

Video Generation with Predictive Latents

Yian Zhao^{1,2}, Feng Wang^{1,†}, Qiushan Guo¹, Chang Liu³, Jian Zhang², Jie Chen²

¹ByteDance Seed, ²Peking University, ³Tsinghua University

[†]Project lead

Abstract

Video Variational Autoencoder (VAE) enables latent video generative modeling by mapping the visual world into compact spatiotemporal latent spaces, improving training efficiency and stability. While existing video VAEs achieve commendable reconstruction quality, continued optimization of reconstruction does not necessarily translate into improved generative performance. *How to enhance the diffusability of video latents remains a critical and unresolved challenge.* In this work, inspired by principles of predictive world modeling, we investigate the potential of predictive learning to improve the video generative modeling. To this end, we introduce a simple and effective predictive reconstruction objective that unifies predictive learning with video reconstruction. Specifically, we randomly discard future frames and encode only partial past observations, while training the decoder to reconstruct the observed frames and predict future ones simultaneously. This design encourages the latent space to encode temporally predictive structures and build a more coherent understanding of video dynamics, thereby improving generation quality. Our model, termed Predictive Video VAE (**PV-VAE**), achieves superior performance on video generation, with **52% faster** convergence and a **34.42 FVD improvement** over the Wan2.2 VAE on UCF101. Furthermore, comprehensive analyses demonstrate that PV-VAE not only exhibits favorable scalability, with generative performance improving alongside VAE training, but also yields consistent gains in downstream video understanding, underscoring a latent space that effectively captures temporal coherence and motion priors.

Date: March 7, 2026

Correspondence: wangfeng.eve@bytedance.com

1 Introduction

Video generation has achieved extraordinary breakthroughs [16, 22, 49, 57, 67], with contemporary models producing content of cinematic brilliance that often surpasses professional-grade cinematography and production standards. This rapid progress stems from the ability to represent the visual world within compact latent spaces, largely driven by advances in Latent Video Diffusion Models (LVDMs) [5] and Video Variational Autoencoders (VAEs) [32]. LVDMs operate not on raw pixels, but on the compact spatiotemporal latent spaces created by video VAEs. These latents not only reduce computational overhead, but more importantly, they provide a structured space for video generative modeling, making video VAEs one of the key components of video generation systems.

The common practice for developing video is to extend well-trained image VAEs and continue training them on video corpora. Modern video VAE [22, 49] typically adopt CNN-based architectures. They are first trained as 2D image VAEs on large-scale image datasets, after which the 2D convolutions are inflated into 3D

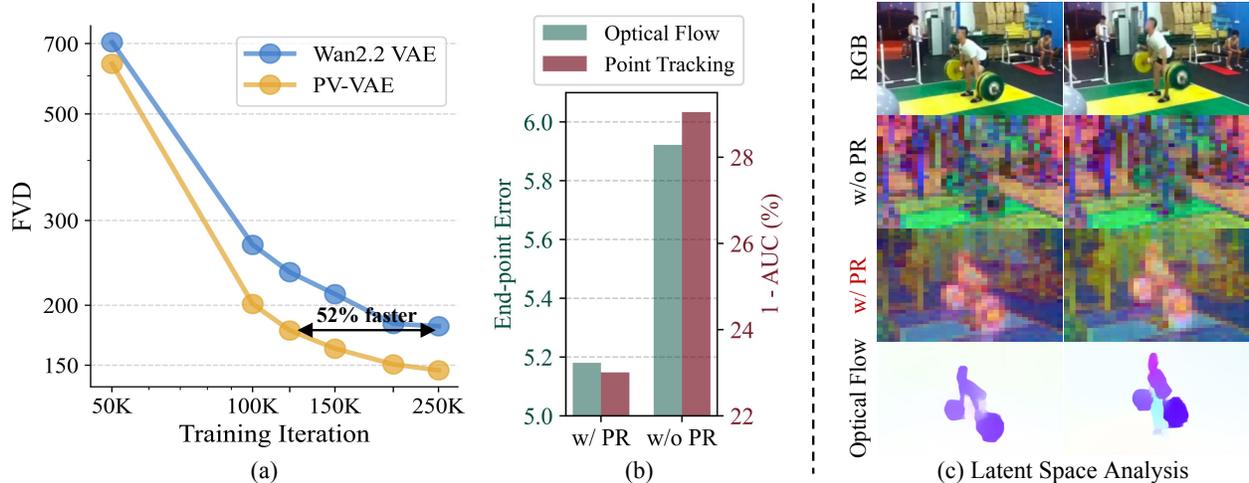


Figure 1 Our PV-VAE achieves 52% faster convergence and 34.42 FVD gain over Wan2.2 VAE on UCF101. Optical flow and point tracking probing tasks show that the Predictive Reconstruction (PR) objective enhances the spatiotemporal understanding of latent space. Latent visualizations further reveal that PV-VAE captures clear motion-aware structures aligned with video dynamics (visualized via optical flow).

causal convolutions to inherit the spatial compression capability [10], followed by video training to achieve joint spatiotemporal compression. While existing video VAEs achieve commendable reconstruction quality, continued optimization of reconstruction does not necessarily translate into improved generative performance. How to enhance the diffusability [38] of video latents remains a critical and unresolved challenge.

Different from images, video modeling requires capturing spatiotemporal representations that describe both the visual content and the underlying temporal dynamics from discrete frame sequences. These representations are essential for generating motion-consistent and temporally coherent videos. Recent studies [47, 70] have shown that the representations learned by video generative models yield meaningful results on various video understanding tasks (*e.g.*, depth estimation, tracking, and segmentation), underscoring the crucial role of well-structured video representations in achieving high-quality video generation. These findings raise a natural question: *what kind of latent spaces enable video generative models to learn temporally structured representations more effectively?* Inspired by the principle of predictive world modeling [23], which frames future-state prediction as a powerful means of acquiring temporal and causal structures of videos, we investigate how predictive learning can improve the generative modeling of latent spaces in video VAEs.

Specifically, we introduce a predictive reconstruction objective that unifies video reconstruction with predictive learning. At each step, we randomly discard future frames, enabling the encoder to observe only partial temporal context, while requiring the decoder to reconstruct the complete video sequence. This design forces the model to jointly capture fine-grained visual details and long-term video dynamics, thereby enriching the latent space with robust motion priors that substantially bolster video generation. Notably, our approach seamlessly integrates into existing video VAE pipelines without altering the original loss composition or introducing additional hyperparameters. Additionally, to prevent “copy-shortcut” from dominating the optimization, a motion-aware objective is incorporated as a targeted constraint, directing the model’s attention toward structural motion and fostering more effective predictive learning.

To validate the effectiveness of our approach, we evaluate both class-conditional and unconditional video generation, and show that our model, termed Predictive Video VAE (**PV-VAE**), consistently achieves notable improvements. For instance, our PV-VAE achieves 52% faster convergence and 34.42 FVD improvement over Wan2.2 VAE [49] on UCF101 [39] (*cf.* figure 1(a)). To further understand the source of these gains, we examine the learned latent spaces through the lens of diffusion features, which have been shown to serve as reliable intermediate indicators of generative capability [40, 61]. Surprisingly, we find that the diffusion features learned with our PV-VAE exhibit stronger performance across several downstream video understanding tasks,

including optical flow estimation [15], next-frame prediction [69], and point tracking [13] (*cf.* figure 1(b)). PCA visualizations of the latent space further reveal that PV-VAE captures motion-aware structures that align well with the underlying video dynamics (*cf.* figure 1(c)). These observations indicate that our method strengthens the temporal understanding and motion sensitivity of the learned latent space, leading to improved video generation quality.

In summary, our main contributions are as follows:

- We investigate the diffusability of video latent spaces and propose a predictive reconstruction objective. By integrating predictive learning into the VAE framework, our method enriches the latent space with robust temporal priors and motion awareness.
- We develop Predictive Video VAE, which achieves significant improvements across both class-conditional and unconditional video generation, validating the efficacy of our approach.
- We provide a comprehensive diagnostic of the latent spaces, establishing a clear link between predictive accuracy and generative quality, showing the data scalability of PV-VAE, and demonstrating consistent gains across multiple downstream video understanding tasks.

2 Related Work

Video VAE. Video VAE [32] serves as a fundamental component in modern video generative pipelines. By employing an encoder–decoder architecture, it maps high-dimensional data into a compact latent space, thereby enhancing the training efficiency and stability of generative models [33]. Early video generative models [5, 29] directly reused image VAEs to spatially compress individual frames or inserted 1D temporal convolutions into image VAEs to mitigate inter-frame flickering. Sora [6] first proposed a video compression network for joint spatiotemporal compression to reduce the inference cost. However, training a video VAE from scratch remains computationally expensive and inefficient. To leverage pretrained image VAEs while enabling temporal compression, the community has explored various hybrid designs. Open-Sora [67] employs a cascade VAE to separately perform spatial and temporal compression. CV-VAE [65] introduces latent space alignment between video VAE and image VAE. OD-VAE [10] inflates 2D convolutions of image VAEs into 3D causal convolutions. CogVideoX’s VAE [57] adopts parallel algorithms for long video processing, while IV-VAE [52] introduces additional channels for temporal compression. For improved efficiency, Lite-VAE [35] and WF-VAE [25] utilize wavelet-based methods, whereas LeanVAE [11] and H3AE [53] prioritize structural lightweighting and decoding acceleration. Additionally, some works [50, 60, 62] decouple motion dynamics from static content to bolster temporal modeling and reduce redundancy. Recently, many advanced video generative models [16, 22, 42, 49] have developed unified image-video VAEs. Despite these advances, little attention has been paid to how the latent spaces can be structured to explicitly benefit video generation. In this work, we take a step toward addressing this challenge by introducing a predictive reconstruction objective.

Diffusability of latent space. Diffusability refers to the suitability of a latent space for the diffusion process. Incorporating structured constraints into the latent space has emerged as a promising approach to improve this. In the image domain, many frameworks [24, 59, 64, 66] internalize semantic priors from pre-trained encoders (*e.g.*, DINOv2 [30]), while VTP [58] advocates for a joint representation-reconstruction learning paradigm. Conversely, video-level exploration remains hampered by architectural and computational bottlenecks. SSSVAE [26] relies on hand-crafted heuristic constraints to shape the latent manifold. In contrast, our proposed predictive reconstruction encourages the latent space to autonomously capture structured temporal dynamics.

Predictive learning. Predictive learning, which aims to predict future states by modeling existing information, has demonstrated powerful representation learning and modeling capabilities across diverse tasks. Its applications span from sequence, action, and trajectory prediction [34, 48] to masked language/visual modeling (MLM/MVM) [7, 12, 19, 54]. SiameseMAE [18] combines predictive learning with masked modeling to learn fine-grained correspondences from randomly sampled video frames. JEPA (Joint Embedding Predictive Architecture) [23] further proposes that predictive latent learning serves as a fundamental pathway toward understanding the visual world and constructing world models. Subsequent works [1–4] have demonstrated

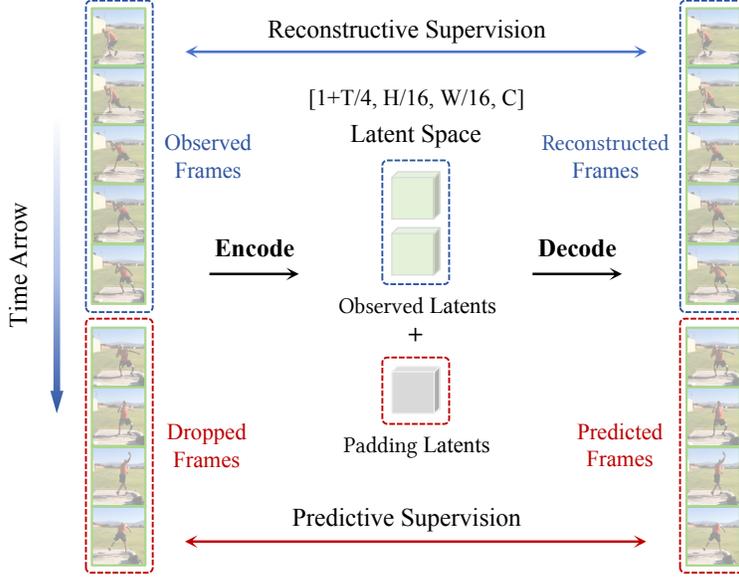


Figure 2 Overall pipeline of the proposed PV-VAE. PV-VAE randomly discards future frames and encodes only observed ones. The padded latents are then decoded to reconstruct the full video, enabling the model to learn visual reconstruction and temporal understanding jointly from reconstructive and predictive supervision.

powerful capabilities in visual understanding, prediction, and planning under predictive learning objectives, further validating the effectiveness of this paradigm. Most recently, Cambrian-S [56] posits predictive sensing as a promising direction for next-generation intelligent agents, offering a proof-of-concept via next-latent-frame prediction. Building upon these insights, our approach integrates predictive learning with video reconstruction, enabling the model to simultaneously reconstruct visual details and predict future states. This design enhances the temporal dynamics and motion understanding of latent spaces, thereby facilitating more effective video generative modeling.

3 Approach

Our goal is to enhance the diffusability of the latent spaces by jointly learning predictive and reconstruction objectives. Let $\mathbf{x} \in \mathbb{R}^{(1+T) \times H \times W \times 3}$ denotes a video clip with $1+T$ frames in pixel space, and $\mathbf{z} \in \mathbb{R}^{(1+t) \times h \times w \times c}$ denotes the sampled video latents. Here, $p_s = H/h = W/w$ and $p_t = T/t$ are the spatial and temporal compression ratios, and c denotes the latent channel. The initial extra frame serves to ensure a unified processing pipeline for image ($T=0$) and video data, following common practice [49, 57].

3.1 Framework

Integrating predictive learning into reconstruction. To incorporate predictive learning, we reformulate the VAE training procedure by introducing a partial-to-complete reconstruction task. Specifically, we divide the video clip into two parts along the time dimension, denoted as $\mathbf{x} = \langle \mathbf{x}_{obs}, \mathbf{x}_{drop} \rangle$. The model is trained to reconstruct the entire clip \mathbf{x} conditioned on the observed portion \mathbf{x}_{obs} . At each training step, we first partition the video clip into $G = 1 + T/p_t$ groups based on the temporal compression ratio p_t , where the first group consists of the first frame, and each subsequent group includes p_t frames. We then sample the number of dropped groups, $k \sim U\{0, \dots, \lfloor (G-1) \cdot r \rfloor\}$, where r is a predefined maximum dropping ratio. The retained preceding frames $\mathbf{x}_{obs} \in \mathbb{R}^{(1+T-k \cdot p_t) \times H \times W \times 3}$ are fed into the encoder to obtain the corresponding observed latent $\mathbf{z}_{obs} \in \mathbb{R}^{(G-k) \times h \times w \times c}$. Given that the decoder shares symmetric spatiotemporal scaling factors with the encoder, it requires a full-length latent sequence to reconstruct the entire video sequence. As a result, we pad \mathbf{z}_{obs} by temporally concatenating it with padding vectors $\mathbf{z}_{pad} \in \mathbb{R}^{k \times h \times w \times c}$, which are sampled from an uninformative prior (*i.e.*, containing no input information). This complete latent sequence is passed through the decoder to reconstruct the entire video \mathbf{x} . Since the dropped frames \mathbf{x}_{drop} are entirely withheld from the

encoder, the model is compelled to infer the subsequent video evolution from the past observations \mathbf{x}_{obs} and encode this predictive information into its latent spaces. The overall pipeline of our method is illustrated in [figure 2](#). Under this learning objective, the model not only learns to reconstruct fine-grained visual details but also develops a deeper understanding of temporal dynamics and motion awareness in videos, thereby improving the latent representations to facilitate better generative modeling.

Model design. We implement PV-VAE with 3D causal convolutions, employing $16\times$ spatial and $4\times$ temporal downsampling, with a latent channel dimension of 64. For the encoder, we first perform two stages of spatiotemporal downsampling, reducing both the temporal and spatial dimensions by a factor of 4. Then, while keeping the temporal length fixed, we apply two additional spatial downsampling operations, resulting in an overall $16\times$ spatial reduction. The decoder is symmetric to the encoder, first conducting two stages of spatial upsampling followed by two stages of spatiotemporal upsampling.

3.2 Implementation

Training. PV-VAE is first pretrained on multi-resolution image data for 300K steps at resolutions of 256×256 , 384×384 , and 512×512 . Following this pretraining, it is further trained for 50K steps on video data at 256×256 and 512×512 resolutions using the proposed predictive reconstruction objective. During training, each process randomly samples a varying number of images or videos based on the resolution to maintain a balanced computational load across processes. Since the decoder requires reconstructing videos from complete video latents during inference, a training–inference gap arises. To address this issue, we introduce an additional *decoder fine-tuning stage*. Specifically, we freeze the encoder, disable the random frame-dropping operation, and train the decoder for another 50K steps to perform standard video reconstruction. This stage substantially improves reconstruction quality and provides a stronger foundation for high-fidelity video generation.

Loss functions. We adopt a combination of losses commonly used in video VAEs [49, 57], including a mean squared error (MSE) loss, a learned perceptual image patch similarity (LPIPS) loss [63], an adversarial (GAN) loss [17], and a KL regularization term. The GAN loss is activated from step 5,000 during training and remains enabled throughout the entire decoder fine-tuning stage. To prevent the “copy-shortcut” of non-motion regions from dominating the optimization, we incorporate an additional *motion-aware objective*. Specifically, the model is required to reconstruct not only the raw pixels but also the temporal differences between adjacent frames. This design effectively filters out static backgrounds and compels the video VAE to prioritize the learning of structural motion and temporal evolution. The total loss is formulated as follows:

$$\mathcal{L}_{total} = \lambda_{rec}(\mathcal{L}_{MSE} + \mathcal{L}_{Diff}) + \lambda_{lpiPs}\mathcal{L}_{LPIPS} + \lambda_{gan}\mathcal{L}_{GAN} + \lambda_{kl}\mathcal{L}_{KL}, \quad (1)$$

where each λ controls the relative contribution of its corresponding component.

4 Experiments

4.1 Experimental setups

Evaluation details. We evaluate PV-VAE on three widely used benchmarks: UCF101 [39], RealEstate10K [68], and Kinetics-400 [21]. For video generation, we follow prior work [10, 52] and adopt the Latte architecture [29], a Transformer-based latent diffusion model that supports both unconditional and class-conditional generation. We use UCF101 for class-conditional generation and RealEstate10K for unconditional generation. All videos are converted into 17-frame clips at 256×256 resolution for both training and testing. For video reconstruction, we randomly sample 2,048 videos from Kinetics-400, which offers better visual quality and higher resolution than UCF-101, making it better suited for assessing reconstruction fidelity. We take the first 17 frames of each video and evaluate the model at 256×256 and 512×512 resolutions to assess its ability to reconstruct inputs across different spatial scales, which is crucial for video generation.

To assess generation quality, we report Frchet Video Distance (FVD) and Kernel Video Distance (KVD) [46]. For UCF101, we additionally report the Inception Score (IS) [36] computed using the pre-trained C3D model from [45], following the evaluation protocol of [10]. All metrics are computed over 2048 generated samples. To assess reconstruction quality, we report reconstruction FVD (rFVD), Peak Signal-to-Noise Ratio (PSNR) [20], Learned Perceptual Image Patch Similarity (LPIPS) [63], and Structural Similarity Index Measure (SSIM) [51].

Table 1 Comparison of generation performance on the UCF101 and RealEstate10K datasets at 17-frame 256×256 resolution. The best and second-best are indicated in **bold** and underlined. The notation $tTsScC$ denotes a temporal downsampling factor of T , a spatial downsampling factor of $S \times S$, and a latent channel dimension of C .

| Method | Latent config | UCF101 | | | RealEstate10K | | TSpeed (it/s) | TMem (GiB) | Param (M) |
|-------------------------|---------------|---------------|--------------|--------------|---------------|-------------|------------------|---------------|--------------|
| | | FVD ↓ | KVD ↓ | IS ↑ | FVD ↓ | KVD ↓ | | | |
| CogX-VAE [57] | t4s8c16 | 176.90 | 16.47 | 64.19 | 94.12 | 10.41 | 0.76 | 85.93 | 216 |
| IV-VAE [52] | t4s8c16 | 175.74 | 22.32 | 64.51 | 92.37 | <u>8.35</u> | 1.28 | 88.34 | 242 |
| WF-VAE-L [25] | t4s8c16 | 188.19 | 33.01 | <u>67.49</u> | 107.26 | 12.56 | 2.52 | 87.36 | 317 |
| Hunyuan-VAE [22] | t4s8c16 | 210.30 | 52.81 | 66.40 | 83.45 | 13.23 | 1.64 | 87.36 | 246 |
| Wan2.1 VAE [49] | t4s8c16 | <u>167.10</u> | 11.54 | 66.04 | 83.84 | 10.64 | 1.88 | 86.44 | 127 |
| Wan2.2 VAE [49] | t4s16c48 | 180.79 | 17.80 | 67.32 | 87.15 | 10.11 | 4.96 | 30.90 | 705 |
| SSVAE [26] | t4s16c48 | 168.68 | 19.71 | 66.39 | <u>79.08</u> | 8.79 | 3.92 | 34.00 | 315 |
| PV-VAE | t4s16c64 | 146.37 | <u>14.52</u> | 69.72 | 72.50 | 4.06 | 4.40 | 33.34 | 661 |

We further measure the training speed (TSpeed) and training memory consumption (TMem) of the generation model along with the inference speed (ISpeed) and inference memory consumption (IMem) of the video VAE. All speed and memory metrics are measured on 17-frame 256×256 video clips with a batch size of 4 using a single NVIDIA H20 GPU. To ensure numerical stability, TSpeed and ISpeed are averaged over 100 steps following 50 warm-up steps.

Training details. We adopt the AdamW optimizer [28] with a base learning rate of 5×10^{-5} . The learning rate is linearly warmed up and decayed by a factor of 10 using a cosine schedule. During random dropping, the first frame is always retained, and the maximum dropping ratio r is set to 1.0. For generation, we remove the patchify downsampling module of the Latte model [29] to accommodate the higher spatiotemporal compression rate following [9]. The generation model is trained using rectified flow [27] for 250K steps with a learning rate of 1×10^{-4} and a global batch size of 64, and is evaluated with an Euler sampler using 100 steps. All experiments are conducted on 16 NVIDIA H800 GPUs.

4.2 Comparison

We compare PV-VAE with several representative video VAEs, including CogVideoX VAE (CogX-VAE)[57], IV-VAE[52], WF-VAE [25], HunyuanVideo VAE (Hunyuan-VAE)[22], Wan2.1 VAE, Wan2.2 VAE[49], and SSVAE [26].

Comparison on generation. table 1 reports the generation performance on UCF101 [39] and RealEstate10K [68] dataset. Our PV-VAE achieves the best overall performance among all models. Notably, compared with video VAEs using a $4 \times 8 \times 8$ downsampling factor, PV-VAE not only attains superior generation quality but also delivers substantial improvements in training speed and memory efficiency. Taking UCF-101 as an example, PV-VAE outperforms Hunyuan-VAE by 63.93 FVD and achieves a $2.68\times$ speedup in training while reducing memory consumption by 62%. Compared to Wan2.2 VAE / SSVAE, PV-VAE delivers a 34.42 / 22.31 FVD improvement despite using a higher latent-channel dimension. These results suggest that PV-VAE learns a richer and more structured latent spaces of motion and temporal dynamics, making it highly effective for video generative modeling.

Comparison on reconstruction. table 2 presents the reconstruction results on the Kinetics-400 [21] dataset. Video VAEs with $4 \times 8 \times 8$ compression typically yield better reconstruction metrics. In the context of $4 \times 16 \times 16$ models, PV-VAE delivers reconstruction performance comparable to existing video VAEs. It slightly underperforms relative to Wan2.2 VAE but consistently outperforms SSVAE. We also test the inference speed and memory consumption of different models at 256×256 resolution. Compared to Hunyuan-VAE / Wan2.2 VAE, PV-VAE achieves 38% / 19% faster inference while reducing memory consumption by 64% / 15%.

Qualitative comparison. We further present qualitative comparison results in figure 3. Under the same generative training settings, PV-VAE demonstrates superior visual fidelity over the Wan2.2 VAE, while

Table 2 Comparison of reconstruction performance on the Kinetics-400 validation set at different resolutions.

| Method | $17 \times 256 \times 256$ | | | | $17 \times 512 \times 512$ | | | | ISpeed (it/s) | IMem (GiB) |
|-------------------------|----------------------------|-------|-------|--------|----------------------------|-------|-------|--------|------------------|---------------|
| | rFVD↓ | PSNR↑ | SSIM↑ | LPIPS↓ | rFVD↓ | PSNR↑ | SSIM↑ | LPIPS↓ | | |
| CogX-VAE [57] | 4.90 | 33.78 | 0.97 | 0.027 | 1.79 | 36.00 | 0.99 | 0.024 | 0.46 | 13.64 |
| IV-VAE [52] | 2.78 | 34.08 | 0.96 | 0.019 | 0.97 | 37.24 | 0.96 | 0.016 | 0.32 | 5.39 |
| WF-VAE-L [25] | 3.06 | 33.48 | 0.96 | 0.023 | 1.08 | 35.93 | 0.96 | 0.023 | 0.87 | 5.00 |
| Hunyuan-VAE [22] | 2.96 | 34.30 | 0.97 | 0.016 | 0.90 | 37.13 | 0.97 | 0.015 | 0.50 | 22.00 |
| Wan2.1 VAE [49] | 2.92 | 33.21 | 0.95 | 0.018 | 1.02 | 36.15 | 0.97 | 0.017 | 0.60 | 6.77 |
| Wan2.2 VAE [49] | 3.42 | 33.78 | 0.96 | 0.015 | 1.22 | 36.75 | 0.97 | 0.015 | 0.58 | 9.36 |
| SSVAE [26] | 7.50 | 31.18 | 0.96 | 0.036 | 2.16 | 34.45 | 0.97 | 0.028 | 0.64 | 7.63 |
| PV-VAE | 3.45 | 32.26 | 0.95 | 0.020 | 1.88 | 35.03 | 0.97 | 0.020 | 0.69 | 7.97 |

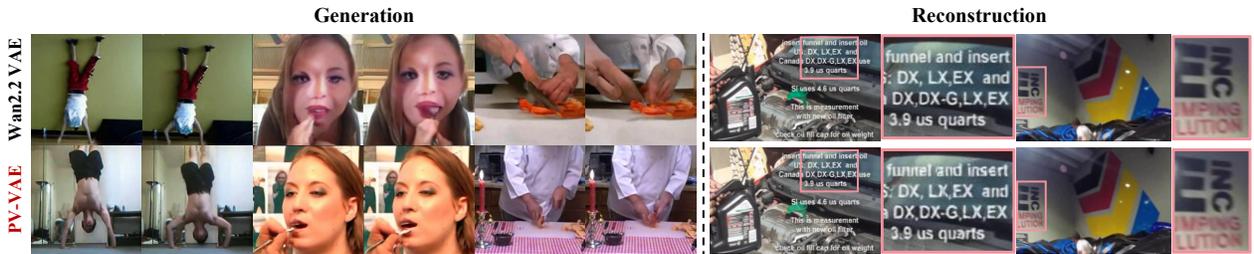


Figure 3 Qualitative comparison of generation and reconstruction. PV-VAE exhibits enhanced generative quality over Wan2.2 VAE while preserving competitive reconstruction fidelity.

exhibiting fewer motion artifacts and enhanced temporal coherence in video content. For reconstruction, we select two challenging cases. Notably, PV-VAE exhibits subtle limitations in reconstructing dense text, a performance gap likely stemming from the scarcity of text-heavy samples in our current data distribution [43]. Moving forward, we aim to incorporate more diverse datasets to further elevate the performance upper bound of PV-VAE.

4.3 Analysis

To better understand how the proposed predictive reconstruction works, we conduct extensive qualitative and quantitative analyses. Specifically, we dissect the latent space structure via principal component analysis (PCA), demonstrate the correlation between frame prediction accuracy and generation performance, investigate the scaling behaviors, and examine the latent temporal properties. Furthermore, we analyze the sources of PV-VAE’s performance gains using diffusion features on several downstream video understanding tasks. Finally, we provide visualizations of both reconstruction and future frame prediction to validate the effectiveness of our predictive reconstruction learning.

PCA analysis of latent space. To investigate the impact of predictive reconstruction on the structure of latent spaces, we perform PCA along the channel dimension of the latents and visualize the top three principal components as RGB images, as shown in figure 4. We randomly sample several videos from the Kinetics-400 [21] validation set and compare the PCA visualizations obtained from the baseline model and PV-VAE, alongside the corresponding optical flow computed by RAFT [41]. For each video, we display two non-adjacent frames to illustrate temporal dynamics. PV-VAE exhibits a clear correspondence between latents and underlying motion, with activation patterns strongly aligned with optical flow. Regions with high activation coincide with large motion vectors. In the left visualizations, the person doing push-ups and the one performing a long jump exhibit noticeably stronger activations than the background. Similarly, in the right visualizations, the hands of the cello player and the arms and hands of the person playing cards receive higher attention, indicating that the model effectively concentrates spatiotemporal saliency within its latent space. Moreover, we observe that the background regions with small motion vectors exhibit reduced noise compared to the

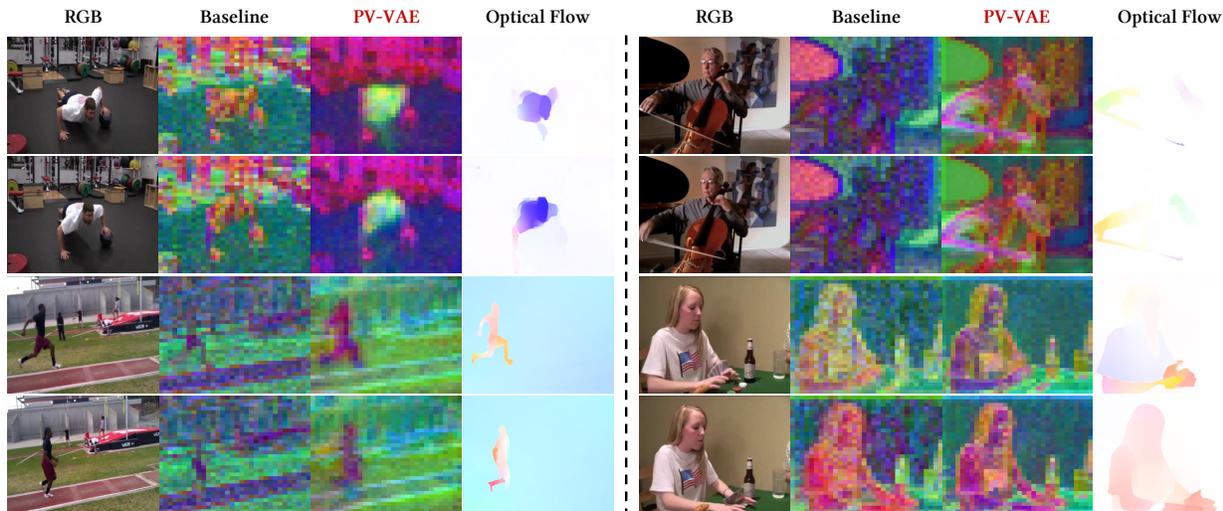


Figure 4 PCA analysis of latent space structure. PV-VAE exhibits a clear correspondence between latent activations and underlying video motion, with activation patterns strongly aligned with optical flow, indicating that our model effectively concentrates spatiotemporal saliency within its latent representations.

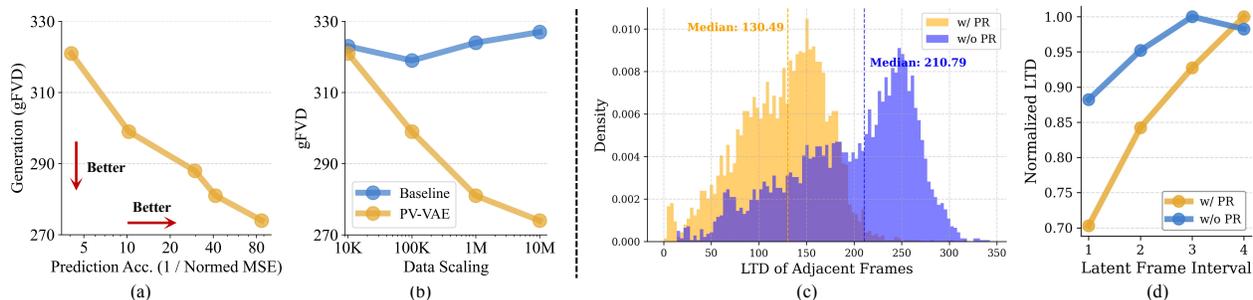


Figure 5 (a): Correlation between generation and prediction accuracy. **(b):** Scalability of the predictive reconstruction objective. **(c):** Short-term temporal smoothness. PV-VAE achieves higher adjacency coherence than the baseline. **(d):** Long-term temporal dynamics. PV-VAE demonstrates a monotonic latent trajectory across expanding frame intervals. These results collectively validate the effectiveness of predictive reconstruction in imposing structured temporal constraints on video latents.

baseline, suggesting that PV-VAE encourages the latent space to allocate more representational bandwidth to dynamic foregrounds while maintaining smoother, lower-variance representations for static areas.

Correlation study and scaling behaviors. To verify the synergy between future prediction and generation, we conduct a correlation study as shown in figure 5(a). The results confirm that improved predictive accuracy consistently translates into superior generative performance, justifying our core motivation. On this basis, we further investigate the scaling behavior of PV-VAE in figure 5(b). We observe consistent performance gains as training data scales, a trend notably absent with the pure reconstruction objective, highlighting the superior scalability of our predictive reconstruction paradigm.

Latent temporal coherence. To evaluate temporal coherence, we introduce the Latent Temporal Distance (LTD) metric, computed as the average L_2 distance between latents across varying intervals for 1,000 Kinetics-400 validation videos. As shown in figure 5(c), PV-VAE exhibits a lower median and a sharper histogram peak in adjacent-frame LTD compared to the baseline, suggesting smoother temporal transitions. Furthermore, as frame intervals grow, PV-VAE demonstrates a consistent monotonic increase in normalized LTD, whereas the baseline lacks this trend, as shown in figure 5(d). This reveals a smoothly evolving latent trajectory that effectively captures continuous video dynamics, confirming the role of predictive reconstruction in promoting temporal consistency.

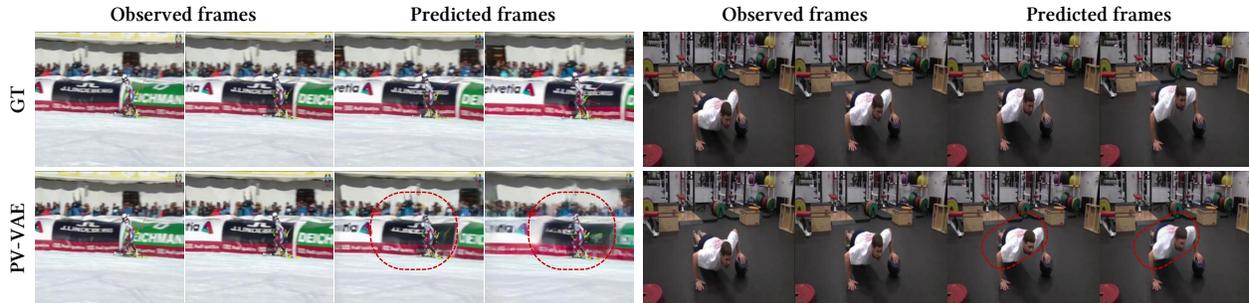


Figure 6 Frame Prediction Validation. PV-VAE generates plausible future frames aligned with underlying temporal evolutions. Red dotted circles highlight shifts in relative object positioning (best viewed under zoom-in).

Table 3 Probing results on three video understanding tasks. Compared to the baseline model, PV-VAE achieves consistent gains across all tasks, indicating that our method enhances the learned representations with stronger video understanding.

| Method | EPE ↓ | MSE ↓ | AUC(%) ↑ |
|--------|------------------------|-----------------------|----------------------|
| w/o PR | 5.9223 | 0.0314 | 70.95 |
| w/ PR | 5.1805 (+12.5%) | 0.0289 (+8.0%) | 76.99 (+8.5%) |

Probing video understanding in the latent space. To dissect the sources of performance gains, we examine the learned latent spaces through the lens of diffusion features across three representative video understanding tasks: optical flow estimation [15], next-frame prediction [69], and point tracking [13], as shown in table 3. Features are extracted from the 14th layer (out of 28) of the LVDM for all tasks, with specific configurations detailed below:

- **Optical flow estimation:** We utilize the Sintel [8] dataset, employing a task-specific decoder with 3D convolutions and pixel-shuffle operations to upsample LVDM features to the original resolution. Performance is quantified by the Average End-Point Error (EPE).
- **Next-frame prediction:** Evaluating on Kinetics-400 [21], we adapt the flow decoder by adjusting its output channels to three for RGB prediction, reporting the Mean Squared Error (MSE).
- **Point tracking:** We evaluate on the TAP-Vid-DAVIS [31] dataset, which contains 30 videos annotated with query points and corresponding ground-truth trajectories. We report the Area Under the Curve (AUC) of tracking accuracy across error thresholds from 0 to 10 pixels.

Compared to the baseline model, PV-VAE achieves consistent improvements across all three tasks, demonstrating that its latent space encodes superior video dynamics and motion-aware representations. These findings suggest that the enhanced generative performance stems from a more robust understanding of fundamental video properties, highlighting the potential of predictive reconstruction as a promising direction for video modeling.

Predictive reconstruction visualization. Finally, we showcase the prediction capabilities of PV-VAE in figure 6. For each video, we discard the latter half of the frames and task the model with reconstructing the entire sequence. Two observed frames from the observed half and two predicted frames from the unobserved half are shown. PV-VAE not only reconstructs the observed frames but also generates plausible future frames that align with the underlying video dynamics. For instance, the model accurately predicts relative spatial shifts between subjects and backgrounds while capturing the temporal progression of actions. These results provide compelling evidence that PV-VAE effectively captures complex temporal dependencies in video data.

4.4 Ablation study

Incremental ablation of PV-VAE. We first perform an incremental ablation study to dissect the contribution of each key component in PV-VAE, as summarized in table 4. The introduction of predictive reconstruction

Table 4 Incremental ablation of PV-VAE. Generation (UCF-101) and reconstruction (Kinetics-400) performance across different configurations. All results are measured at 256×256 resolution.

| Method | gFVD ↓ | rFVD ↓ | PSNR ↑ | SSIM ↑ | LPIPS ↓ |
|-----------------------------|--------|--------|--------|--------|---------|
| Baseline | 174.81 | 3.03 | 33.44 | 0.96 | 0.017 |
| + Predictive Reconstruction | 156.33 | 5.66 | 31.47 | 0.94 | 0.026 |
| + Motion-aware Objective | 150.10 | 5.79 | 31.38 | 0.94 | 0.026 |
| + Decoder Fine-tuning | 146.37 | 3.45 | 32.26 | 0.95 | 0.020 |

Table 5 Ablation on maximum dropping ratio (MDR).

| MDR | gFVD ↓ | KVD ↓ | IS ↑ |
|------|---------------|--------------|--------------|
| 50% | 159.82 | 14.67 | 69.35 |
| 75% | 154.06 | 16.93 | 70.27 |
| 100% | 146.37 | 14.52 | 69.72 |

Table 6 Ablation on padding strategy for latents.

| Padding | gFVD ↓ | KVD ↓ | IS ↑ |
|-----------|---------------|--------------|--------------|
| Gaussian | 150.68 | 11.87 | 68.01 |
| Learnable | 146.37 | 14.52 | 69.72 |

markedly enhances generation performance, while the motion-aware objective also yields positive gains. Furthermore, decoder fine-tuning significantly improves reconstruction quality, from which the generation metrics also derive a slight benefit.

Maximum dropping ratio. We also investigate the impact of the maximum dropping ratio r , as shown in table 5. Specifically, we set r to 50%, 75%, and 100%, respectively. Since reconstruction shows marginal differences following decoder fine-tuning, we focus on comparing the generation performance on the UCF-101 dataset. The results show that generative performance consistently improves with higher perturbation levels, indicating that stronger predictive regularization encourages the learning of more robust and higher-quality representations. Therefore, we set the maximum dropping ratio r to 100% in our training setup.

Padding strategies for latents. We further conduct an ablation study on how to pad the video latents. Specifically, we compare two strategies: (i) sampling \mathbf{z}_{pad} from a standard Gaussian distribution, and (ii) using learnable tokens following masked modeling practice [44]. As shown in table 6, the learnable tokens yield slightly better generation quality.

5 Discussion and Conclusion

Generation vs. Reconstruction. The trade-off between reconstruction and generation remains central to tokenizer design. Rich latent information enhances reconstruction fidelity yet complicates generative modeling, while a highly compressed latents facilitates generation but sacrifices detail. Previous works typically adopt a low-dimensional latent space to strike a balance. Recent advancements [37, 58, 59] show that high-dimensional latents can actually facilitate generation if they are structured by pre-trained or self-supervised priors. However, we argue that for video, a structured latent space must encompass both semantics and motion. Our PV-VAE shifts the focus from “what is in the frame” to “what happens next”. By employing a predictive reconstruction objective, we ensure the latent space is motion-aware rather than a mere pixel container. Notably, this predictive philosophy can be generalized to masked modeling, such as frame infilling or joint spatio-temporal prediction. Such self-supervised paradigms could further bolster the robustness and versatility of video latent spaces, a direction we intend to explore in future work.

Advantages of multi-stage training. We introduce an additional decoder fine-tuning stage, a strategic design aimed at further enhancing reconstruction, drawing inspiration from successful approaches in the image domain [37, 55]. Our empirical observations reveal that this stage serves as an *effective “free lunch” with bounded gains*. It consistently refines reconstruction quality while preserving latent diffusability by keeping the encoder frozen. In addition, since the encoder remains unchanged during this phase, the decoder fine-tuning can be conducted in parallel with the diffusion backbone (*e.g.*, DiT) training, thereby substantially accelerating the overall development and iteration efficiency.

Rethinking video VAE Architecture. Next, we discuss the architectural design of video VAEs. Despite the

Table 7 Performance and efficiency comparison between CNN and Transformer-based video VAEs. Results are evaluated at 256×256 resolution. * denotes our optimized Transformer-based variant.

| Padding | UCF101 | | | Kinetics-400 | | | | ISpeed (it/s) |
|---------|--------|-------|-------|--------------|--------|--------|---------|------------------|
| | gFVD ↓ | KVD ↓ | IS ↑ | rFVD ↓ | PSNR ↑ | SSIM ↑ | LPIPS ↓ | |
| PV-VAE | 146.37 | 14.52 | 69.72 | 3.45 | 32.26 | 0.95 | 0.020 | 0.69 |
| PV-VAE* | 178.86 | 20.66 | 69.80 | 4.03 | 33.02 | 0.95 | 0.022 | 1.29 |

dominance of Vision Transformers (ViT) [14] across most vision tasks, existing video VAEs [22, 25, 49] still predominantly rely on 3D causal convolutions. This reliance prevents video VAEs from leveraging the vast ecosystem of modern techniques optimized for Transformer architectures, while also incurring heavy computational overhead and lacking global modeling capabilities. To address these limitations, we explore a minimalist, plain Transformer-based video VAE under the same spatiotemporal compression ratio ($4 \times 16 \times 16$, $C=64$). The input is first divided into $4 \times 16 \times 16$ spatiotemporal patches, then processed by a stack of Transformer blocks. The decoder directly upsamples the representations back to the original resolution using a pixel-shuffle operation. Both the encoder and decoder consist of 12 layers each, featuring 16 heads with a 128 *head_dim*, amounting to a total parameter count of roughly 1.2B. We compare the reconstruction and generative performance, and inference speed against its CNN-based counterpart, as shown in table 7. Our findings indicate that while the Transformer-based PV-VAE* achieves comparable reconstruction fidelity, its generative capability remains limited.

Despite the current generative gap compared to CNN-based models, we contend that Transformer-based video VAEs hold significant promise for future research, primarily for the following two reasons: (i) *Computational efficiency*: Despite having a larger number of parameters, the Transformer variant achieves 87% faster inference speed, effectively mitigating the efficiency bottleneck of current video VAEs, especially when processing long video sequences. (ii) *Representational flexibility*: The Transformer architecture naturally integrates with various video representation learning paradigms, allowing it to flexibly incorporate diverse self-supervised objectives [44] and further improve latent representations for generative modeling. In future work, we will further explore optimized architectural configurations and training recipes for Transformer-based video VAEs to fully unlock their latent potential.

Conclusion. In this work, we present Predictive Video VAE (PV-VAE), which incorporates a predictive reconstruction objective to jointly optimize visual fidelity and temporal dynamics. This approach yields a more temporally structured and generation-ready video latent space. Extensive downstream evaluations and in-depth analyses demonstrate that PV-VAE effectively captures motion-aware representations, leading to substantial gains in video generation performance. We hope our findings provide meaningful insights for future video VAE research and help push the frontiers of video generative modeling.

References

- [1] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 15619–15629, 2023.
- [2] Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zhohus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. arXiv preprint arXiv:2506.09985, 2025.
- [3] Federico Baldassarre, Marc Szafraniec, Basile Terver, Vasil Khalidov, Francisco Massa, Yann LeCun, Patrick Labatut, Maximilian Seitzer, and Piotr Bojanowski. Back to the features: Dino as a foundation for video world models. arXiv preprint arXiv:2507.19468, 2025.
- [4] Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mido Assran, and Nicolas Ballas. V-jepa: Latent video prediction for visual representation learning. 2023.
- [5] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127, 2023.
- [6] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. OpenAI Blog, 1(8):1, 2024.
- [7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. Advances in neural information processing systems, 33:1877–1901, 2020.
- [8] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In A. Fitzgibbon et al. (Eds.), editor, European Conf. on Computer Vision (ECCV), Part IV, LNCS 7577, pages 611–625. Springer-Verlag, October 2012.
- [9] Junyu Chen, Han Cai, Junsong Chen, Enze Xie, Shang Yang, Haotian Tang, Muyang Li, Yao Lu, and Song Han. Deep compression autoencoder for efficient high-resolution diffusion models. arXiv preprint arXiv:2410.10733, 2024.
- [10] Liuhan Chen, Zongjian Li, Bin Lin, Bin Zhu, Qian Wang, Shenghai Yuan, Xing Zhou, Xinhua Cheng, and Li Yuan. Od-vae: An omni-dimensional video compressor for improving latent video diffusion model. arXiv preprint arXiv:2409.01199, 2024.
- [11] Yu Cheng and Fajie Yuan. Leanvae: An ultra-efficient reconstruction vae for video diffusion models. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 15692–15702, 2025.
- [12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers), pages 4171–4186, 2019.
- [13] Carl Doersch, Ankush Gupta, Larisa Markeeva, Adria Recasens, Lucas Smaira, Yusuf Aytar, Joao Carreira, Andrew Zisserman, and Yi Yang. Tap-vid: A benchmark for tracking any point in a video. Advances in Neural Information Processing Systems, 35:13610–13626, 2022.
- [14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020.
- [15] David Fleet and Yair Weiss. Optical flow estimation. In Handbook of mathematical models in computer vision, pages 237–257. Springer, 2006.
- [16] Yu Gao, Haoyuan Guo, Tuyen Hoang, Weilin Huang, Lu Jiang, Fangyuan Kong, Huixia Li, Jiashi Li, Liang Li, Xiaojie Li, et al. Seedance 1.0: Exploring the boundaries of video generation models. arXiv preprint arXiv:2506.09113, 2025.
- [17] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. Communications of the ACM, 63(11):139–144, 2020.

- [18] Agrim Gupta, Jiajun Wu, Jia Deng, and Fei-Fei Li. Siamese masked autoencoders. Advances in Neural Information Processing Systems, 36:40676–40693, 2023.
- [19] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 16000–16009, 2022.
- [20] Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In 2010 20th international conference on pattern recognition, pages 2366–2369. IEEE, 2010.
- [21] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. arXiv preprint arXiv:1705.06950, 2017.
- [22] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. arXiv preprint arXiv:2412.03603, 2024.
- [23] Yann LeCun. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. Open Review, 62(1): 1–62, 2022.
- [24] Xingjian Leng, Jaskirat Singh, Yunzhong Hou, Zhenchang Xing, Saining Xie, and Liang Zheng. Repa-e: Unlocking vae for end-to-end tuning of latent diffusion transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 18262–18272, 2025.
- [25] Zongjian Li, Bin Lin, Yang Ye, Liuhan Chen, Xinhua Cheng, Shenghai Yuan, and Li Yuan. Wf-vae: Enhancing video vae by wavelet-driven energy flow for latent video diffusion model. In Proceedings of the Computer Vision and Pattern Recognition Conference, pages 17778–17788, 2025.
- [26] Shizhan Liu, Xinran Deng, Zhuoyi Yang, Jiayan Teng, Xiaotao Gu, and Jie Tang. Delving into latent spectral biasing of video vaes for superior diffusability. arXiv preprint arXiv:2512.05394, 2025.
- [27] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. arXiv preprint arXiv:2209.03003, 2022.
- [28] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101, 2017.
- [29] Xin Ma, Yaohui Wang, Gengyun Jia, Xinyuan Chen, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation. arXiv preprint arXiv:2401.03048, 2024.
- [30] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193, 2023.
- [31] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 724–732, 2016.
- [32] Lucas Pinheiro Cinelli, Matheus Araújo Marins, Eduardo Antúnio Barros da Silva, and Sérgio Lima Netto. Variational autoencoder. In Variational methods for machine learning with applications to deep networks, pages 111–149. Springer, 2021.
- [33] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 10684–10695, 2022.
- [34] Michael S Ryoo. Human activity prediction: Early recognition of ongoing activities from streaming videos. In 2011 international conference on computer vision, pages 1036–1043. IEEE, 2011.
- [35] Seyedmorteza Sadat, Jakob Buhmann, Derek Bradley, Otmar Hilliges, and Romann M Weber. Litevae: Lightweight and efficient variational autoencoders for latent diffusion models. Advances in Neural Information Processing Systems, 37:3907–3936, 2024.
- [36] Masaki Saito, Shunta Saito, Masanori Koyama, and Sosuke Kobayashi. Train sparsely, generate densely: Memory-efficient unsupervised training of high-resolution temporal gan. International Journal of Computer Vision, 128(10):2586–2606, 2020.

- [37] Qingyu Shi, Size Wu, Jinbin Bai, Kaidong Yu, Yujing Wang, Yunhai Tong, Xiangtai Li, and Xuelong Li. Rectok: Reconstruction distillation along rectified flow. [arXiv preprint arXiv:2512.13421](#), 2025.
- [38] Ivan Skorokhodov, Sharath Girish, Benran Hu, Willi Menapace, Yanyu Li, Rameen Abdal, Sergey Tulyakov, and Aliaksandr Siarohin. Improving the diffusability of autoencoders. [arXiv preprint arXiv:2502.14831](#), 2025.
- [39] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. [arXiv preprint arXiv:1212.0402](#), 2012.
- [40] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. *Advances in Neural Information Processing Systems*, 36:1363–1389, 2023.
- [41] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pages 402–419. Springer, 2020.
- [42] Hansi Teng, Hongyu Jia, Lei Sun, Lingzhi Li, Maolin Li, Mingqiu Tang, Shuai Han, Tianning Zhang, WQ Zhang, Weifeng Luo, et al. Magi-1: Autoregressive video generation at scale. [arXiv preprint arXiv:2505.13211](#), 2025.
- [43] Shengbang Tong, Boyang Zheng, Ziteng Wang, Bingda Tang, Nanye Ma, Ellis Brown, Jihan Yang, Rob Fergus, Yann LeCun, and Saining Xie. Scaling text-to-image diffusion transformers with representation autoencoders. [arXiv preprint arXiv:2601.16208](#), 2026.
- [44] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022.
- [45] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.
- [46] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. [arXiv preprint arXiv:1812.01717](#), 2018.
- [47] Pedro Vélez, Luisa F Polanía, Yi Yang, Chuhan Zhang, Rishabh Kabra, Anurag Arnab, and Mehdi SM Sajjadi. From image to video: An empirical study of diffusion representations. [arXiv preprint arXiv:2502.07001](#), 2025.
- [48] Tuan-Hung Vu, Catherine Olsson, Ivan Laptev, Aude Oliva, and Josef Sivic. Predicting actions from static scenes. In *European Conference on Computer Vision*, pages 421–436. Springer, 2014.
- [49] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. [arXiv preprint arXiv:2503.20314](#), 2025.
- [50] Yuchi Wang, Junliang Guo, Xinyi Xie, Tianyu He, Xu Sun, and Jiang Bian. Vidtwins: Video vae with decoupled structure and dynamics. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22922–22932, 2025.
- [51] 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.
- [52] Pingyu Wu, Kai Zhu, Yu Liu, Liming Zhao, Wei Zhai, Yang Cao, and Zheng-Jun Zha. Improved video vae for latent video diffusion model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18124–18133, 2025.
- [53] Yushu Wu, Yanyu Li, Ivan Skorokhodov, Anil Kag, Willi Menapace, Sharath Girish, Aliaksandr Siarohin, Yanzhi Wang, and Sergey Tulyakov. H3ae: High compression, high speed, and high quality autoencoder for video diffusion models. [arXiv preprint arXiv:2504.10567](#), 2025.
- [54] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9653–9663, 2022.
- [55] Jiawei Yang, Tianhong Li, Lijie Fan, Yonglong Tian, and Yue Wang. Latent denoising makes good tokenizers. In *The Fourteenth International Conference on Learning Representations*, 2026.

- [56] Shusheng Yang, Jihan Yang, Pinzhi Huang, Ellis Brown, Zihao Yang, Yue Yu, Shengbang Tong, Zihan Zheng, Yifan Xu, Muhan Wang, et al. Cambrian-s: Towards spatial supersensing in video. [arXiv preprint arXiv:2511.04670](#), 2025.
- [57] 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.
- [58] Jingfeng Yao, Yuda Song, Yucong Zhou, and Xinggang Wang. Towards scalable pre-training of visual tokenizers for generation. [arXiv preprint arXiv:2512.13687](#), 2025.
- [59] Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In [Proceedings of the Computer Vision and Pattern Recognition Conference](#), pages 15703–15712, 2025.
- [60] Xiangchen Yin, Jiahui Yuan, Zhangchi Hu, Wenzhang Sun, Jie Chen, Xiaozhen Qiao, Hao Li, and Xiaoyan Sun. Deco-vae: Learning compact latents for video reconstruction via decoupled representation. [arXiv preprint arXiv:2511.14530](#), 2025.
- [61] Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. [arXiv preprint arXiv:2410.06940](#), 2024.
- [62] Sihyun Yu, Weili Nie, De-An Huang, Boyi Li, Jinwoo Shin, and Anima Anandkumar. Efficient video diffusion models via content-frame motion-latent decomposition. [arXiv preprint arXiv:2403.14148](#), 2024.
- [63] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In [Proceedings of the IEEE conference on computer vision and pattern recognition](#), pages 586–595, 2018.
- [64] Shilong Zhang, He Zhang, Zhifei Zhang, Chongjian Ge, Shuchen Xue, Shaoteng Liu, Mengwei Ren, Soo Ye Kim, Yuqian Zhou, Qing Liu, et al. Both semantics and reconstruction matter: Making representation encoders ready for text-to-image generation and editing. [arXiv preprint arXiv:2512.17909](#), 2025.
- [65] 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.
- [66] Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. [arXiv preprint arXiv:2510.11690](#), 2025.
- [67] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all. [arXiv preprint arXiv:2412.20404](#), 2024.
- [68] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. [arXiv preprint arXiv:1805.09817](#), 2018.
- [69] Yufan Zhou, Haiwei Dong, and Abdulmoteleb El Saddik. Deep learning in next-frame prediction: A benchmark review. [IEEE Access](#), 8:69273–69283, 2020.
- [70] Zixin Zhu, Xuelu Feng, Dongdong Chen, Junsong Yuan, Chunming Qiao, and Gang Hua. Exploring pre-trained text-to-video diffusion models for referring video object segmentation. In [European Conference on Computer Vision](#), pages 452–469. Springer, 2024.