Delving into Latent Spectral Biasing of Video VAEs for Superior Diffusability

Shizhan Liu[*1], Xinran Deng[*1,3], Zhuoyi Yang[†2], Jiayan Teng[2], Xiaotao Gu[1], and Jie Tang[‡2]

[1] Zhipu AI, Beijing, China
[2] Tsinghua University, Beijing, China
[3] University of Chinese Academy of Sciences, Beijing, China

Abstract. Latent diffusion models pair VAEs with diffusion backbones, and the structure of VAE latents strongly influences the difficulty of diffusion training. However, existing video VAEs typically focus on reconstruction fidelity, overlooking latent structure. We present a statistical analysis of video VAE latent spaces and identify two spectral properties essential for diffusion training: a channel-wise eigenspectrum dominated by a few modes, and a spatio-temporal frequency spectrum biased toward low frequencies. To induce these properties, we propose two lightweight, backbone-agnostic regularizers: Latent Masked Reconstruction and Local Correlation Regularization. Experiments show that our Spectral-Structured VAE (SSVAE) achieves a 3× speedup in text-to-video generation convergence and a 10% gain in video reward, outperforming strong open-source VAEs. Code is available at: https://github.com/zai-org/SSVAE.

Keywords: Video VAE · Diffusability · Text-to-Video Diffusion Model

1 Introduction

Latent video diffusion models [11, 24, 27] have recently advanced text-to-video (T2V) generation by coupling a 3D VAE-based tokenizer with a diffusion backbone. A growing body of evidence indicates that the VAE strongly shapes downstream diffusion training dynamics [4, 23, 28].

Unfortunately, most existing video VAEs, including those adopted in SOTA T2V models, pursue better temporal compression and high-fidelity reconstruction through architectural design [15,17,26] and the optimization of reconstruction-based objectives [11,24,27] (e.g., MSE, adversarial, LPIPS). The latent structure, however, receives limited attention in their optimization. This objective-target mismatch leads to a well-known phenomenon: stronger reconstruction fidelity does not necessarily translate into better generative utility [28].

Prior work on image VAEs has proposed regularizers and analyses from various perspectives, yet the central question remains open: **What properties should a latent space possess to facilitate downstream generative**

* Equal contribution. †Project leader. ‡Corresponding author.

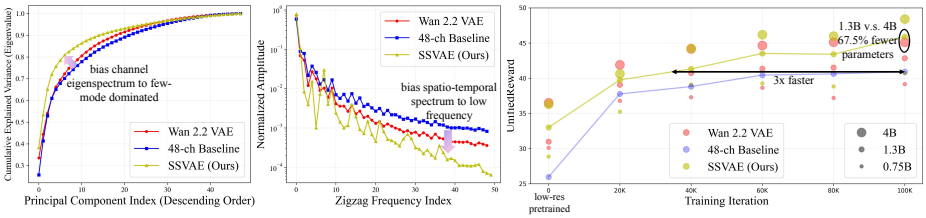


Fig. 1: We identify that both a low-frequency biased spatio-temporal frequency spectrum and a few-mode biased channel eigenspectrum facilitate diffusion training. By inducing low-frequency bias, few-mode bias and enhancing decoder robustness, our SSVAE achieves a $3\times$ convergence speedup over the baseline on $17\times512\times512$ generation, using prompts from VBench.

training? Existing analyses often rely on indirect proxy metrics, such as ImageNet linear probing accuracy [4] or similarity to DINOv2 features [14,18], yielding only an indirect characterization of VAE latents. Optimizing VAEs for these metrics often requires specialized designs. Other representation-level explanations, such as latent clustering [28], provide useful intuition but lack measurable quantities that connect the intuition to generation quality. Moreover, existing analyses are usually studied in isolation, leaving their overlap and complementarity unclear, which makes it difficult to distill a principled set of 'properties beneficial for generation.'

In this work, we seek to identify properties of video VAE latent spaces that facilitate diffusion training, which hereafter referred to as *diffusability*, through statistical analysis of VAE latents. As direct characterizations of the latent space, statistical measures can not only offer actionable targets for designing regularizers, but also shed light on the mechanisms behind existing proxy-metric approaches. Moreover, the overlap and complementarity among various statistical analyses open up possibilities for combining different regularizations. Specifically, for video VAE latents with temporal, height, width, and channel dimensions, we analyze two complementary spectra: the channel-wise eigenspectrum and the spatio-temporal frequency spectrum. **Our key finding is that biased, rather than uniform, spectra lead to improved diffusability.**

For the eigenspectrum, we show that a few-mode-biased channel-wise eigenspectrum enhances diffusability. In other words, the latent code has a low effective rank and can be closely approximated by a linear combination of only a few basis vectors. By analyzing the learning dynamics of diffusion models along each eigenvector, we reveal the mechanism by which few-mode bias accelerates convergence. Finally, we propose Latent Masked Reconstruction (LMR), which simultaneously promotes few-mode bias and improves decoder robustness, resulting in substantially faster convergence of diffusion models.

We then turn to the complementary spatio-temporal frequency spectrum. SER [23] has shown that spatial frequency spectra biased toward low-frequency components improve latent-space diffusability, and we extend this analysis to

the spatio-temporal domain, demonstrating that two common strategies in image VAEs, scale-equivariant regularization [12, 23] and foundation-model alignment [14, 28], both promote low-frequency bias. However, while they introduce non-negligible computational overhead, they do not adequately address the temporal dimension in video latents. From a statistical perspective, we find that the proportion of low-frequency energy is positively correlated with the local spatio-temporal correlation of the latents. This motivates our development of Local Correlation Regularization (LCR), a computationally efficient regularizer that explicitly enhances local correlation, thereby increasing low-frequency energy.

Based on LMR and LCR, we train a **Spectral-Structured VAE (SSVAE)** that significantly improves diffusability, as shown in Fig. 1. We summarize our contributions as the following:

- We present the first detailed statistical analysis of the latent space in video VAEs, revealing two spectral properties that are crucial for generative training: a channel eigenspectrum dominated by few modes, and a spatio-temporal frequency spectrum biased toward low frequencies. Our findings provide new insights into the underlying mechanisms of existing methods and offer actionable design targets for regularization.
- We propose Latent Masked Reconstruction (LMR) to promote few-mode bias and decoder robustness to noise, and Local Correlation Regularization (LCR) to bias latents toward low frequencies. Both are lightweight, backbone-agnostic, and easy to implement.
- Extensive experiments across diffusion backbones and parameter scales show that our training recipe substantially accelerates convergence by 3× and improves video reward by 10%, consistently outperforming strong open-source VAE baselines for text-to-video.

2 Related Work

2.1 Video VAE

Existing video VAEs primarily focus on improving spatio-temporal compression and video reconstruction. Early efforts extend image VAEs to video by leveraging temporal redundancy for compression [7, 30]. Subsequent work explores keyframe-based temporal compression and grouped causal convolutions to mitigate inter-frame imbalance [26], and multi-level wavelet transforms to enhance temporal coherence with efficient architectures [15]. In large-scale text-to-video models, 3D CNN-based VAEs have become standard tokenizers, as adopted by CogVideoX [27], HunyuanVideo [11], and Wan2.1 [24].

Despite solid progress on compression and reconstruction fidelity, most video VAEs are still trained with reconstruction-centric objectives (e.g., L1/L2, LPIPS, adversarial losses), with limited attention to the latent space properties that affect downstream diffusion training. Consequently, they are ill-equipped to address the well-known trade-off between reconstruction and generation [28]. Our work focuses on shaping video VAE latents for generation.

2.2 Generation-Oriented Image Tokenizer

Numerous analyses and regularizers have been proposed for generation-oriented image tokenizers. Some works design and optimize proxy metrics linked to VAE latents, such as MAEtok [4], which correlates ImageNet linear probing accuracy with generative quality, and REPA-E [14], which uses diffusion backbone–DINOv2 feature similarity. However, the relationships between proxy metrics and VAE latents are often complex and indirect, making it difficult to identify underlying principles. Other works visualize the latent space and offer qualitative interpretations. For example, VA-VAE [28] claims latent clustering is detrimental, and DC-AE-1.5 [6] emphasizes preserving object structure. Yet their methods only indirectly optimize the objectives motivated by these analyses, which limits the extent to which analytical improvements translate into better generation. There are also works that perform frequency-based analyses, such as SER [23], which suppress high-frequency components indirectly by enforcing scale equivariance, but at notable computational cost.

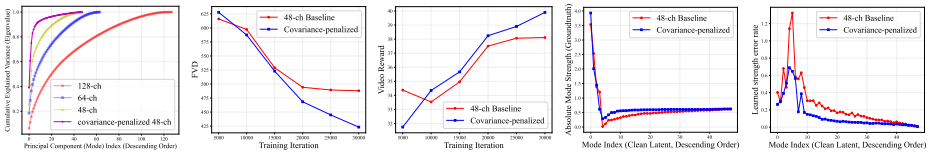
Despite these diverse perspectives, most analyses are conducted in isolation. As a result, a principled set of generative-beneficial properties is still lacking.

3 Channel Eigenspectrum Shaping

We first analyze the channel eigenspectrum of video VAEs. We observe that VAEs exhibiting Few-Mode Bias (FMB) and enhanced decoder robustness possess better diffusability. Next, we detail the impact of FMB on diffusion training in Sec. 3.1, and analyze its convergence accelerating mechanism from a cross-correlation view in Sec. 3.2. Finally, Sec. 3.3 introduces the Latent Masked Reconstruction (LMR) to simultaneously promotes FMB and decoder robustness.

3.1 Impact of FMB on Diffusion Training

We begin by analyzing the channel eigenspectrum of VAEs with different channel counts, motivated by the well-known observation that downstream generative training converges slower on VAEs with more channels. For a C-channel VAE, let $\mathbf{u}^0 \in \mathbb{R}^{1\times C}$ denote a latent vector sampled from the normalized latent $\tilde{\mathbf{z}}$, where the superscript 0 denotes the clean latent at diffusion timestep 0. The channel-wise autocorrelation matrix $\Sigma_{\mathbf{uu}} \in \mathbb{R}^{C\times C}$ is then computed as $\Sigma_{\mathbf{uu}} = \mathbb{E}[(\mathbf{u}^0)^\top \mathbf{u}^0]$, where the expectation is taken over the sample and spatio-temporal dimensions. Subsequently, we apply principal component analysis (PCA) to $\Sigma_{\mathbf{uu}}$ to analyze its eigenvalue spectrum. We train 48-, 64-, and 128-channel VAEs using standard losses (L1, KL, LPIPS, and adversarial), and compare the cumulative explained variance curves in Fig. 2a. Here, 'explained variance' corresponds to eigenvalues sorted in descending order. The differences are substantial: **higher-channel VAEs tend to distribute eigenvalues more evenly across eigenvectors, whereas lower-channel VAEs concentrate eigenvalues in a few dominant eigenvectors.** We term this phenomenon



(a) Mode strength comparison of VAE latent covariance. **(b)** Generation quality comparison by FVD. **(c)** Generation quality comparison by UnifiedReward. **(d)** Mode strength comparison of cross correlation. **(e)** Mode-wise learned error in cross correlation.

Fig. 2: Comparative analysis of the VAE latent channel covariance matrix and the diffusion output–input cross-correlation matrix. A few-mode-biased latent space is associated with a lower FVD, higher generation quality, and faster convergence.

'Few-Mode Bias' (FMB). Here, 'mode' refers to the eigenvector. A highly few-mode biased latent space implies a low effective rank across channels and can be well approximated by a linear combination of a few eigenvectors.

We further investigate the impact of FMB on generation training in VAEs with equal channel counts. We train a baseline 48-channel VAE and a covariance-regularized variant by penalizing the total eigenvalues outside the top three principal components. Our findings are as follows: (i) **The FVD trend is consistent with the eigenvalue distribution.** The covariance-penalized VAE effectively induces a few-mode bias and achieves a lower FVD, as shown in Fig. 2a and Fig. 2b. (ii) **Few-mode biased models achieve higher video rewards**, as depicted in Fig. 2c. Based on these observations, we hypothesize that a few-mode biased latent space benefits diffusion training and proceed to analyze the underlying reasons.

3.2 A Cross-Correlation View of FMB

To uncover why a few-mode biased latent space accelerates diffusion training, we analyze the learning dynamics using output-input cross-correlation, which is employed in the convergence analysis framework of deep linear networks [21] and has recently been shown to be effective for non-linear diffusion models [2]. Output-input cross-correlation captures the statistical dependencies that the model needs to learn. By analyzing its eigenspectrum, we can quantitatively evaluate training progress for each 'correlation mode', namely the eigenvectors of the groundtruth cross-correlation matrix. A mode is considered well learned if the strength of the learned cross-correlation along its corresponding eigenvector approaches the groundtruth eigenvalue.

Specifically, let $\Sigma_{\mathbf{vu}}(t) \in \mathbb{R}^{C\times C}$ denote the channel-wise output-input cross-correlation matrix of the diffusion backbone at timestep t, and let λ_l denote the l-th largest eigenvalue of the autocorrelation $\Sigma_{\mathbf{uu}}$ that we analyzed in Sec. 3.1. Then we have the following theorem under flow-matching and velocity prediction.

Theorem 1. *$\Sigma_{\mathbf{vu}}(t)$ has the same eigenvectors as $\Sigma_{\mathbf{uu}}$, and its eigenvalue on the l-th eigenvector of $\Sigma_{\mathbf{uu}}$ is given by:*

$$s_l(t) = t - (1-t)\lambda_l. \tag{1}$$

Please refer to the supplementary Sec. 1 for the proof and scope of Theorem 1.

This eigenspectrum relationship bridges the structure of VAE latents and the diffusion model's training dynamics. Since $\Sigma_{\mathbf{vu}}(t)$ and $\Sigma_{\mathbf{uu}}$ share eigenvectors, convergence analysis of $\Sigma_{\mathbf{vu}}(t)$ allows us to assess the diffusion backbone's learning progress in each mode of the clean latent. Specifically, to evaluate overall model convergence across all timesteps, we employ Monte Carlo simulation to estimate the expected eigenvalues, $\bar{s}_l$, of the ground-truth cross-correlation matrix, under the widely adopted Logit-normal timestep sampling strategy [9] (see Fig. 2d). By comparing these ground-truth mode strengths $\bar{s}_l$ to those learned by the model, we quantitatively evaluate convergence rates for different correlation modes (see Fig. 2e). This analysis leads to two key findings: (i) **Few-mode biasing can result in all absolute mode strengths, $|\bar{s}_l|$, exceeding those of the baseline**. This phenomenon arises from the non-monotonic relationship between $|\bar{s}_l|$ and λ_l (see Eq. 1): as λ_l decreases, the corresponding $|\bar{s}_l|$ first decreases and then increases. (ii) **Modes with larger absolute strengths tend to converge faster**, extending the insights of [21] to diffusion training. Fig. 2e shows that the covariance-penalized version has lower errors in the learned strength for each mode than the baseline.

Taken together, these results suggest that **few-mode biased latent spaces can accelerate diffusion model convergence by amplifying the absolute mode strengths in the output-input cross-correlation matrix.**

3.3 Latent Masked Reconstruction

To reliably promote a few-mode biased VAE latent space, we introduce Latent Masked Reconstruction (LMR) during training. Unlike the covariance penalization method in Sec. 3.1, **LMR avoids explicit eigenvalue decomposition, reducing computational overhead and improving numerical stability**. The intuition behind LMR is that, to minimize the influence of random masking on reconstruction, the VAE encoder must concentrate essential information into a few modes, thus promoting FMB.

Specifically, LMR randomly replaces latent vectors with mask tokens across spatio-temporal dimensions, feeding the masked latents to the decoder for reconstruction, as shown in Fig. 3(a). The masking ratio is randomly selected from $\{0, 0.25, 0.5, 0.75\}$, where 0 indicates no masking. Let $\mathcal{M} \in \mathbb{R}^{T\times H\times W\times 1}$ denote the spatio-temporal mask and $\mathbf{m} \in \mathbb{R}^{1\times 1\times 1\times C}$ the mask token. With the broadcast mechanism, this process can be formulated as:

$$\mathcal{L}_{LMR} = \mathcal{D}(\mathbf{x}, Dec(\mathbf{z} \odot \mathcal{M} + (1-\mathcal{M}) \odot \mathbf{m})), \tag{2}$$

here $\mathbf{x}$ denotes the video, $\mathbf{z}$ its corresponding latent code, $Dec(\cdot)$ the VAE Decoder, and $\odot$ the Hadamard product. The video reconstructed via LMR is then used to compute the distance metrics, including the L1, LPIPS, and GAN losses.

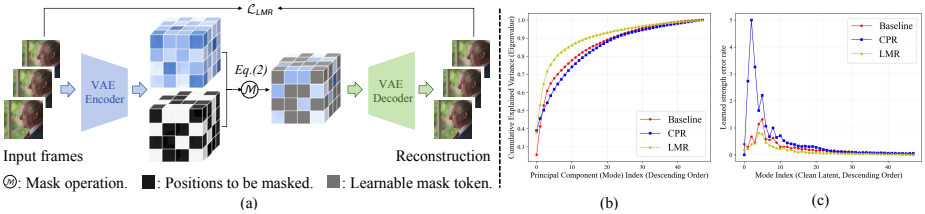


Fig. 3: (a) LMR introduces a few-mode bias by reconstructing videos using spatio-temporally masked latents. We omit the channel dimension for simplicity. (b) Cumulative explained variance comparison between CPR and LMR. (c) Eigenspectrum shaping effects comparison of LMR and CPR.

Beyond facilitating FMB, an equally important role of LMR is to enhance decoder robustness. Video VAEs in T2V models often use a very small KL loss weight [24] to produce latents easier to model. While this sufficiently constrains the posterior mean, it can cause posterior variance collapse. Wan 2.1 VAE, for example, uses a KL loss weight of 1e−6, yielding a posterior variance as low as 1e−7. In contrast, the typical reconstruction error in diffusion models is at the 10^{-1} magnitude, which makes it difficult for the VAE decoder to generate reasonable videos from highly noisy diffused samples. The masking operation in LMR, which replaces latent features with mask tokens, can also be interpreted as a kind of noise injection. Consequently, LMR enhances both the mode distribution and decoder robustness. We further discuss the influence of decoder robustness in Sec. 5.3 and Tab. 2.

To further justify our design choices, we compare LMR with other regularizations. Unlike the image tokenizer MAEtok [4], which performs random masking in pixel space, LMR masks latents directly and is thus compatible with latent regularizers such as LCR. Compared to the Channel-wise Progressive Reconstruction (CPR) in DC-AE-1.5 [6], which randomly retains the first C' channels for reconstruction, we find that random spatial-temporal masking more effectively encourages few-mode bias. Reproducing CPR under the same masking ratios and probabilities shows that it mainly amplifies the primary eigenvalue but distributes the remaining ones more uniformly than the baseline (see Fig. 3(b)), leading to slower convergence on most modes (Fig. 3(c)). This may explain why CPR requires **modifying the diffusion training** for progressive channel-wise denoising, whereas LMR remains effective as a plug-and-play regularization without altering the subsequent diffusion training.

4 Spatio-Temporal Frequency Spectrum Shaping

Complementary to the channel eigenspectrum, we further analyze the spatio-temporal frequency spectrum. Existing work [23] has shown that biasing the spatial frequency spectrum toward low frequencies improves diffusability, and we extend this analysis to the spatio-temporal frequency spectrum, identify

commonalities and limitations in existing methods, and propose Local Correlation Regularization, a regularizer founded on the statistical link between low-frequency energy and local correlation.

4.1 Preliminary

To analyze the frequency characteristics of latents, a widely adopted approach is first apply a frequency transform to map the latents into the frequency domain, and then compute the energy at each frequency to obtain the power spectral density (PSD). Existing work [23] employs a 2D discrete cosine transform (DCT) to analyze the spatial frequency spectrum, averaging across the temporal and channel dimensions. Following this procedure, we instead adopt a 3D DCT to analyze the spatio-temporal frequency spectrum, averaging only across the channel dimension. For visualization, we linearize the 3D frequency grid via zigzag ordering, bin consecutive frequencies, sum the energy within each bin, and normalize the spectral energy, as depicted in Fig. 4(a).

SER [23] observes that **a spatial spectrum biased toward low frequencies (or, equivalently, the suppression of high-frequency components) correlates with improved diffusion training.** Intuitively, high-SNR low-frequency components facilitate the recovery of low-SNR high-frequency details during denoising, simplifying optimization. We find that this low-frequency biasing principle likewise applies to the spatio-temporal frequency spectrum.

4.2 From PSD to Local Correlation

Existing works encourage low-frequency bias by enforcing spatial scale equivariance in latents [12,23]. Notably, we observe that foundation-model alignment [28] similarly increases low-frequency energy, although it is not claimed. This may partly explain its effectiveness. However, neither approach addresses temporal frequencies, failing to adequately bias the spatiotemporal spectrum toward low frequencies, as shown in Fig. 4(a). Moreover, both incur non-trivial computational overhead. Thus, we seek a new regularizer that operates directly on the spatio-temporal spectrum while remaining computationally efficient.

Our key insight is that **low-frequency energy is largely governed by the similarity of latent vectors at neighboring spatio-temporal positions.** When adjacent positions have highly similar channel vectors, the latent becomes smoother, concentrating more power in low frequencies. This follows the Wiener–Khinchin theorem [13], which states that the power spectral density (PSD) and the autocorrelation function form a Fourier pair. Consequently, boosting small-lag correlations increases low-frequency power. Guided by this observation, we propose a simple yet effective Local Correlation Regularization (LCR), which explicitly enhances similarities within small spatio-temporal patches, is computationally efficient, and naturally addresses the temporal dimension. A discussion of the relationships between the autocorrelation function, low-frequency energy, and local correlation, along with a complexity analysis of LCR, is presented in supplementary Sec. 2.

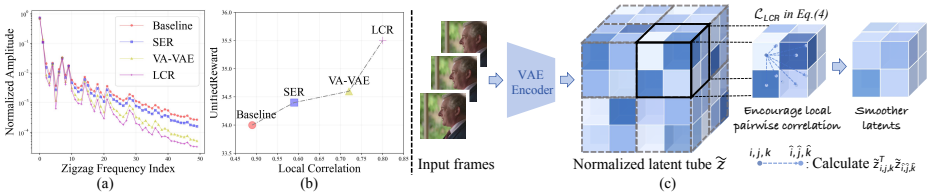


Fig. 4: (a) Comparison of PSD curves. Low zigzag frequency indexes correspond to low frequencies. LCR exhibits the strongest but controlled low-frequency bias. (b) Comparison of UnifiedReward over local correlation value. Steeper PSDs correspond to larger local correlation, and result in better video generation quality. (c) LCR introduces a low-frequency bias by promoting pairwise correlations within each spatio-temporal local patch in the normalized latents. We omit the channel dimension for simplicity.

Concretely, given a per-channel standardized latent, we partition it spatio-temporally into non-overlapping small patches and measure the average pairwise correlation within each patch. We denote the standardized latent as $\tilde{\mathbf{z}} \in \mathbb{R}^{T\times H\times W\times C}$, where T, H, W, C denote temporal length, height, width, and number of channels. During training, the mean and variance of each channel are computed over (B, T, H, W) to match the diffusion backbone's preprocessing, where B is the batch size. Using (i, j, k) to index time, height and width, and letting p denote the set of spatio-temporal coordinates within a patch, the average local correlation in patch p is

$$\tilde{R}(p) \;= \mathbb{E}_{\substack{(i,j,k),(\hat{i},\hat{j},\hat{k})\in p,\\ (i,j,k)\neq(\hat{i},\hat{j},\hat{k})}} \big[\, \tilde{\mathbf{z}}^{\top}_{i,j,k}\, \tilde{\mathbf{z}}_{\hat{i},\hat{j},\hat{k}} \,\big]. \tag{3}$$

We define Local Correlation Regularization (LCR) as a hinge loss on the patch-averaged $\tilde{R}(p)$:

$$\mathcal{L}_{LCR} \;=\; \mathrm{ReLU}\Big(\alpha - \mathbb{E}_p\big[\tilde{R}(p)\big]\Big)\,, \tag{4}$$

where α is a saturation threshold that prevents over-smoothing; together with reconstruction losses, it keeps the bias **controlled**, as excessive bias can hurt generation (see supplementary Sec. 6). In practice, we set the patch size to 2, dividing spatial patches in the first frame of the latent, and spatio-temporal patches in the remaining frames. We apply LCR at the **batch level** rather than per-sample, which preserves the flexibility of local correlations, accommodating videos with high dynamics and complex details. We further observe that replacing the dot product in Eq. 3 with cosine similarity preserves effective autocorrelation optimization while better balancing local correlations between the first and subsequent frames. Accordingly, this choice yields the Pearson correlation coefficient within each local patch and is adopted in our implementation. LCR is also illustrated in Fig. 4(c).

In Fig. 4(b), we observe that for 64-channel VAEs, **local correlation both reflects the steepness of the PSD and correlates positively with the reward model score**. UnifiedReward [25] is employed as the reward model

and is found to generally align with human preferences. By effectively increasing low-frequency components (equivalent to suppressing high-frequency components under fixed latent energy), LCR improves generation quality.

5 Experiments

5.1 Experiment Setup

Architecture. We adopt a ResNet-based 3D VAE with 3D causal convolutions (similar to CogVideoX [27]) and extend DC-AE's [5] residual autoencoding to 3D for up/down sampling. We find that the architecture primarily impacts reconstruction quality, with limited effects on statistics like frequency or eigenspectrum. Default settings include 16× spatial and 4× temporal downsampling, a diffusion patch size of $(1 \times 1 \times 1)$, and a latent dimension of 48. This follows recent high-compression VAEs [24], which fold the usual 2×2 patchification of 16-channel, 8× spatial-downsampled latents into the encoder. For existing video VAEs [11, 17, 24, 27], the spatial patch size is set to 2 for those with 8× spatial downsampling, and 1 otherwise.

Data. We mainly use open-source datasets to train VAEs: laion2B-en [20] and COYO [3] for images, and webvid [1] and panda70M [8] for videos. For video generation, we train with an internal collection of movie and television clips, reserving 1,348 videos as a disjoint validation set, referred to as **MovieValid**. It covers diverse content, including landscapes, humans, animals, plants, architecture, artworks, black-and-white films, and cartoons.

Evaluation. Videos are generated with prompts from VBench (GPT-enhanced version) [29], MovieGenBench [19], and MovieValid. We assess video quality using reward models and FVD. Given the well-documented shortcomings of FVD for T2V evaluation, including its preference for temporal inconsistency and artifacts [10], as well as its inability to reflect multi-faceted quality [29], we prioritize reward models such as UnifiedReward [25] and VideoAlign [16], which achieve high human agreement (70–80%) [25]. UnifiedReward [25] scores are reported as 'UR', and VideoAlign [16] scores as 'VAR'. Since VideoAlign's raw rewards are unbounded logits, we apply the sigmoid function to each evaluation aspect and average the results to obtain a score in the 0–100 range.

Training Details. We train our VAE on a mixed image-video dataset with a 6:7 ratio. The loss function combines $\mathcal{L}_{LMR}$, KL, LPIPS, GAN, and $\mathcal{L}_{LCR}$ with weights 1, 5e−4, 1, 1, and 0.02, respectively. For $\mathcal{L}_{LCR}$, we use the dynamic gradient-based weighting scheme from [28] with threshold $\alpha = 0.75$. For $\mathcal{L}_{LMR}$, mask ratios $\{0, 0.25, 0.5, 0.75\}$ are assigned selection probabilities $\{0.7, 0.1, 0.1, 0.1\}$. Prior to input to the diffusion backbone, VAE latents are normalized using the mean and standard deviation from MovieValid. Following SD3 [9], we train diffusion models under the flow-matching and velocity prediction setting, sampling timesteps from a logit-normal distribution.

Table 1: Generation comparison across various video VAEs. 'CPR' refers to the spatial and temporal downsampling factors of the VAEs, and 'Chn' is short for channel number. The best generation results are **bolded**, and the second-best are underlined.

Method	CPR, Chn	VBench			MovieGenBench			MovieValid		
		UR↑	VAR↑	FVD↓	UR↑	VAR↑	FVD↓	UR↑	VAR↑	FVD↓
		MMDiT 1.3B, $17 \times 256 \times 256$								
WF-VAE [15]	$(8 \times 8 \times 4)$, 16	26.5	23.7	978	25.5	20.8	859	30.6	27.9	514
IV-VAE [26]	$(8 \times 8 \times 4)$, 16	34.0	27.8	1066	28.8	22.0	784	37.7	30.9	541
Hunyuan VAE [11]	$(8 \times 8 \times 4)$, 16	29.7	25.9	1354	26.3	21.1	1169	32.1	28.9	610
CogvideoX VAE [27]	$(8 \times 8 \times 4)$, 16	30.4	26.6	1040	26.7	21.4	826	32.6	29.9	543
Wan 2.1 VAE [24]	$(8 \times 8 \times 4)$, 16	33.3	27.2	986	28.5	22.1	765	36.5	30.8	517
StepVideo VAE [17]	$(16 \times 16 \times 8)$, 64	29.7	25.6	1004	26.3	20.6	742	31.8	28.8	**418**
Wan 2.2 VAE [24]	$(16 \times 16 \times 4)$, 48	33.6	**27.9**	**873**	28.4	**22.8**	716	36.7	30.8	432
Baseline	$(16 \times 16 \times 4)$, 48	32.4	27.1	909	27.8	21.6	798	36.3	30.9	455
SSVAE (Ours)	$(16 \times 16 \times 4)$, 48	**34.7**	27.7	983	**30.2**	22.6	**703**	**39.1**	**31.6**	511
		MMDiT 1.3B, $17 \times 512 \times 512$								
WF-VAE	$(8 \times 8 \times 4)$, 16	36.3	27.9	1352	29.2	22.2	894	39.8	31.3	667
IV-VAE	$(8 \times 8 \times 4)$, 16	41.8	29.8	997	34.1	23.4	776	46.4	33.2	535
Hunyuan VAE	$(8 \times 8 \times 4)$, 16	36.9	28.1	1166	31.9	23.5	892	41.4	31.5	663
CogvideoX VAE	$(8 \times 8 \times 4)$, 16	39.0	29.1	971	31.2	23.4	675	41.3	32.5	452
Wan 2.1 VAE	$(8 \times 8 \times 4)$, 16	41.6	29.8	978	33.6	23.7	791	46.6	33.9	500
StepVideo VAE	$(16 \times 16 \times 8)$, 64	35.6	28.0	1190	29.4	21.6	866	37.0	30.9	562
Wan 2.2 VAE	$(16 \times 16 \times 4)$, 48	42.8	30.6	1019	33.4	23.7	733	47.9	34.0	502
Baseline	$(16 \times 16 \times 4)$, 48	40.9	29.1	970	32.9	23.3	681	46.5	33.9	504
SSVAE (Ours)	$(16 \times 16 \times 4)$, 48	**45.9**	**30.7**	**828**	**35.7**	**24.3**	**605**	**50.8**	**34.8**	**403**
		MMDiT 1.3B, $81 \times 256 \times 256$								
Wan 2.2 VAE	$(16 \times 16 \times 4)$, 48	35.9	30.8	584	29.8	23.1	600	40.1	35.2	**247**
SSVAE (Ours)	$(16 \times 16 \times 4)$, 48	**37.8**	**33.2**	**560**	**31.6**	**24.5**	**484**	**40.8**	**36.8**	272
		Wan 1.3B, $17 \times 512 \times 512$								
Wan 2.2 VAE	$(16 \times 16 \times 4)$, 48	37.7	27.3	1124	30.6	23.2	736	42.3	31.6	601
SSVAE (Ours)	$(16 \times 16 \times 4)$, 48	**45.5**	**29.5**	**823**	**34.5**	**24.9**	**691**	**48.6**	**33.5**	**466**

5.2 Comparison with Existing Video VAEs

We train 48-channel VAEs with LMR and LCR ('SSVAE (Ours)') and without them ('Baseline'). SSVAE is initially trained for 150k steps at 256×256 resolution. Subsequently, it is fine-tuned for 50k steps at 512×512 resolution with a fixed encoder and both LCR and LMR disabled to enhance high-resolution reconstruction. Generation for $17 \times 256 \times 256$ and $81 \times 256 \times 256$ settings uses a 1.3B MMDiT-like [9] diffusion model trained for 50k steps on images, followed by a further 50k steps on 17-frame or 81-frame videos, respectively. For $17 \times 512 \times 512$, the model is initialized from the $17 \times 256 \times 256$ checkpoint and further trained for 100k steps on 512×512 videos. A total of 200k generation steps matches or exceeds the training budget used in prior works [15, 26].

Generation Comparison across Resolutions and Backbones. Spectra of 48-channel VAEs are shown in Fig. 1, while Tab. 1 compares the generation quality of SSVAE to existing SOTA video VAEs. Key observations include: (i)

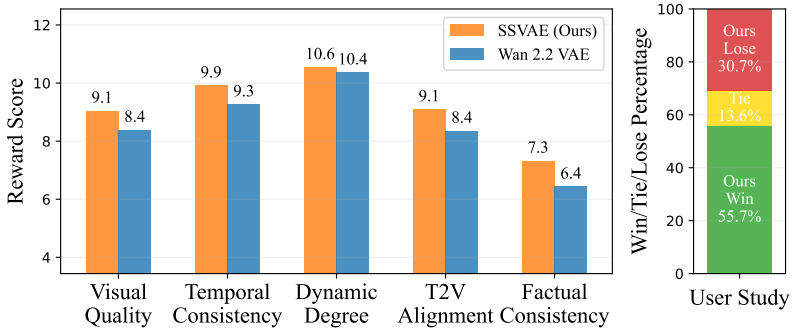


Fig. 5: Human preference and UnifiedReward dimension breakdown.

SSVAE improves diffusability by a large margin. Visual reward models align more closely with human judgments [25], and SSVAE achieves the highest reward scores across almost all settings and datasets. For $17 \times 512 \times 512$ generation, it achieves a $3\times$ UR convergence speedup over the 48-channel baseline (Fig. 1). Although FVD is mixed at 256p, SSVAE improves FVD at 512p and on 81-frame VBench/MovieGenBench, and the overall reward-model and human-preference results support SSVAE. (ii) **Generation quality (diffusability) is highly correlated with spectral properties**. Comparing the 48-channel Baseline, Wan 2.2, and SSVAE in Fig. 1, we observe that video reward scores consistently improve as spectral bias is enhanced, highlighting the significance of spectrum biasing; (iii) **The advantages of SSVAE become more pronounced with longer denoising sequences**. Compared to the previous SOTA Wan 2.2 VAE, SSVAE yields an average improvement of 3.4 in UR and 1.4 in VAR across $17 \times 512 \times 512$ and $81 \times 256 \times 256$ generation tasks, and across all datasets. This may be primarily attributed to the ability of few-mode bias to accelerate convergence. (iv) **SSVAE is diffusion-backbone-agnostic**. It consistently improves generative quality regardless of the diffusion backbone used, including both MMDiT [9] and Wan [24].

Generation Performance across Model Scales. We further compare SSVAE with Wan 2.2 VAE across various diffusion backbone model sizes, including 0.75B, 1.3B, and 4B parameters. As shown in Fig. 1, our experiments on VBench at $17 \times 512 \times 512$ demonstrate that SSVAE with a 1.3B diffusion backbone achieves a higher UnifiedReward score than Wan 2.2 VAE with a 4B diffusion model, while using 67.5% fewer parameters. Fig. 6(a) shows the generation results for the 4B models, where our method achieves higher visual quality and better alignment with the prompts.

Human-Preference Validation. Fig. 5 details the "MMDiT 1.3B, $17 \times 512 \times 512$" setting from Tab. 1: SSVAE improves all five UnifiedReward dimensions and obtains a 55.7% win rate in a 1,200-pair user study using MovieGenBench prompts.

Table 2: Incremental component analysis on VBench and MovieValid.

Components	VBench			MovieValid		
	UR	VAR	FVD	UR	VAR	FVD
Baseline	36.3	26.8	1069	38.1	30.6	487
+①LCR	36.8	27.3	993	39.0	30.6	472
+①+②Covariance Penalize	37.4	27.6	980	39.9	30.9	467
+①+②+③Decoder Finetune	38.0	27.9	961	40.7	31.3	451
+①+④LMR	**39.1**	**28.2**	**958**	**41.4**	**31.7**	**450**
+①+④+③	38.9	28.0	960	41.3	31.6	450

Table 3: Comparison of frequency spectrum biasing techniques. 'LC' denotes local correlation computed on MovieValid.

Components	LC	UR	VAR	FVD
64-ch Baseline	0.49	34.0	28.4	553
+SER	0.59	34.4	28.5	**541**
+VA-VAE	0.73	34.6	28.6	546
+LCR	**0.80**	**35.5**	**29.2**	606
48-ch Baseline	0.67	38.1	30.6	487
+LCR	**0.80**	**39.0**	**30.6**	**472**

5.3 Ablation Studies

All VAEs in the ablation experiments are trained for 100k steps at 256×256 resolution, which we find sufficient for high-resolution generation, although it slightly compromises reconstruction quality. We then pre-train 1.3B MMDiT models on 512×512 images for 40k steps, followed by finetuning on $17 \times 512 \times 512$ videos for 30k steps. Performance is evaluated mostly on the MovieValid benchmark, as its data distribution matches the training set and its prompts are derived from real videos.

Incremental Component Analysis. We ablate the effects of low-frequency biasing, few-mode biasing, and decoder robustness using a 48-channel setting; results are shown in Tab. 2. '*Covariance Penalize*' refers to the few-mode-forcing trick in Sec. 3.1, where the sum of all eigenvalues except the largest three are minimized. '*Decoder Finetune*' adds a training stage where the encoder is fixed and a Gaussian noise with $5e-3$ variance is injected into the latents to finetune the decoder for improved robustness. We observe that: (i) Low-frequency bias, few-mode bias, and decoder finetuning complement each other in boosting generation quality; (ii) Simply applying decoder finetuning to a covariance-penalized VAE yields performance improvements. However, LMR achieves further gains by jointly promoting few-mode bias and enhancing decoder robustness through end-to-end training. Additional finetuning of the decoder after applying LMR offers no further benefit.

Frequency Biasing Ablation. We revisit existing frequency biasing techniques on 64-channel VAEs, including scale-equivariant methods [23] and foundation model alignment approaches [28]. We also validate the effectiveness of LCR on 48-channel VAEs. As shown in Tab. 3 and Fig. 4, we observe: (i) Local correlation positively correlates with generation quality: higher local correlation leads to better generation rewards; (ii) Low-frequency bias may explain the generative benefits of foundation model alignment. However, existing frequency biasing methods cannot sufficiently bias the spatio-temporal frequency spectrum, as they lack explicit mechanisms for handling the temporal dimension. In contrast, LCR efficiently introduces spatio-temporal low-frequency components and allows flexible control of their energy via the threshold α; (iii) The improvements from LCR

Fig. 6: (a) A comparison of $17 \times 512 \times 512$ video generation using 4B diffusion models trained with SSVAE and Wan 2.2 VAE. For simplicity, only the first and last frames are shown. Prompts are sampled from MovieValid, and artifacts are highlighted in blue. Our results demonstrate higher visual quality, fewer artifacts, and better alignment with the prompts. (b) Reconstruction comparison of SSVAE and Wan 2.2 VAE.

are less pronounced in low-channel VAEs compared to high-channel VAEs. We attribute this to low-channel VAEs naturally introducing more low-frequency components. As shown in Tab. 3, the local correlation of the 48-channel baseline (0.67) is higher than that of the 64-channel baseline (0.49).

Reconstruction Comparison and Ablation. We compare SSVAE and other VAEs at $17 \times 512 \times 512$ in Tab. 4. While SSVAE's reconstruction fidelity is not the highest, it surpasses StepVideo and CogVideoX. DAVIS results in Tab. 6 show that SSVAE maintains comparable reconstruction quality in high-frequency, large-motion scenarios, where its PSNR is only 0.15 dB below the 48-channel baseline with matched LPIPS. We also train downstream **I2V** models with channel-concat conditioning and find that SSVAE remains close to Wan 2.2 VAE in **first-frame PSNR (34.95 vs. 35.14, gap 0.19)**, suggesting that diffusion noise reduces the practical reconstruction gap between VAEs. Visualization results in Fig. 6(b) confirm that its reconstruction remains **visually faithful, avoiding perceptible semantic distortions in details** such as the cup's text (Cafe) and clothing patterns.

We further ablate the impact of LCR and LMR on reconstruction in Tab. 5 under identical training regimes and observe: (i) **Smooth latents do not necessitate RGB detail loss.** The LCR, which induces latent smoothness, even slightly improves reconstruction. This aligns with existing 16-channel models like IV-VAE and Wan 2.1, which exhibit similarly steep PSDs (Supp. Fig. A.3) yet maintain high fidelity. (ii) LMR incurs a mild quality drop (0.23 PSNR vs. 48-ch

Table 4: Comparison of reconstruction results on MovieValid at a resolution of $17 \times 512 \times 512$.

Method	PSNR↑	SSIM↑	LPIPS↓
Wan 2.1 VAE	38.57	0.97	0.0231
Wan 2.2 VAE	39.30	0.97	0.0206
WF-VAE	38.61	0.96	0.0352
IV-VAE	40.29	0.97	0.0220
Hunyuan VAE	40.31	0.97	0.0209
CogvideoX VAE	36.65	0.96	0.0398
StepVideo VAE	36.78	0.95	0.0435
48-ch Baseline	38.07	0.96	0.0320
SSVAE (Ours)	37.84	0.96	0.0325

Table 5: Impact of LCR and LMR on MovieValid reconstruction quality at a resolution of $17 \times 512 \times 512$. All VAEs are trained under the same regime.

Method	Param.	PSNR↑	SSIM↑	LPIPS↓
Wan 2.2 VAE-retrain	705M	38.41	0.96	0.0312
48-ch Baseline	315M	38.07	0.96	0.0320
+LCR	315M	38.29	0.96	0.0315
+LCR+LMR	315M	37.84	0.96	0.0325

Table 6: DAVIS reconstruction comparison at 512p.

Method	PSNR↑	LPIPS↓
Wan 2.2 VAE	30.90	0.056
Wan 2.2 VAE-retrain	30.14	0.087
48-ch Baseline	29.84	0.091
SSVAE (Ours)	29.69	0.091

baseline), which likely stems from a trade-off between decoding precision and robustness. (iii) The disparity with the official Wan 2.2 VAE likely arises from its greater parameter count and more extensive training (scale and data). For context, the comparable Seaweed [22] underwent 1.3M training steps, whereas our model used 200k. Notably, the gap narrows significantly when comparing SSVAE to the re-trained Wan 2.2 baseline. Although extended training also improves SSVAE's reconstruction, we observed negligible gains in generation quality and thus prioritized computational efficiency.

In the supplementary material, experiments on public datasets confirm that our gains are independent of diffusion training data. Additionally, we provide LCR and LMR hyperparameter and computational overhead analyses, CPR/LMR residual diagnostics, the relationship between our spectral regularizations and spatial representation learning, extra visualizations, and the UnifiedReward-Thinking evaluation prompt.

6 Conclusion

We analyze video VAE latent spaces and identify two critical spectral properties for diffusion training: few-mode bias in the channel eigenspectrum and low-frequency bias in the spatio-temporal spectrum. We introduce two lightweight regularizers to induce these properties, and experiments show that our approach accelerates convergence and improves video generation quality.

References

1. Bain, M., Nagrani, A., Varol, G., Zisserman, A.: Frozen in time: A joint video and image encoder for end-to-end retrieval. In: IEEE International Conference on Computer Vision (2021)
2. Bonnaire, T., Urfin, R., Biroli, G., Mézard, M.: Why diffusion models don't memorize: The role of implicit dynamical regularization in training. arXiv preprint arXiv:2505.17638 (2025)
3. Byeon, M., Park, B., Kim, H., Lee, S., Baek, W., Kim, S.: Coyo-700m: Image-text pair dataset. `https://github.com/kakaobrain/coyo-dataset` (2022), accessed: June 26, 2026
4. Chen, H., Han, Y., Chen, F., Li, X., Wang, Y., Wang, J., Wang, Z., Liu, Z., Zou, D., Raj, B.: Masked autoencoders are effective tokenizers for diffusion models. In: Forty-second International Conference on Machine Learning (2025)
5. Chen, J., Cai, H., Chen, J., Xie, E., Yang, S., Tang, H., Li, M., Han, S.: Deep compression autoencoder for efficient high-resolution diffusion models. In: The Thirteenth International Conference on Learning Representations (2024)
6. Chen, J., Zou, D., He, W., Chen, J., Xie, E., Han, S., Cai, H.: Dc-ae 1.5: Accelerating diffusion model convergence with structured latent space. arXiv preprint arXiv:2508.00413 (2025)
7. Chen, L., Li, Z., Lin, B., Zhu, B., Wang, Q., Yuan, S., Zhou, X., Cheng, X., Yuan, L.: Od-vae: An omni-dimensional video compressor for improving latent video diffusion model. arXiv preprint arXiv:2409.01199 (2024)
8. Chen, T.S., Siarohin, A., Menapace, W., Deyneka, E., Chao, H.w., Jeon, B.E., Fang, Y., Lee, H.Y., Ren, J., Yang, M.H., Tulyakov, S.: Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)
9. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: Forty-first international conference on machine learning (2024)
10. Ge, S., Mahapatra, A., Parmar, G., Zhu, J.Y., Huang, J.B.: On the content bias in fréchet video distance. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7277–7288 (2024)
11. Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., Wu, K., Lin, Q., Yuan, J., Long, Y., Wang, A., Wang, A., Li, C., Huang, D., Yang, F., Tan, H., Wang, H., Song, J., Bai, J., Wu, J., Xue, J., Wang, J., Wang, K., Liu, M., Li, P., Li, S., Wang, W., Yu, W., Deng, X., Li, Y., Chen, Y., Cui, Y., Peng, Y., Yu, Z., He, Z., Xu, Z., Zhou, Z., Xu, Z., Tao, Y., Lu, Q., Liu, S., Zhou, D., Wang, H., Yang, Y., Wang, D., Liu, Y., Jiang, J., Zhong, C.: Hunyuanvideo: A systematic framework for large video generative models. arXiv preprint arXiv:2412.03603 (2025)
12. Kouzelis, T., Kakogeorgiou, I., Gidaris, S., Komodakis, N.: EQ-VAE: Equivariance regularized latent space for improved generative image modeling. In: Forty-second International Conference on Machine Learning (2025)
13. Kschischang, F.R.: The wiener-khinchin theorem. The Edward S. Rogers Sr. Department of Electrical and Computer Engineering University of Toronto (2017)
14. Leng, X., Singh, J., Hou, Y., Xing, Z., Xie, S., Zheng, L.: Repa-e: Unlocking vae for end-to-end tuning with latent diffusion transformers. arXiv preprint arXiv:2504.10483 (2025)

15. Li, Z., Lin, B., Ye, Y., Chen, L., Cheng, X., Yuan, S., Yuan, L.: Wf-vae: Enhancing video vae by wavelet-driven energy flow for latent video diffusion model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 17778–17788 (June 2025)
16. Liu, J., Liu, G., Liang, J., Yuan, Z., Liu, X., Zheng, M., Wu, X., Wang, Q., Qin, W., Xia, M., et al.: Improving video generation with human feedback. arXiv preprint arXiv:2501.13918 (2025)
17. Ma, G., Huang, H., Yan, K., Chen, L., Duan, N., Yin, S., Wan, C., Ming, R., Song, X., Chen, X., et al.: Step-video-t2v technical report: The practice, challenges, and future of video foundation model. arXiv preprint arXiv:2502.10248 (2025)
18. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., HAZIZA, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning robust visual features without supervision. Transactions on Machine Learning Research (2024)
19. Polyak, A., Zohar, A., Brown, A., Tjandra, A., Sinha, A., Lee, A., Vyas, A., Shi, B., Ma, C.Y., Chuang, C.Y., et al.: Movie gen: A cast of media foundation models. arXiv preprint arXiv:2410.13720 (2024)
20. Project, L.: Laion dataset. `https://laion.ai/projects/` (2022), accessed: June 26, 2026
21. Saxe, A.M., McClelland, J.L., Ganguli, S.: Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. arXiv preprint arXiv:1312.6120 (2013)
22. Seaweed Team, Yang, C., Lin, Z., Zhao, Y., Lin, S., Ma, Z., Guo, H., Chen, H., Qi, L., Wang, S., et al.: Seaweed-7b: Cost-effective training of video generation foundation model. arXiv preprint arXiv:2504.08685 (2025)
23. Skorokhodov, I., Girish, S., Hu, B., Menapace, W., Li, Y., Abdal, R., Tulyakov, S., Siarohin, A.: Improving the diffusability of autoencoders. In: Forty-second International Conference on Machine Learning (2025)
24. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
25. Wang, Y., Li, Z., Zang, Y., Wang, C., Lu, Q., Jin, C., Wang, J.: Unified multimodal chain-of-thought reward model through reinforcement fine-tuning. arXiv preprint arXiv:2505.03318 (2025)
26. Wu, P., Zhu, K., Liu, Y., Zhao, L., Zhai, W., Cao, Y., Zha, Z.J.: Improved video vae for latent video diffusion model. In: Proceedings of the Computer Vision and Pattern Recognition Conference (2024)
27. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer. In: The Thirteenth International Conference on Learning Representations (2024)
28. Yao, J., Yang, B., Wang, X.: Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 15703–15712 (2025)
29. Zhang, F., Tian, S., Huang, Z., Qiao, Y., Liu, Z.: Evaluation agent: Efficient and promptable evaluation framework for visual generative models. arXiv preprint arXiv:2412.09645 (2024)

30. Zhao, S., Zhang, Y., Cun, X., Yang, S., Niu, M., Li, X., Hu, W., Shan, Y.: Cv-vae: A compatible video vae for latent generative video models. In: Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., Zhang, C. (eds.) Advances in Neural Information Processing Systems. vol. 37, pp. 12847–12871. Curran Associates, Inc. (2024)