LFQ [8] improves token compactness through lookup-free binarization but lacks multi-scale adaptability. RVQ [3] introduces residual refinement, yet struggles to retain fine-grained details in high-frequency regions.

Replacing LFQ with standard VQ in our PVQ framework (PSVQ) degrades performance, as the smaller codebook limits expressivity. In contrast, PVQ achieves superior results by leveraging hierarchical residual modeling and pyramidal token fusion. This design enables efficient code allocation across scales, combining coarse global structure with fine detail preservation. Our default PVQ configuration achieves the best scores, demonstrating that multi-resolution quantization with binarized, lookup-free tokens support both compression efficiency and rich representational fidelity.

3.2. Effect of FFT and DWT on Training

The experimental results, as illustrated in Fig. 1 and Fig. 2, demonstrate the efficacy of the proposed PVQ-VAE model and highlight the importance of incorporating both FFT and wavelet (DWT) components. As shown in Fig. 1, PVQ-VAE consistently outperforms its ablated variants in terms of PSNR and SSIM throughout the training process, indicating superior reconstruction quality. Moreover, Fig. 2 shows that the full model exhibits faster and more stable loss convergence, reflecting improved training dynamics. The ablation study further reveals a significant drop in both reconstruction performance and convergence speed when either the FFT or wavelet modules are removed, with the greatest decline observed when both are omitted. These findings validate the complementary benefits of combining frequency-domain (FFT) and multi-resolution (wavelet) representations, establishing PVQ-VAE as a robust and effective framework for high-fidelity image modeling.

3.3. Visual Results on the Effect of Frequency (FFT/DWT) Branches

Fig.5–12 present qualitative ablation results demonstrating the role of the frequency-aware branches—namely Fast Fourier Transform (FFT) and Discrete Wavelet Transform (DWT)—in PVQ-VAE. These components are designed to capture complementary frequency information: FFT encodes global structural patterns, while DWT preserves fine-grained local details.

When both branches are removed $(w/o\ FFT + w/o\ DWT)$, the model fails to capture essential spatiotemporal features, resulting in severe blurring across all frames. The reconstructed outputs lose both global coherence and local texture, leading to perceptually degraded results with missing motion boundaries and faded scene layouts. This highlights the critical role of frequency-aware encoding in preserving motion dynamics and scene fidelity.

Removing only the FFT branch $(w/o\ FFT)$ leads to the degradation of large-scale spatial structures. The absence of global frequency encoding causes inconsistencies in the scene's geometry and layout, especially in dynamic scenarios where maintaining object position and structure is critical. This is evident in cases such as the moving car in Fig.6 sequence or aerial views in Fig. 10, where the horizon lines and object outlines become distorted or unstable.

Conversely, removing only the Wavelet branch $(w/o\ DWT)$ suppresses high-frequency components, resulting in the loss of fine details and texture sharpness. For instance, the fur textures in the dog sequence in Fig. 9 or the foliage in the coastline sequence in Fig. 12 appear overly smoothed, and object boundaries become less distinct. While the overall scene structure remains, the perceptual quality is diminished due to missing details.

In contrast, PVQ-VAE with both FFT and Wavelet branches preserves both global coherence and local details across diverse scenarios, including complex motions, lighting variations, and dynamic environments. These results confirm the complementary nature of frequency-aware processing: FFT ensures structural consistency at large scales, while Wavelet encoding sharpens textures and captures local variations crucial for high-fidelity video reconstruction.

3.4. Reconstruction Quality Comparison of PVQ-VAE Against State-of-the-Art Video VAEs

Fig. 13-18 present a comprehensive frame reconstruction comparison between our proposed PVQ-VAE and state-of-the-art video VAEs, namely 3D-MBQ-VAE [4], CV-VAE [10], and CogVideoX VAE [7]. Across diverse scenes including underwater motion, fast-action flying, high-speed driving, urban occlusions, and complex sci-fi environments, PVQ-VAE consistently reconstructs sharper and more structurally coherent frames. This improvement stems from both architectural innovations and training strategies designed to overcome the limitations of existing methods.

First, PVQ-VAE integrates frequency-aware branches (FFT and DWT) into its encoder, allowing it to preserve both global structural information and fine-grained local textures. Whereas, models like CV-VAE and 3D-MBQ-VAE lack this spectral decomposition, leading to blurred textures and loss of high-frequency details. For instance, in the underwater sequence Fig. 13, the coral textures and diver contours remain crisp in PVQ-VAE, while they appear smoothed and washed out in CV-VAE and 3D-MBQ-VAE reconstructions.

Second, the use of Pyramidal Vector Quantization (PVQ) allows PVQ-VAE to hierarchically fuse multiscale discrete tokens, achieving finer detail recovery without inflating latent size. In contrast, 3D-MBQ-VAE employs a single-level quantization, which cannot balance coarse motion consistency with texture fidelity. This leads to frame degradation under motion, as seen in the fast-motion flying sequence Fig. 14, where PVQ-VAE maintains the bird structures with minimal artifacts, unlike the competing models that produce ghosting or blending effects.

Third, PVQ-VAE avoids the blurring artifacts associated with continuous latent spaces, which affect models like CV-VAE that prioritize diffusion model compatibility over reconstruction sharpness. This is particularly evident in the motorcycle sequences (Fig. 15 and 18), where PVQ-VAE reconstructs crisp road markings and object boundaries, whereas CV-VAE outputs lack fine spatial detail, resulting in texture loss and temporal smearing.

Moreover, while CogVideoX VAE employs a 3D causal latent model, its focus on text-video alignment over reconstruction fidelity leads to degraded visual quality under complex motion. For example, in the urban traffic sequence (Fig. 16) and dynamic sci-fi scene (Fig. 17), PVQ-VAE successfully reconstructs occluded objects, intricate mechanical structures, and background textures, whereas CogVideoX outputs exhibit visible distortions and loss of structural coherence.

Finally, PVQ-VAE leverages cross-modal alignment with a high-resolution 2D VAE (FLUX) to further enhance perceptual quality. This contrastive supervision encourages the model to retain image-level sharpness in the video domain. By aligning frame-wise representations with high-fidelity image priors, the model learns to reconstruct sharper textures and cleaner motion boundaries. As a result, PVQ-VAE maintains both temporal consistency and high-frequency detail, even in challenging scenarios.

Overall, these results demonstrate that PVQ-VAE achieves a superior rate-distortion trade-off, producing compact latents while preserving rich visual details across diverse motion sequences, as consistently validated by Fig. 13-18.

References

[1] Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th international conference on pattern recognition*, pages 2366–2369. IEEE, 2010.

[2] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

[3] Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11523–11532, 2022.

[4] Onkar Susladkar, Jishu Sen Gupta, Chirag Sehgal, Sparsh Mittal, and Rekha Singhal. Motionaura: Generating high-quality and motion consistent videos using discrete diffusion. *arXiv preprint arXiv:2410.07659*, 2024.

[5] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

[6] 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.

[7] 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.

[8] Lijun Yu, José Lezama, Nitesh B Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Vighnesh Birodkar, Agrim Gupta, Xiuye Gu, et al. Language model beats diffusion–tokenizer is key to visual generation. *arXiv preprint arXiv:2310.05737*, 2023.

[9] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of

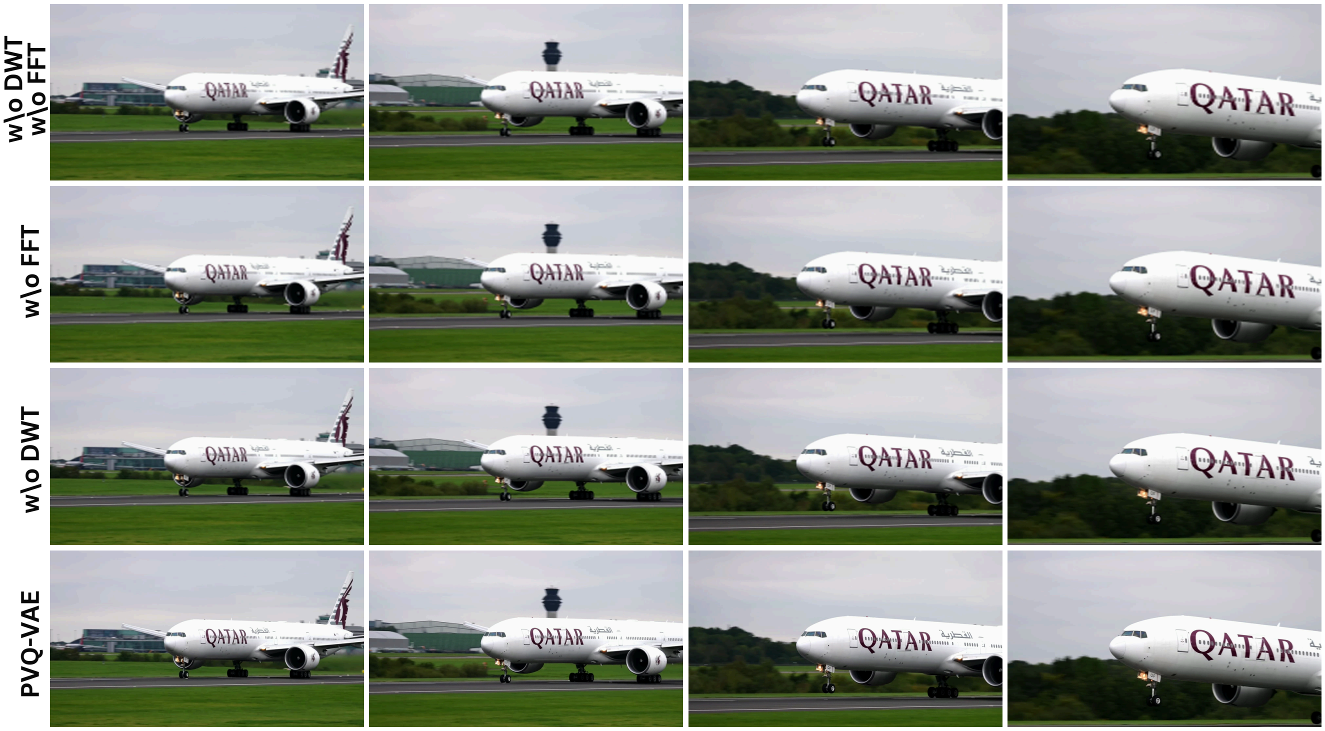

Figure 5. Ablation on a moving aircraft sequence. See Video Comparison

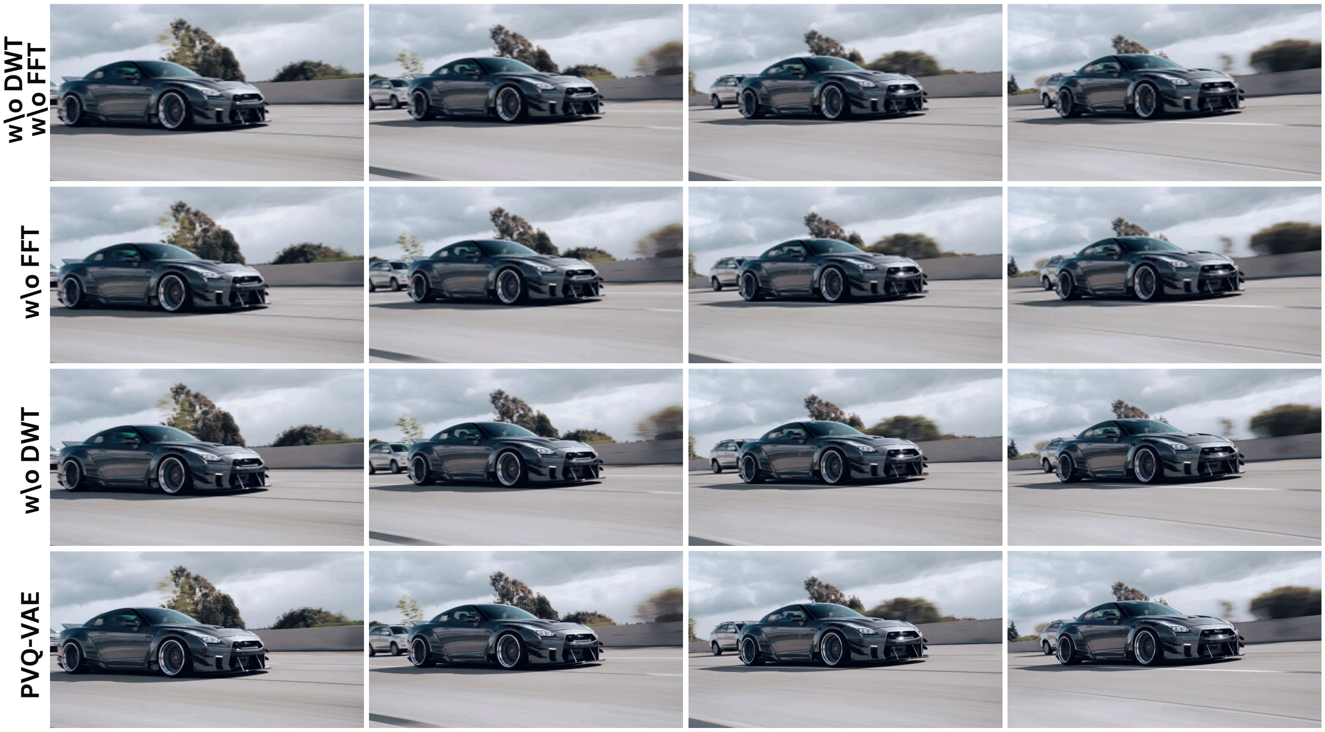

Figure 6. Ablation of frequency-aware branches on a car motion sequence.See Video Comparison

deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recogni-*

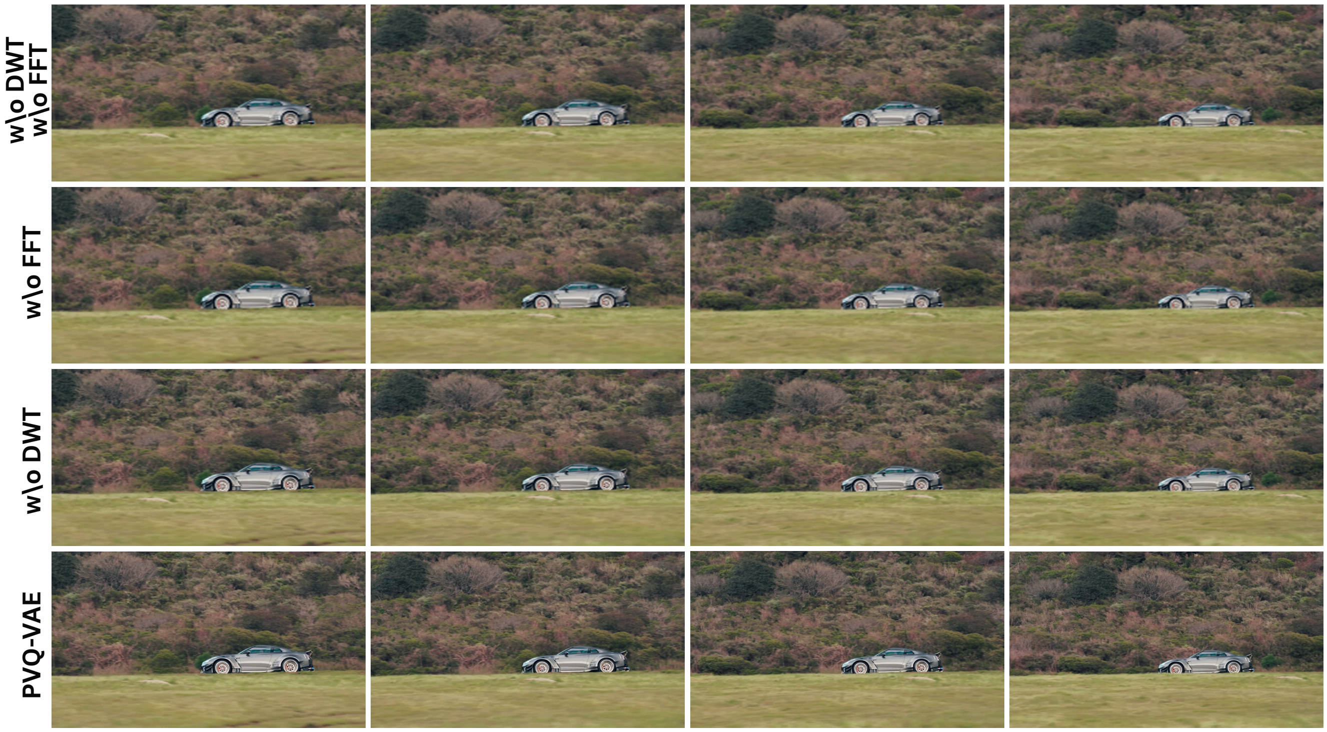

Figure 7. Ablation of frequency-aware branches on a car motion sequence in natural scenery.See Video Comparison

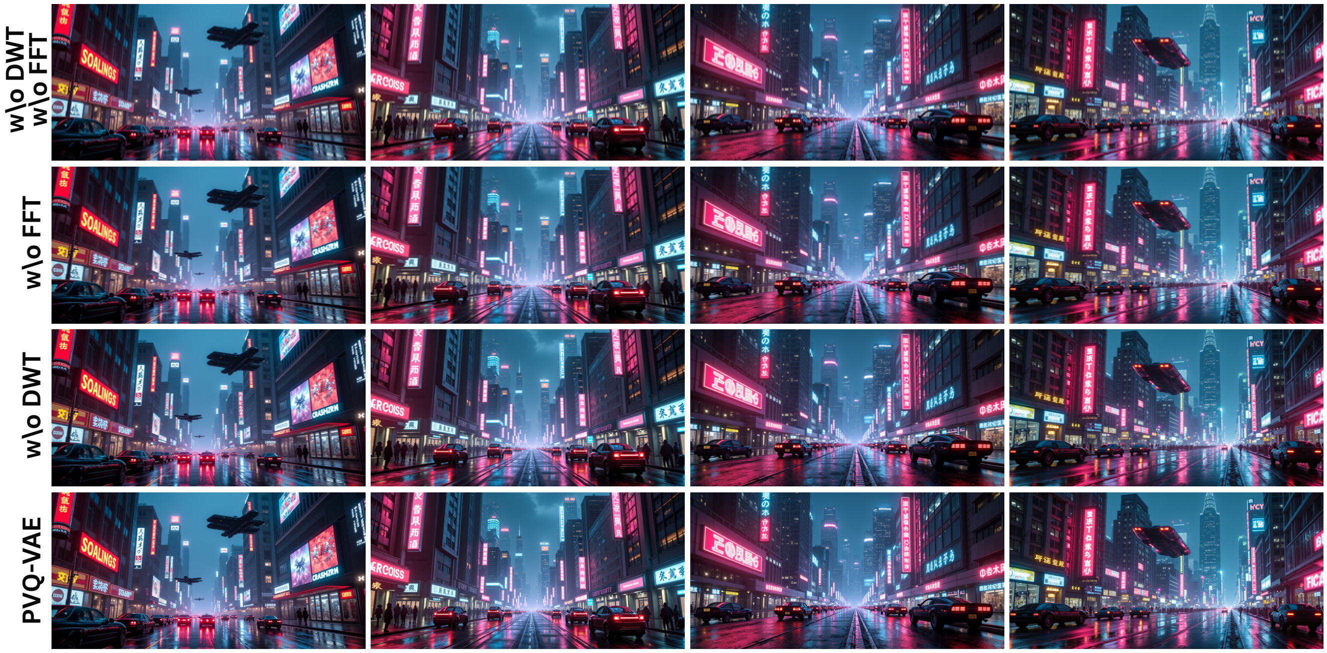

Figure 8. Ablation of frequency-aware branches on a night city sequence. See Video Comparison

tion, pages 586–595, 2018.

[10] Sijie Zhao, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Muyao Niu, Xiaoyu Li, Wenbo Hu, and Ying Shan. Cv-vae: A compatible video vae for latent generative video models. *Advances in Neural Information Processing Systems*, 37:12847–12871, 2024.

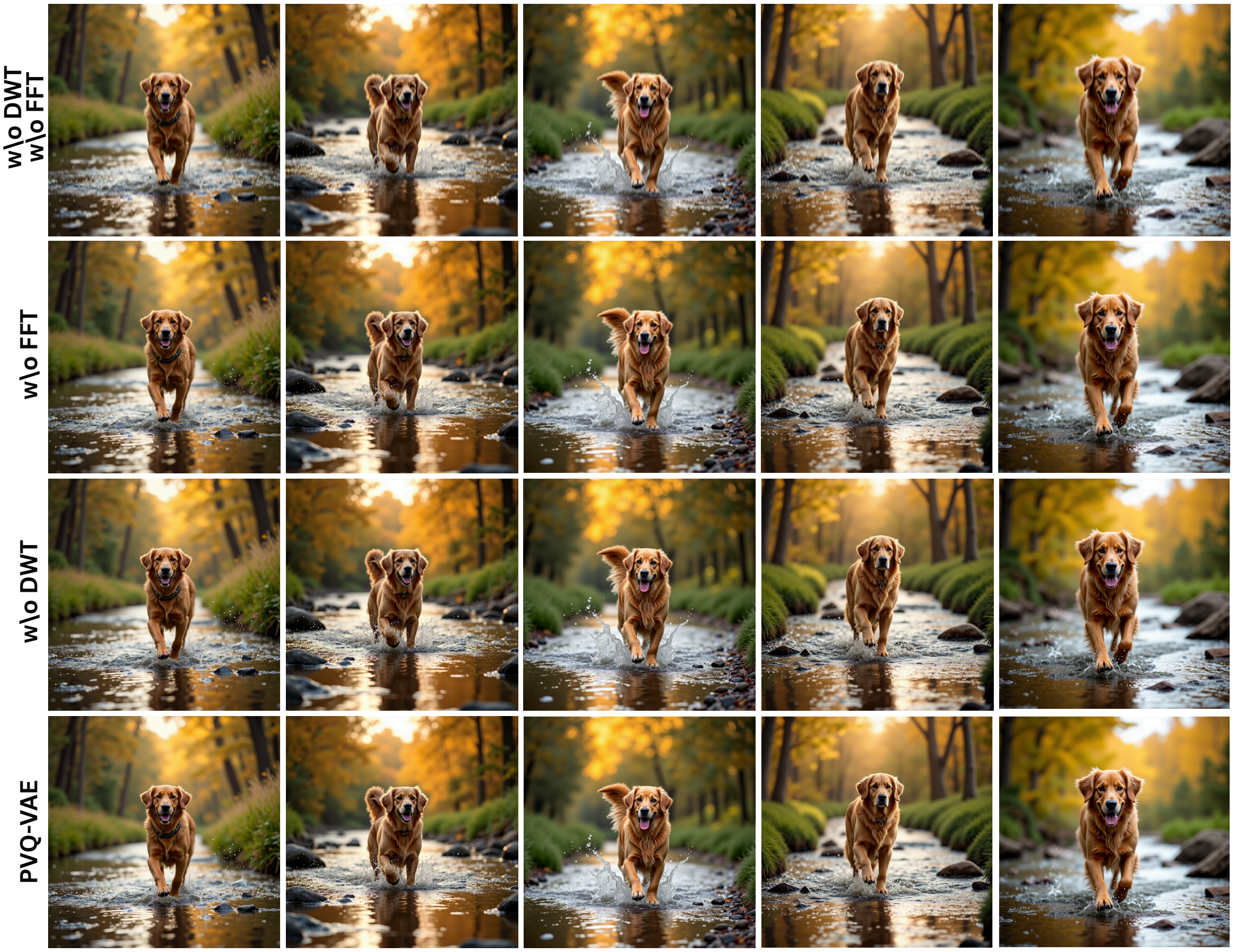

Figure 9. Frequency-aware branch ablation on a dynamic dog sequence.See Video Comparison

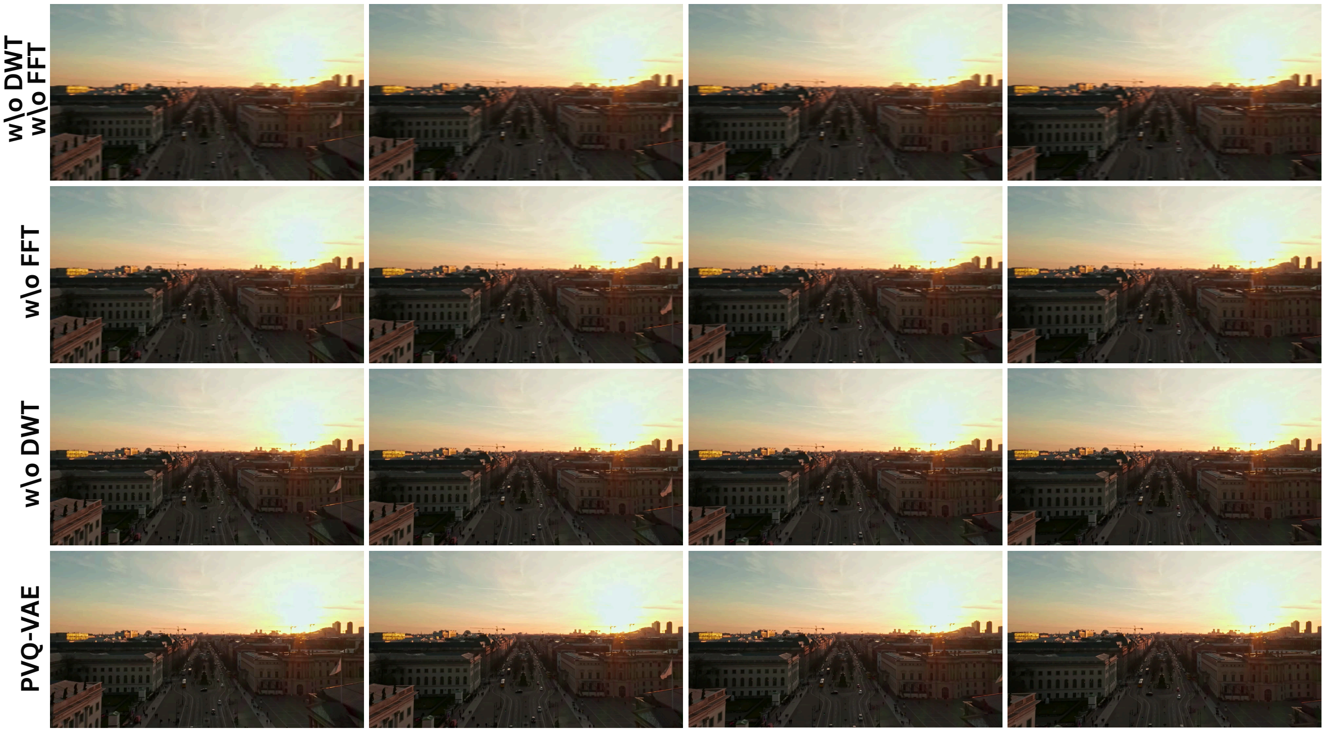

Figure 10. Frequency-aware branch ablation on a sunset cityscape sequence. See Video Comparison

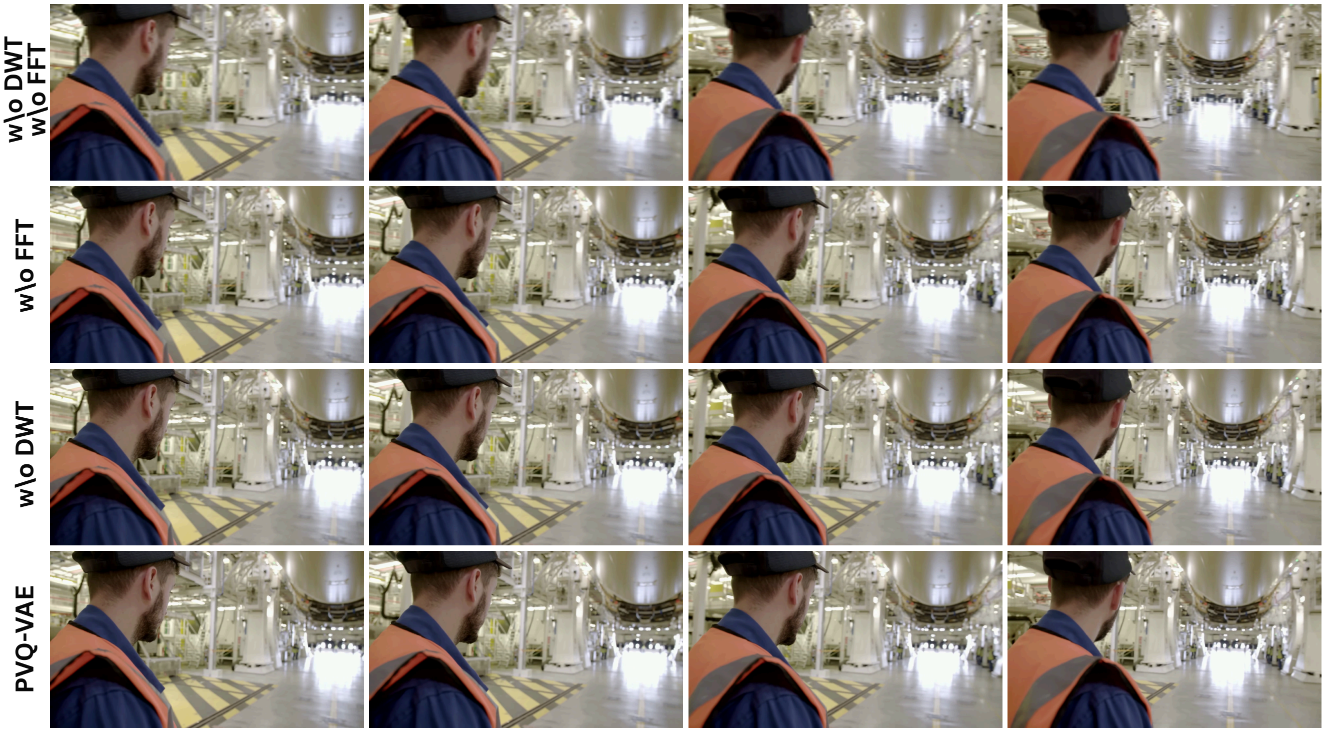

Figure 11. Frequency-aware branch ablation on an industrial walkthrough sequence.See Video Comparison

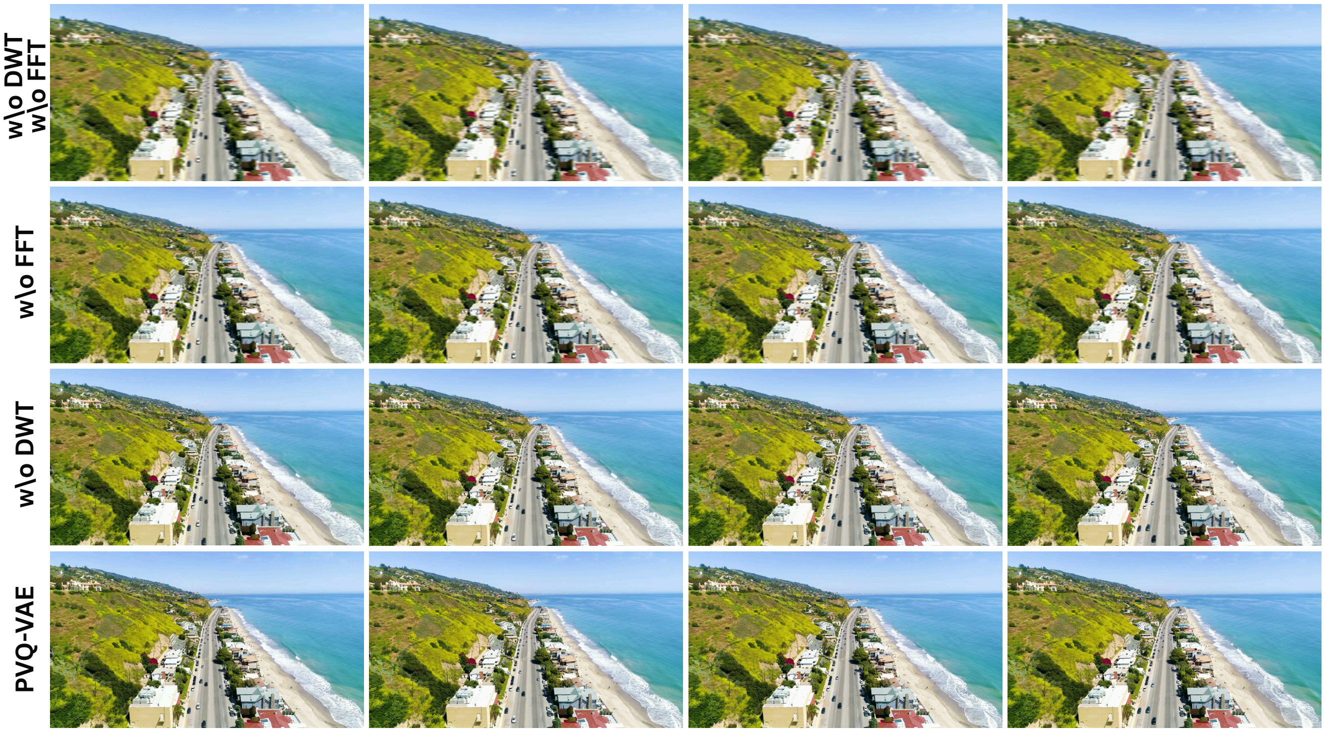

Figure 12. Frequency-aware branch ablation on an aerial coastline sequence.See Video Comparison

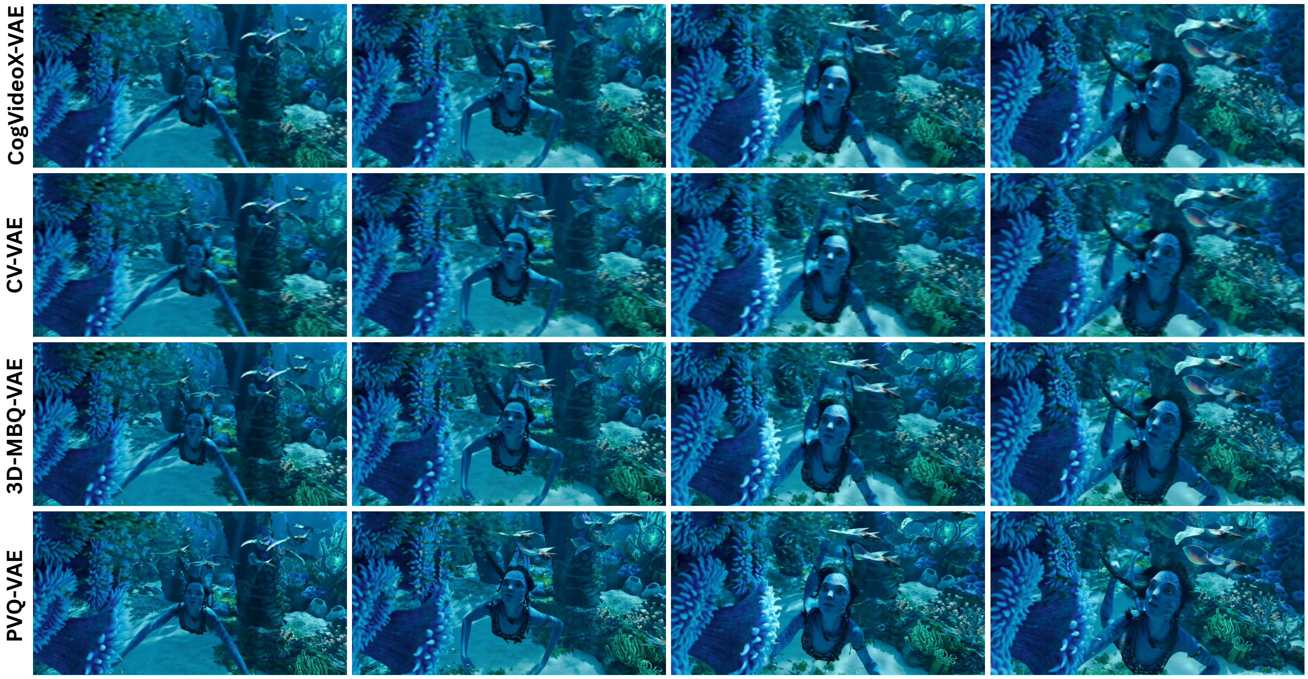

Figure 13. Qualitative comparison of frame reconstruction against state-of-the-art methods on an underwater sequence.See Video Comparison

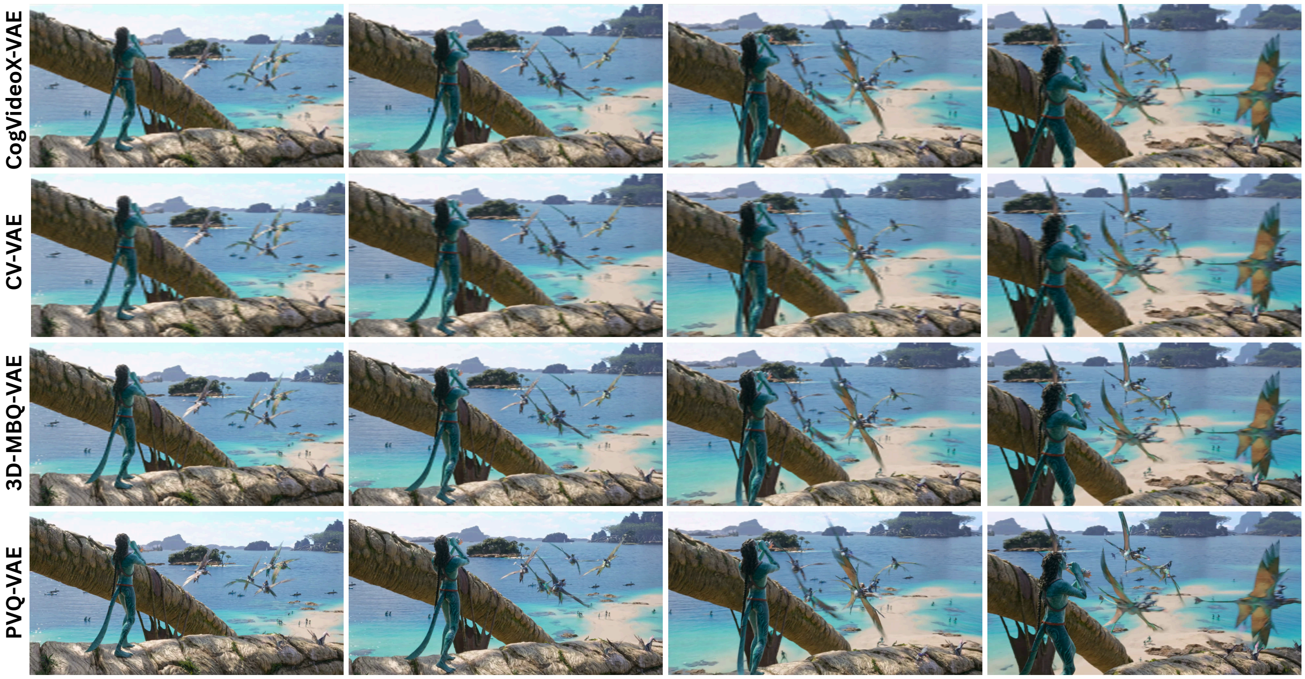

Figure 14. Qualitative comparison of frame reconstruction on a fast-motion flying sequence.See Video Comparison

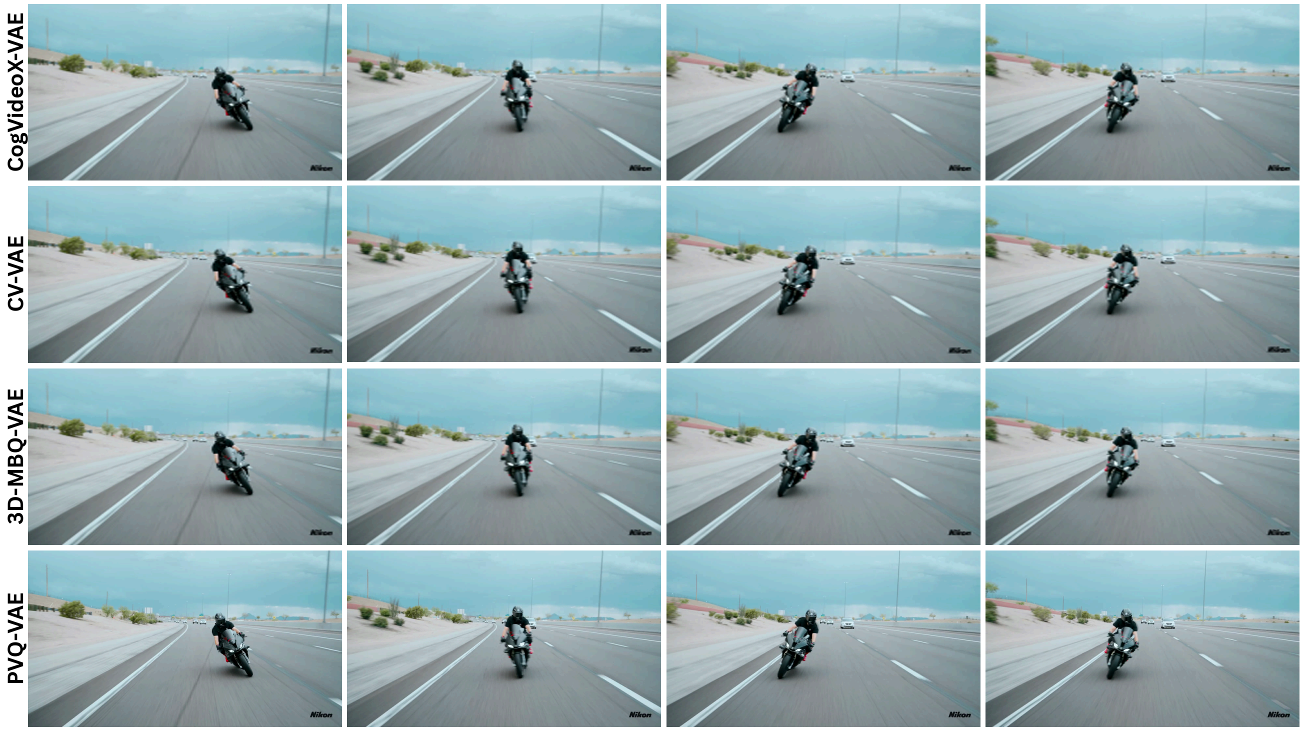

Figure 15. Qualitative comparison of frame reconstruction on a high-speed motorcycle motion sequence.See Video Comparison

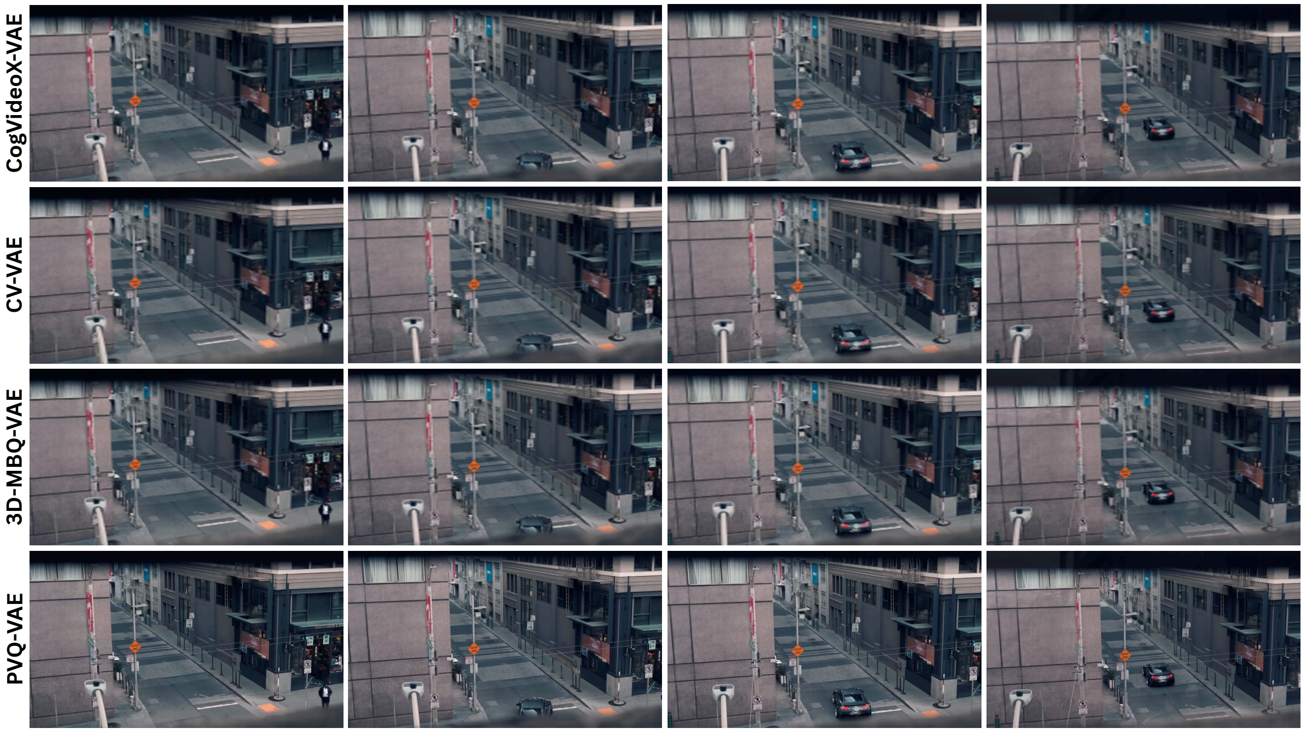

Figure 16. Qualitative comparison of frame reconstruction on an urban traffic sequence with complex motion and occlusions.See Video Comparison

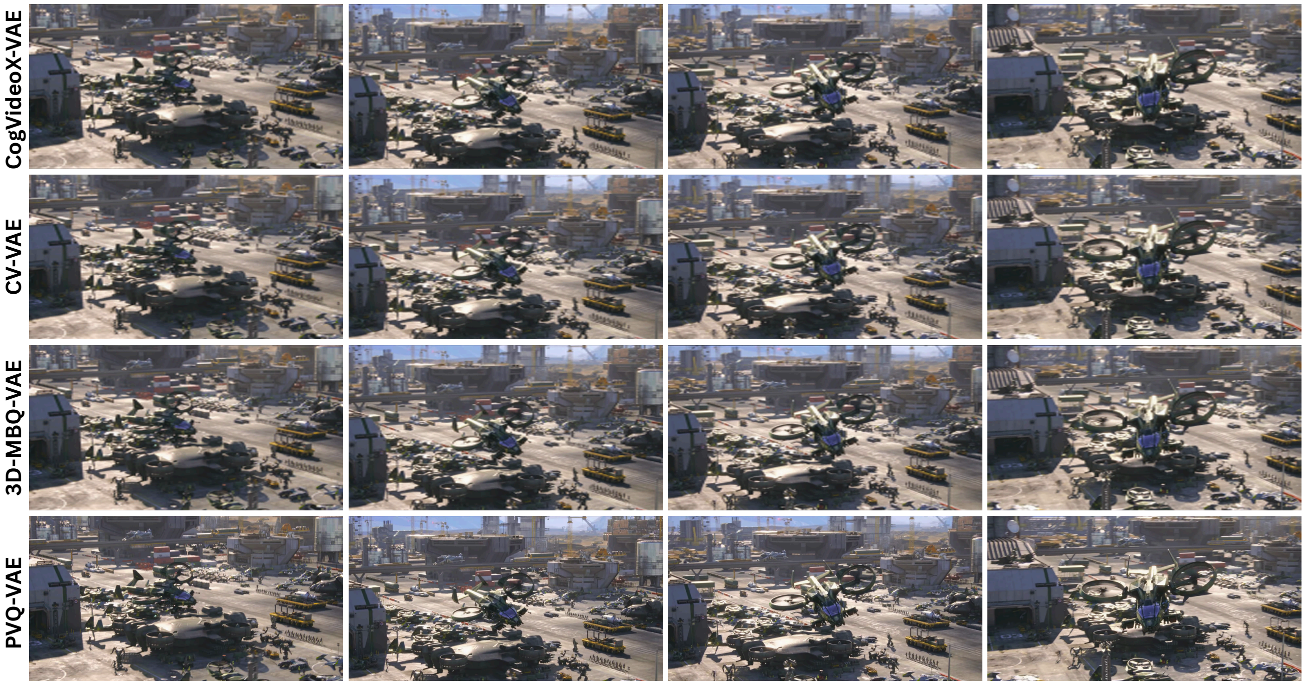

Figure 17. Qualitative comparison of frame reconstruction on a dynamic sci-fi scene with complex object motion and fine structural details.See Video Comparison

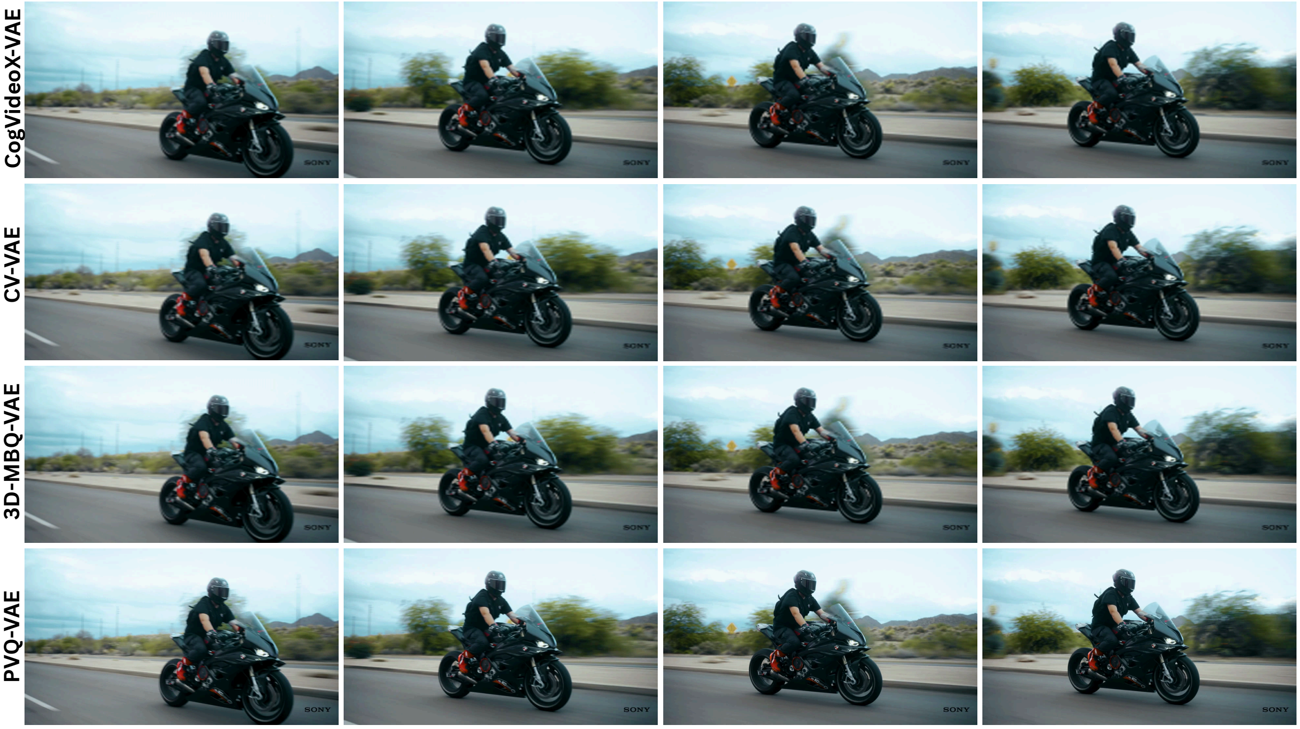

Figure 18. Qualitative comparison of frame reconstruction on a high-speed motorcycle riding sequence.See Video Comparison