

VIDEORAE: TAMING VIDEO FOUNDATION MODELS FOR GENERATIVE MODELING VIA REPRESENTATION AUTOENCODERS

Zhihao Xie^{1,2,*} Junfeng Wu^{2,*} Xinting Hu⁴ Junchao Huang^{1,3} Li Jiang^{1,3,†}

¹ The Chinese University of Hong Kong, Shenzhen

² Huazhong University of Science and Technology ³ Shenzhen Loop Area Institute

⁴ University of Science and Technology of China

ABSTRACT

The rapid advancement of video generative modeling has been largely driven by diffusion and autoregressive models operating in the latent spaces of 3D Variational Autoencoders (3D-VAEs). However, conventional 3D-VAEs are mainly optimized for pixel-level reconstruction, and this objective often limits the semantic and spatio-temporal structure captured by their latent spaces, constraining downstream synthesis quality. Concurrently, Video Foundation Models (VFM) such as V-JEPA 2 and VideoMAEv2 have demonstrated exceptional semantic understanding capabilities. Yet, whether such frozen video foundation representations can be transformed into compact, reconstruction-capable, and generation-friendly video latents remains largely unexplored. In this work, we answer this question affirmatively with VideoRAE. VideoRAE leverages multi-scale hierarchical features from a frozen video foundation encoder and employs a lightweight 1D self-attention projector to compress them into a highly compact latent space. The resulting latents support both continuous representations for Diffusion Transformers (DiTs) and discrete representations for autoregressive models via multi-codebook high-dimensional quantization. During decoding, VideoRAE incorporates a local-and-global representation alignment objective with the frozen VFM teacher, which improves semantic preservation and enables training without KL regularization. Comprehensive experiments show that VideoRAE achieves strong reconstruction in both continuous and discrete regimes. On UCF-101, it achieves state-of-the-art class-to-video gFVDs of 40 and 93 with AR and DiT generators, respectively, while converging approximately 5x faster than competing autoencoder baselines. In a controlled 2B-scale text-to-video experiment, replacing LTX-VAE with VideoRAE leads to faster convergence and consistently better VBench performance. These results validate frozen VFM representations as versatile and generation-friendly video latents. The model and code will be [released](#).

1 INTRODUCTION

The field of generative modeling has witnessed a paradigm shift, transitioning from pixel-space generation (Van Den Oord et al., 2016; Goodfellow et al., 2014) to latent-space generation (Rombach et al., 2022; Sun et al., 2024; Kingma & Welling, 2014; Van Den Oord et al., 2017). For video synthesis, both Variational Autoencoders (VAEs) (Cheng & Yuan, 2025; Wan et al., 2025; Yang et al., 2025) and discrete tokenizers (Wang et al., 2024b; Hong et al., 2022) typically employ 3D visual encoders to compress raw video clips into compact spatio-temporal tokens. This approach effectively reduces computational complexity. However, existing video tokenizers are still predominantly optimized for pixel-level reconstruction, which can underemphasize semantic and long-range spatio-temporal structure. Driven by pixel-level Mean Squared Error and adversarial losses, the latent spaces of these visual tokenizers are biased towards optimizing local appearance, rather than

*Equal contribution.

†Corresponding author.

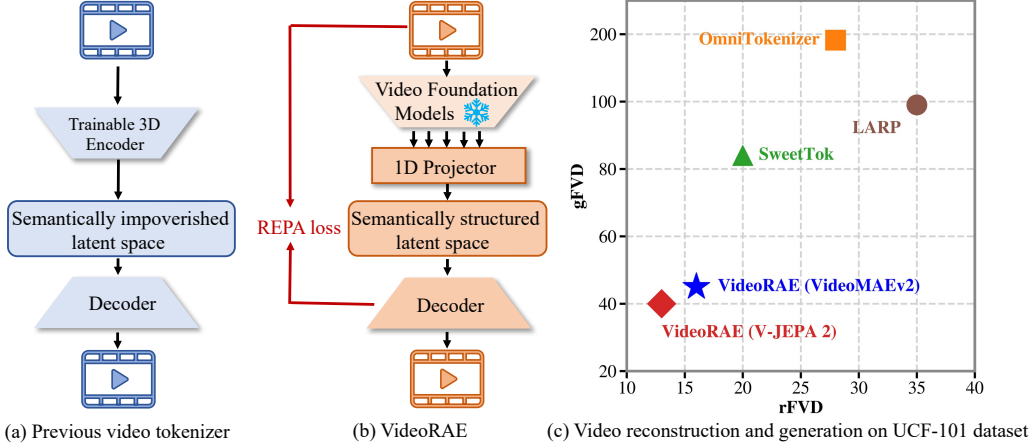


Figure 1: Conceptual comparison between traditional video tokenizers and VideoRAE. Traditional 3D-VAEs are trained from scratch and driven solely by pixel-level MSE and adversarial losses. In contrast, VideoRAE directly utilizes a frozen video foundation model as its core feature extractor. VideoRAE provides a unified and semantically rich latent space that perfectly supports both continuous and discrete generative paradigms, achieving high-fidelity reconstruction and high-quality generation.

capturing high-level semantics or physical structures. Consequently, downstream generative models are burdened with the heavy lifting of learning complex spatio-temporal dynamics and global semantics, increasing both the difficulty and data requirements of training.

Concurrently, visual representation learning has made substantial progress. Video Foundation Models, represented by V-JEPA 2 (Assran et al., 2025) and VideoMAEv2 (Wang et al., 2023), have shown strong capabilities in extracting structured and predictive spatio-temporal features. In the image domain, recent representation autoencoders such as RAE (Zheng et al., 2025b) and VFM-Tok (Zheng et al., 2025a) suggest that pretrained visual representations can be converted into decodable and generation-friendly latent spaces. However, extending this idea to video is nontrivial. Videos contain richer temporal dynamics, stronger redundancy, and stricter compression requirements than images, making it unclear whether frozen VFM features can be directly used as compact and generation-compatible video latents. Existing tokenizer studies for video generation (Guo et al., 2025; Zhang et al., 2025; Xiong et al., 2026) have mainly incorporated semantic alignment as auxiliary supervision within conventional pixel-driven VAE frameworks, while largely retaining their original encoder architectures and latent construction mechanisms. Therefore, whether frozen VFM representations can be transformed into compact latent spaces that support both high-fidelity reconstruction and downstream video generation remains underexplored.

To address this question, we introduce VideoRAE, a unified video autoencoding framework that uses a frozen VFM as its encoder. VideoRAE aggregates hierarchical VFM features to capture both coarse semantics and fine-grained spatio-temporal dynamics, and compresses these redundant representations with a lightweight 1D self-attention projector to meet the stringent compression requirements of video generation. During decoding, we further apply a local-and-global representation alignment objective with the frozen VFM teacher, which encourages the decoder to preserve the semantic structure of the foundation representation and enables stable training without KL regularization. The same framework supports both major latent generative paradigms: continuous latents for diffusion transformers, and discrete tokens for autoregressive models via a multi-codebook high-dimensional quantization strategy that preserves the capacity of VFM representations while making them autoregressively tractable.

We comprehensively evaluate VideoRAE across both continuous and discrete latent regimes, covering video reconstruction and downstream generative modeling. On UCF-101 and TokenBench, VideoRAE achieves state-of-the-art discrete reconstruction and highly competitive continuous reconstruction, demonstrating that frozen VFM representations can remain reconstruction-capable after strong compression. For class-to-video generation on UCF-101, its discrete AR and continuous DiT variants achieve state-of-the-art gFVDs of 40 and 93, respectively. Notably, the AR variant outperforms prior discrete tokenizers while using fewer tokens and a smaller generator, and both AR

and DiT generators converge approximately $5\times$ faster than their respective autoencoder baselines. We further conduct a controlled 2B-scale text-to-video replacement study, where VideoRAE leads to faster convergence than the LTX-VAE baseline under the same training framework and comparable settings. Together, these results demonstrate that the advantages of VideoRAE hold across latent formulations, generative paradigms, and task complexity.

Our main contributions are summarized as follows:

- We diverge from the conventional paradigm that restricts frozen Video Foundation Models (VFMs) primarily to understanding tasks. Instead, we systematically demonstrate for the first time that frozen VFMs can be unleashed to serve as robust, directly applicable autoencoder foundations for high-fidelity video generation.
- We propose VideoRAE, featuring a semantic-driven 1D projector and a local-and-global representation alignment objective tailored to frozen VFM features. It unifies continuous KL-free latents and discrete multi-codebook latents within a single framework.
- Extensive experiments demonstrate that VideoRAE not only matches traditional VAEs in reconstruction metrics but also provides a structured semantic latent space that significantly accelerates the convergence of downstream generative models and improves generation quality. Our work confirms the immense feasibility and potential of visual foundation model representations in video generative tasks.

2 RELATED WORKS

2.1 VIDEO AUTOENCODERS AND TOKENIZERS

Video variational autoencoders (VAEs) and tokenizers are crucial components for compressing high-dimensional video data into compact latent spaces, which is essential for realizing efficient generative modeling. In the realm of continuous spaces, VFRTok (Zhong et al., 2026) attempts to enhance model robustness through variable frame rate training, VidTwin (Wang et al., 2025) improves reconstruction quality by decoupling structure and dynamics, and LeanVAE (Cheng & Yuan, 2025) boosts training efficiency via lightweight network design and wavelet transforms. In the domain of discrete tokenization, LARP (Wang et al., 2024a) pioneers a 1D tokenizer architecture and elevates generation quality through an autoregressive prior, whereas SweetTok (Tan et al., 2025) mitigates quantization loss by utilizing a dual-stream spatio-temporal codebook. Despite their success, these models are typically trained entirely from scratch, relying heavily on a combination of pixel-level reconstruction and adversarial losses. Unlike these pixel-driven approaches, VideoRAE explores a semantic-driven compression paradigm by directly leveraging frozen VFMs, thereby providing a highly structured latent space that alleviates the learning burden on downstream generators.

2.2 VISUAL FOUNDATION MODELS

Visual Foundation Models (VFMs) (Oquab et al., 2023; Siméoni et al., 2025; Assran et al., 2025; Mur-Labadia et al., 2026; Wang et al., 2023; 2022) aim to learn general, transferable representations from large-scale, diverse data. In the image domain, self-supervised learning and language-supervised pre-training on image-text pairs have endowed VFMs with rich and semantically grounded representations. Extending to the video domain, the training of video foundation models is predominantly based on scalable self-supervised learning, leveraging the inherent spatio-temporal structures of videos. Notably, VideoMAEv2 (Wang et al., 2023) scales masked video autoencoding to billion-level parameters via a dual-masking strategy, forcing the model to develop robust spatio-temporal reasoning capabilities from extremely sparse inputs. Similarly, V-JEPA 2 (Assran et al., 2025) employs a joint-embedding predictive architecture. By having the target encoder predict the latent features of masked regions, V-JEPA 2 (Assran et al., 2025) deliberately discards low-level visual noise to primarily focus on high-level semantics and physical motion logic. While these VFMs exhibit exceptional understanding capabilities, their highly abstract nature has historically deterred their direct application in high-fidelity pixel reconstruction, which leaves open whether such abstract video representations can be directly used for high-fidelity video generation.

2.3 REPRESENTATION LEARNING AND GENERATION

Traditionally, representations learned for visual understanding (Tschannen et al., 2025; Wang et al., 2023; Siméoni et al., 2025; Oquab et al., 2023) and those optimized for generation (Sun et al., 2024; Esser et al., 2021; Podell et al., 2023; Yao et al., 2025; Kingma & Welling, 2014; Van Den Oord et al., 2017) have evolved along two distinct trajectories. Understanding-oriented representations focus on extracting high-level semantics while discarding low-level noise, whereas generation-oriented latents prioritize capturing high-fidelity textures. However, with the widespread adoption of representation alignment, recent advancements are rapidly blurring this boundary. In the image domain, works such as RAE (Zheng et al., 2025b) and VFMTok (Zheng et al., 2025a) demonstrate that features extracted from frozen visual foundation models can serve as highly expressive latent spaces for image generation. Furthermore, MAETok (Chen et al., 2025) bridges self-supervised features with discrete token-based generation, while works like UniTok (Ma et al., 2025) attempt to unify generation and understanding tasks. These pioneering efforts share a common insight: operating within a highly semantic latent space can significantly improve both generation quality and training efficiency. Despite these advances, in the specific realm of video autoencoding, semantic features are still typically relegated to mere auxiliary supervision signals. In this paper, we introduce VideoRAE to demonstrate that features from video foundation models can inherently serve as exceptional representations for visual generation.

3 METHOD: VIDEORAE

The core philosophy of VideoRAE is to abandon the traditional pixel-driven paradigm of training video tokenizers from scratch. Instead, we propose a semantic-driven compression framework built directly upon frozen Video Foundation Models (VFM). Given an input video (Sec. 3.1), we first utilize a pre-trained and frozen VFM (e.g., V-JEPA 2) to extract hierarchical semantic features spanning from shallow to deep layers. Next, we design a lightweight 1D projector to perform spatio-temporal fusion and downsampling on these features, generating highly compact tokens. Subsequently (Sec. 3.2), depending on the requirements of downstream generative tasks, these tokens are seamlessly mapped into either a continuous representation (for Diffusion Transformers) or a discrete representation (for autoregressive models). Finally (Sec. 3.3), a decoder is tasked with reconstructing the original pixels from these latent representations. During this process, we introduce a Representation Alignment (REPA) module to guide the decoder’s learning. By explicitly aligning semantics, we intrinsically maintain a high-quality topological structure in the latent space, which completely replaces the traditional KL-divergence constraint.

3.1 SEMANTIC-DRIVEN ENCODER AND 1D PROJECTOR

Given an input video clip $X \in \mathbb{R}^{B \times 3 \times T \times H \times W}$, we leverage an off-the-shelf, frozen VFM as the primary feature extractor. To simultaneously capture low-level temporal dynamics and high-level abstract semantics, we extract hierarchical intermediate features from shallow to deep layers of the VFM and explicitly fuse them via element-wise summation:

$$F_{\text{VFM}} = \sum_{l \in \mathcal{S}} f_l(X), \quad F_{\text{VFM}} \in \mathbb{R}^{B \times N_{\text{vfm}} \times D_{\text{token}}} \quad (1)$$

where $\mathcal{S}$ represents the set of selected layers, and $f_l(\cdot)$ denotes the mapping function of the l -th VFM layer. Here, N_{vfm} represents the total number of flattened 3D spatio-temporal tokens output by the VFM, and D_{token} is the embedding dimension.

Although F_{VFM} encapsulates highly expressive hierarchical semantics, its extensive sequence length (N_{vfm}) and inherent spatio-temporal redundancy pose significant computational bottlenecks for downstream generative modeling. To address this, we introduce a lightweight 1D Projector based on self-attention mechanisms, denoted as $\mathcal{E}_{\text{1D}}$. Rather than repeatedly processing the redundant 3D structure, this module models the global dependencies within F_{VFM} and dynamically condenses the long spatio-temporal sequence into a highly compact 1D sequence of base tokens:

$$Z_{\text{base}} = \mathcal{E}_{\text{1D}}(F_{\text{VFM}}), \quad Z_{\text{base}} \in \mathbb{R}^{B \times N_{\text{latent}} \times D_{\text{token}}} \quad (2)$$

where N_{latent} is the length of the compressed 1D latent sequence ($N_{\text{latent}} \ll N_{\text{vfm}}$), and D_{token} is the hidden dimension of the projector.

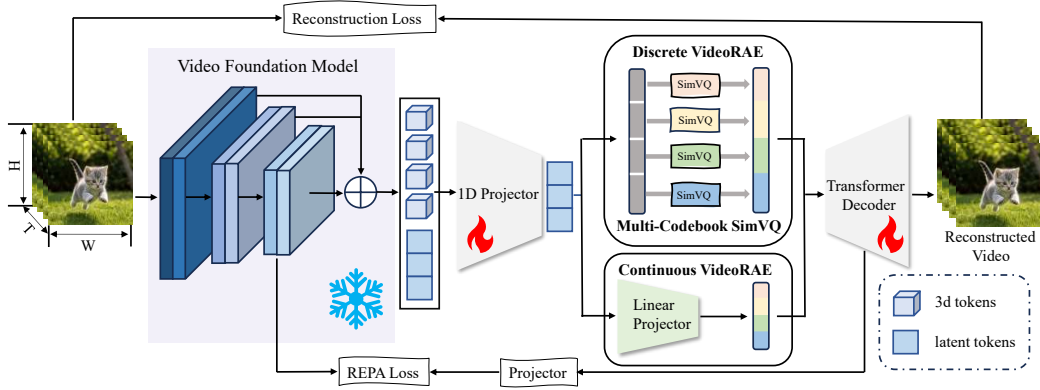


Figure 2: **Overall architecture of VideoRAE.** Given an input video, multi-scale hierarchical features are first extracted from a frozen VFM and fused, followed by a 1D self-attention projector that dynamically compresses them into compact base tokens. These tokens are then formatted into either a continuous latent space via linear projection for Diffusion Transformers, or a discrete latent space via Multi-Codebook SimVQ for Autoregressive models. Finally, during decoding, the Representation Alignment module explicitly aligns intermediate decoder features with the VFM teacher at both local and global scales, intrinsically regularizing the semantic structure of the latent manifold.

3.2 VERSATILE GENERATIVE LATENT SPACES

To cater to the distinct preferences of various generative architectures, VideoRAE is designed to format the base tokens into specific latent spaces.

Continuous VideoRAE for Diffusion Models. For generative architectures like Diffusion Transformers (DiTs), a smooth and continuous latent space is preferred. Traditional continuous 3D-VAEs heavily rely on the KL-divergence penalty to explicitly regularize the latent space into a standard Gaussian distribution. However, this rigid constraint often restricts the expressive capacity of the latents, leading to the over-smoothing of the visual details and causing severe optimization conflicts with the pixel-reconstruction objective. In contrast, VideoRAE completely bypasses the need for the KL-divergence penalty. By inheriting the highly structured semantic topology from the powerful VFM encoder, and implicitly regularizing the representation through our subsequent Representation Alignment module, the latent space is intrinsically maintained on a well-behaved semantic manifold. Consequently, without imposing any artificial distribution assumptions, a simple linear layer effectively projects Z_{base} into a latent space, yielding a highly expressive and DiT-friendly continuous latent Z_{cont} .

Discrete VideoRAE for Autoregressive Models. Autoregressive generative models rely on discrete token sequences. To break the quantization bottleneck and maximally preserve the rich semantics inherited from the VFM, a straightforward solution is to increase both the codebook size and the latent dimensionality. However, existing studies indicate that simply expanding the codebook yields diminishing returns. Moreover, under strong semantic constraints, simple vector quantization easily causes a large portion of codewords to remain underutilized during training, ultimately triggering severe codebook collapse.

To address this, we propose the **Multi-Codebook SimVQ** strategy, operating on a divide-and-conquer principle. Specifically, the continuous latent vector Z_{base} is first evenly partitioned along the channel dimension into K sub-vectors: $Z_{\text{base}} = [Z_1, Z_2, \dots, Z_K]$, where $Z_k \in \mathbb{R}^{B \times N_{\text{latent}} \times (D_{\text{token}}/K)}$. To stabilize the quantization of high-dimensional semantic features, we integrate the SimVQ paradigm into each sub-codebook. Unlike standard VQ that directly optimizes the codebook embeddings, SimVQ initializes a frozen base codebook E_k and maps it to the target latent space via a learnable Multi-Layer Perceptron (MLP). Thus, the effective sub-codebook $\mathcal{C}_k$ is generated dynamically, and only the MLP parameters are updated during training. For each chunk Z_k , its quantized representation $\hat{Z}_k$ is obtained by searching for the nearest neighbor in the mapped sub-codebook $\mathcal{C}_k$:

$$\mathcal{C}_k = \text{MLP}_k(E_k), \quad \hat{Z}_k = \arg \min_{c \in \mathcal{C}_k} \|Z_k - c\|_2^2 \quad (3)$$

The final discrete latent Z_{disc} is constructed by concatenating these quantized chunks: $Z_{\text{disc}} = [\hat{Z}_1, \hat{Z}_2, \dots, \hat{Z}_K]$.

Compared to conventional methods, the multi-codebook strategy effectively scales up the vocabulary size while keeping individual codebooks within an easily optimizable range. More importantly, preserving a high-dimensional latent space is crucial for maintaining the semantic integrity of the VFM. The semantic features extracted by VFMs are inherently complex and highly entangled. Forcing them into a low-dimensional bottleneck inevitably discards critical context and fine-grained physical dynamics. By maintaining a high overall dimensionality through the multi-codebook mechanism, the quantized discrete representations can comfortably accommodate the rich information provided by the VFM. This design not only substantially prevents model collapse but also provides ample semantic guidance for the subsequent decoding process under REPA constraints.

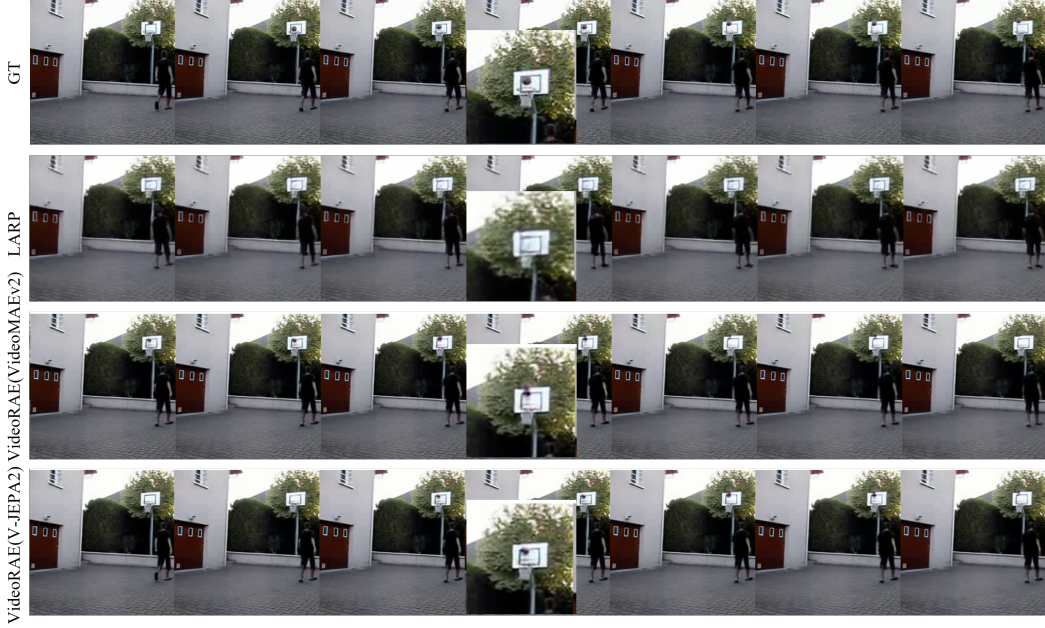


Figure 3: Visualization comparison of the Discrete-token reconstruction on the UCF-101 dataset.

3.3 DECODER AND REPRESENTATION ALIGNMENT

The decoder $\mathcal{D}_{\text{sem}}$ is responsible for reconstructing original video pixels from the highly compressed latent representations. Specifically, $\mathcal{D}_{\text{sem}}$ is designed as a Transformer-based architecture. It accepts the sequence of 1D latent tokens and concatenates them with 1024 learnable 3D tokens. After processing these concatenated tokens through stacked self-attention blocks, the learnable 3D tokens are separated and mapped back into 3D feature maps via a depatchify operation for subsequent pixel reconstruction.

Beyond traditional reconstruction losses, we propose the Representation Alignment (REPA) objective. Let H_{stu} denote the shallow 3D intermediate features of the decoder (i.e., the spatial features derived shortly after the depatchify operation). We project and spatially interpolate H_{stu} across a 3D grid to align with the final layer features of the VFM, F_{target} , yielding the aligned feature H_{align} .

The REPA loss enforces semantic consistency at both local and global scales:

$$\mathcal{L}_{\text{local}} = -\frac{1}{N_{\text{vfm}}} \sum_{i=1}^{N_{\text{vfm}}} \frac{H_{\text{align}}^{(i)} \cdot F_{\text{target}}^{(i)}}{\|H_{\text{align}}^{(i)}\| \|F_{\text{target}}^{(i)}\|} \quad (4)$$

$$\mathcal{L}_{\text{global}} = -\frac{\text{Pool}(H_{\text{align}}) \cdot \text{Pool}(F_{\text{target}})}{\|\text{Pool}(H_{\text{align}})\| \|\text{Pool}(F_{\text{target}})\|} \quad (5)$$

$$\mathcal{L}_{\text{REPA}} = \mathcal{L}_{\text{local}} + \lambda \mathcal{L}_{\text{global}} \quad (6)$$

where λ is the weight for the global loss. Crucially, this semantic alignment implicitly regularizes the geometry of the latent space, forcing the representations to reside on a well-behaved semantic manifold. It forces the latent features to adhere to the high-quality semantic topological structure of the foundation model. This structural preservation is the fundamental reason why we can discard the KL-divergence loss used in traditional 3D-VAEs in our VideoRAE.

3.4 TRAINING OBJECTIVES

The end-to-end training of VideoRAE integrates pixel-level reconstruction, adversarial perception, and our semantic alignment constraint. The overall loss function is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{adv}} + \lambda_{\text{repa}} \mathcal{L}_{\text{REPA}} (+\mathcal{L}_{\text{vq}}) \quad (7)$$

where the reconstruction loss $\mathcal{L}_{\text{rec}}$ comprises L1 and LPIPS perceptual losses to ensure basic pixel correctness. The generative adversarial loss $\mathcal{L}_{\text{adv}}$ (GAN loss) is retained and specifically dedicated to recovering high-frequency visual textures. Meanwhile, $\mathcal{L}_{\text{REPA}}$ injects macroscopic structure and physical semantics. Finally, $\mathcal{L}_{\text{vq}}$ is included exclusively when training the discrete VideoRAE, accounting for codebook commitment losses.

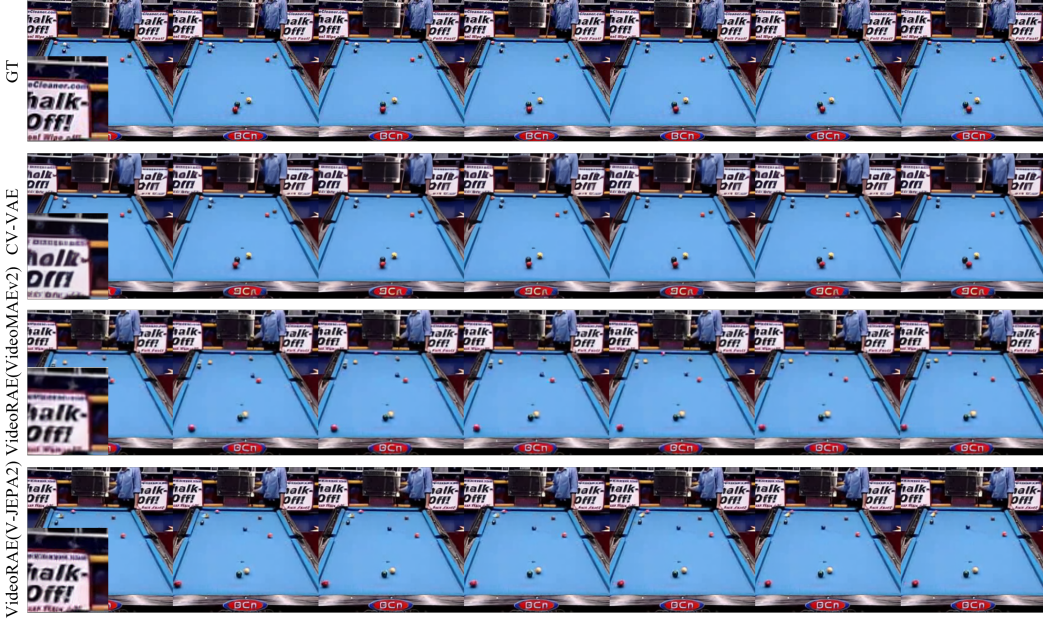


Figure 4: Visualization comparison of the Continuous-space reconstruction on the UCF-101 dataset.

4 EXPERIMENTS

Datasets: Building upon previous works, we train the VideoRAE on the UCF-101 (Soomro et al., 2012) and Kinetics-600 datasets (Carreira et al., 2018), using $16 \times 256 \times 256$ video clips for both training and evaluation; meanwhile, the text-to-video (T2V) generation models are trained specifically on the VideoUFO dataset (Wang & Yang, 2026). For the assessment of reconstruction performance, evaluations are conducted on the UCF-101 (Soomro et al., 2012) and TokenBench (Agarwal et al., 2025) datasets. Furthermore, we evaluate class-to-video (C2V) generation capabilities on the UCF-101 dataset (Soomro et al., 2012), and assess text-to-video generation performance on VBench (Huang et al., 2024). Following established evaluation protocols (Wang et al., 2024a), we sample 10,000 cases to assess the C2V generation of discrete tokenizers, whereas for continuous VAE models, we evaluate using a subset of 2,048 sampled cases.

Implementation details: For feature extraction, we employ pre-trained and frozen V-JEPA 2 (Assran et al., 2025) and VideoMAEv2 (Wang et al., 2023) as the teacher models, extracting multi-level intermediate features. The pre-trained video foundation models utilize a $16 \times$ spatial downsampling and a $2 \times$ temporal downsampling. Subsequently, our 1D self-attention projector compresses the

input 3D features into a highly concentrated sequence of 512 or 1024 latent tokens. For the continuous latent space, the channel dimension is linearly projected to a 32/64-dimensional bottleneck. For the discrete latent space, we adopt the Multi-Codebook SimVQ strategy, setting the number of sub-codebooks to $K = 4$, with a vocabulary size of $V = 4096$ for each sub-codebook. To evaluate the generative performance of the structured latent spaces, we train standard Diffusion Transformers (DiTs) (Peebles & Xie, 2023) on the continuous VideoRAE, and LLaMA-style Transformer-based autoregressive (AR) models (Sun et al., 2024) on the discrete VideoRAE. All models are trained on NVIDIA H100 GPUs.

Baselines. To comprehensively evaluate the performance of VideoRAE across diverse tasks and architectures, we select appropriate state-of-the-art (SOTA) baselines tailored to each specific experimental setting. Notably, to guarantee fairness, we reproduced and re-tested all open-source baseline models under identical evaluation protocols and environments:

- *Video Reconstruction.* Under the discrete representation paradigm, we compare our method against leading video tokenizers, including TATS (Ge et al., 2022), OmniTokenizer (Wang et al., 2024b), LARP (Wang et al., 2024a), and SweetTok (Tan et al., 2025). For continuous autoencoding, we benchmark against not only CV-VAE (Zhao et al., 2024) and LeanVAE (Cheng & Yuan, 2025), but also the most recent powerful autoencoders such as LTX-VAE (HaCohen et al., 2024), LTX2 (HaCohen et al., 2026), CogVideoX-VAE (Yang et al., 2025), and WAN2.1-VAE (Wan et al., 2025).
- *Class-Conditional Generation.* On the UCF-101 dataset, we carefully align the baselines with the downstream generative frameworks. For Autoregressive (AR) models operating in the discrete space, we evaluate against CogVideo (Hong et al., 2022), TATS (Ge et al., 2022), VideoLaVIT (Jin et al., 2024), OmniTokenizer (Wang et al., 2024b), LARP (Wang et al., 2024a), and SweetTok (Tan et al., 2025). For Diffusion Transformers (DiT) in the continuous space, our comparisons encompass Latte (Ma et al., 2024), WF-VAE (Li et al., 2025), DeCo-VAE (Yin et al., 2025), as well as generative pipelines built upon LeanVAE (Cheng & Yuan, 2025), LTX-VAE (HaCohen et al., 2024), CogVideoX-VAE (Yang et al., 2025), and WAN2.1-VAE (Wan et al., 2025).
- *Text-to-Video Generation.* Finally, for the complex open-domain T2V generation task evaluated on the VBench suite, we compare our continuous VideoRAE directly with the state-of-the-art open-source continuous model, LTX-VAE (HaCohen et al., 2024).

Table 1: Reconstruction performance of discrete tokenizers on UCF-101 and TokenBench.

Method	#rToken	UCF-101			TokenBench-256×256		
		PSNR↑	LPIPS↓	rFVD↓	PSNR↑	LPIPS↓	rFVD↓
<i>AR generative models with discrete video tokenizers</i>							
TATS	1024	-	-	162	-	-	-
OmniTokenizer	5120	29.25	0.11	28	25.55	0.11	44
LARP-L-Long	1024	27.88	0.12	35	28.65	0.11	45
SweetTok	1280	29.27	0.07	20	-	-	-
VideoRAE(VideoMAEv2)	1024	29.94	<u>0.10</u>	<u>16</u>	30.69	0.09	<u>33</u>
VideoRAE(V-JEPA 2)	1024	<u>29.39</u>	<u>0.10</u>	13	<u>29.93</u>	<u>0.10</u>	28

4.1 MAIN RESULTS

Reconstruction Quality Comparison. We comprehensively evaluate the reconstruction capabilities of VideoRAE under both discrete and continuous latent paradigms on the UCF-101 (Soomro et al., 2012) and TokenBench (Agarwal et al., 2025) datasets. As delineated in Table 1, under the discrete tokenizer configuration, VideoRAE establishes new state-of-the-art benchmarks. Specifically, VideoRAE(V-JEPA 2) achieves an excellent rFVD of 13 on UCF-101 (Soomro et al., 2012) and 28 on TokenBench (Agarwal et al., 2025), significantly outperforming recent strong baselines such as LARP (Wang et al., 2024a) and SweetTok (Tan et al., 2025). Meanwhile, VideoRAE(VideoMAEv2) demonstrates superior spatial fidelity, securing the highest PSNR across both datasets. This verifies the superiority of high-dimensional semantic representations derived from video foundation models.

Table 2: Reconstruction performance of continuous autoencoders on UCF-101 and TokenBench.

Method	Config	UCF-101			TokenBench-256×256		
		PSNR↑	LPIPS↓	rFVD↓	PSNR↑	LPIPS↓	rFVD↓
<i>Diffusion generative models with continuous video tokenizers</i>							
CV-VAE	4096×4	30.11	0.12	57	30.80	0.11	61
LTX-VAE	128×128	32.02	0.10	35	33.74	0.08	32
LTX2	128×128	28.88	0.12	15	29.36	0.11	36
LeanVAE	4096×4	30.99	0.11	10	31.79	0.10	23
VideoRAE(VideoMAEv2)	512×32	<u>31.23</u>	0.08	13	<u>32.06</u>	0.08	<u>25</u>
VideoRAE(V-JEPA 2)	512×32	30.40	<u>0.09</u>	<u>14</u>	31.19	<u>0.09</u>	<u>25</u>
CogvideoX-VAE	4096×16	36.22	0.05	7	<u>36.35</u>	0.04	<u>9</u>
WAN2.1-VAE	4096×16	35.70	0.05	4	37.02	0.04	6
VideoRAE(VideoMAEv2)	1024×64	33.64	<u>0.06</u>	<u>5</u>	34.25	<u>0.07</u>	11
VideoRAE(V-JEPA 2)	1024×64	32.14	0.07	7	32.56	0.08	13

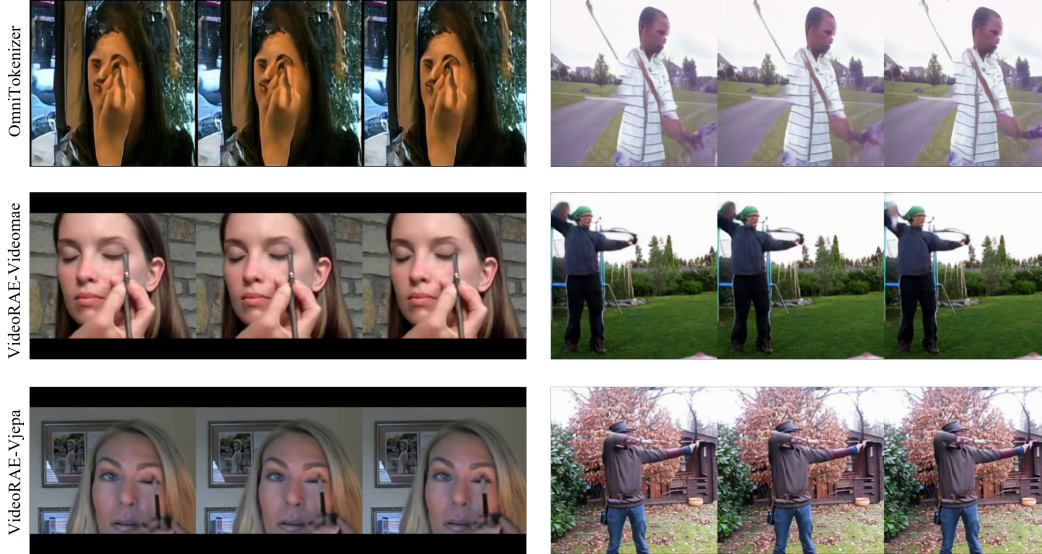


Figure 5: Visualization of the Discrete-token generation results on the UCF-101 dataset.

Moreover, it demonstrates that our Multi-Codebook SimVQ strategy can successfully reconstruct high-fidelity pixels even within a highly compressed discrete semantic space.

For the continuous latent space presented in Table 2, VideoRAE also delivers highly competitive performance, exhibiting a strong balance between spatial texture recovery and perceptual quality. We observe that while models like LTX-VAE (HaCohen et al., 2024) achieve high PSNR, they struggle with spatio-temporal consistency (resulting in a sub-optimal rFVD of 35). Conversely, although LeanVAE (Cheng & Yuan, 2025) achieves a low rFVD, it relies heavily on a massive latent resolution. In contrast, using a highly compact 512×32 bottleneck, VideoRAE delivers top-tier perceptual quality alongside competitive PSNR performance. Even in comparison with industry-grade VAEs, VideoRAE demonstrates competitive performance despite being trained solely on the K600 (Carreira et al., 2018) and UCF-101 (Soomro et al., 2012) datasets. The aforementioned reconstruction results across both discrete and continuous models effectively validate our core hypothesis: representations from frozen video foundation models do not hinder reconstruction. Instead, they can surpass traditional pixel-driven autoencoders in reconstructing videos with exceptional perceptual quality and fidelity.

Discrete-token class-conditional generation. To evaluate the generative capability of the discrete representations, we train autoregressive (AR) models on our tokenized latent space. As shown in Ta-

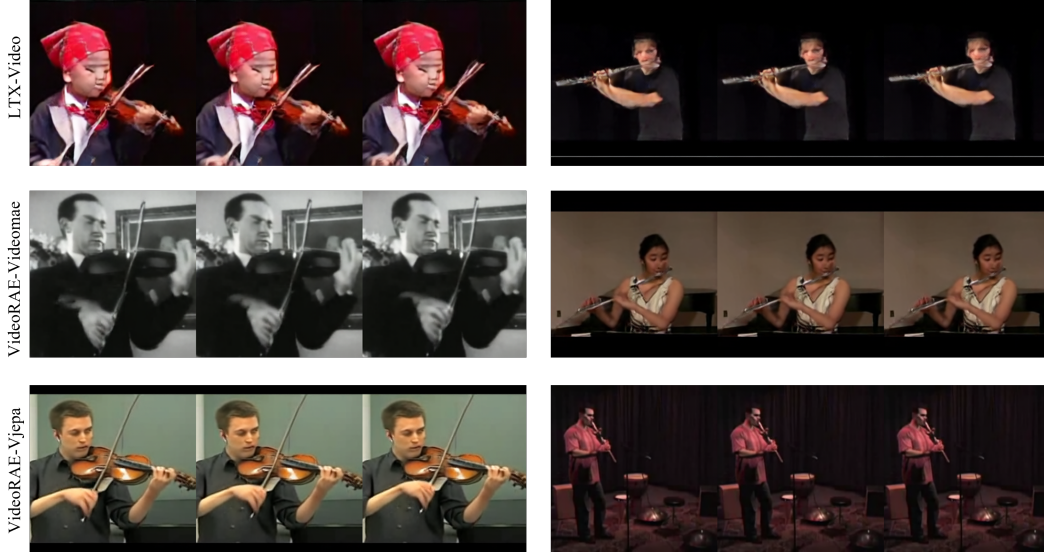


Figure 6: Visualization of the Continuous-space generation results on the UCF-101 dataset.

ble 3, VideoRAE demonstrates highly competitive performance on the UCF-101 class-conditional generation task. Notably, VideoRAE(V-JEPA 2) achieves a state-of-the-art gFVD of 40. Compared to the previously leading method, SweetTok(Tan et al., 2025) (which obtains a gFVD of 65 using 1.9B parameters and 1280 tokens), VideoRAE secures a decisive advantage while utilizing fewer tokens and a smaller generator size. The feature space derived from visual foundation models is inherently suited for generative tasks. Furthermore, the discrete tokens obtained through our Multi-Codebook SimVQ quantization and semantic alignment strategy effectively preserve the rich semantics of the VFM, making the downstream autoregressive prediction tasks more efficient and accurate.

Continuous-space class-conditional generation. For the continuous latent space, we adopt a Diffusion Transformer (DiT)(Rombach et al., 2022) architecture to evaluate the generation quality. As presented in Table 4, VideoRAE exhibits strong generative capabilities. While recent advanced continuous VAEs (such as LeanVAE(Cheng & Yuan, 2025) and LTX-VAE(HaCohen et al., 2024)) have improved the baseline gFVD to 164 and 161, respectively, VideoRAE further reduces this metric to 99 (VideoMAEv2) and 93 (V-JEPA 2). Furthermore, we extended industry-grade models, including WAN2.1-VAE(Wan et al., 2025) and CogVideoX(Yang et al., 2025), to class-to-video generation on the UCF-101 dataset. Although they outperform existing methods like Deco-VAE(Yin et al., 2025), VideoRAE still delivers a distinct superiority in generation performance, even in comparison with them. This substantial improvement indicates that the KL-free and semantic-aligned continuous latent space provided by VideoRAE offers a structurally sound and generation-friendly manifold for diffusion models, thereby facilitating the synthesis of high-fidelity videos.

Discussion: Reconstruction Fidelity vs. Generative Quality. Across both discrete and continuous modalities, we observe a thought-provoking phenomenon: VideoRAE equipped with VideoMAEv2 consistently exhibits superior pixel-level reconstruction fidelity (i.e., higher PSNR and lower LPIPS). However, our empirical results reveal that ultimate fine-grained reconstruction does not strictly translate to higher generation quality. Conversely, VideoRAE(V-JEPA 2) consistently achieves better gFVD scores across all generative tasks. We attribute this discrepancy to the fundamentally distinct pre-training paradigms of the two video foundation models. VideoMAEv2 is driven by masked pixel reconstruction, which forces the model to heavily focus on fine-grained local motion details. In contrast, V-JEPA 2 utilizes a latent-space prediction mechanism, an objective that compels its representations to be more aligned with high-level semantics and macroscopic spatio-temporal dynamics. Although this deep semantic abstraction slightly compromises absolute pixel-level reconstruction precision, it provides a significantly more generation-friendly latent space that eases the modeling burden on downstream generators. This leads to a crucial conclusion: for achieving high-quality video generation, endowing the latent space with stronger semantic under-

standing and visual perception is substantially more beneficial than pursuing rigid low-level pixel fidelity.

Table 3: Class-conditional generation evaluation (Discrete AR) on UCF-101, evaluated on 10K generated samples.

Tokenizer	#Tokens	#Params	gFVD ↓
CogVideo	6800	9.4B	626
TATS	4096	321M	332
Video-LaVIT	512	7B	280
OmniTokenizer	5120	650M	191
LARP-L	1024	632M	99
SweetTok	1280	650M	84
SweetTok	1280	1.9B	65
VideoRAE(VideoMAEv2)	1024	1.3B	<u>45</u>
VideoRAE(V-JEPA 2)	1024	1.3B	40

Table 4: Class-conditional generation evaluation (Continuous DiT) on UCF-101, evaluated on 2K generated samples.

Method	Configs	gFVD ↓
Latte	4096*4	478
LeanVAE	4096*4	164
LTX-VAE	128*128	161
VideoRAE(VideoMAEv2)	512*32	<u>99</u>
VideoRAE(V-JEPA 2)	512*32	93
WF-VAE	4096*16	371
LeanVAE	4096*16	175
DeCo-VAE	4096*16	166
CogvideoX-VAE	4096*16	151
WAN2.1-VAE	4096*16	126
VideoRAE(VideoMAEv2)	1024*64	<u>103</u>
VideoRAE(V-JEPA 2)	1024*64	94

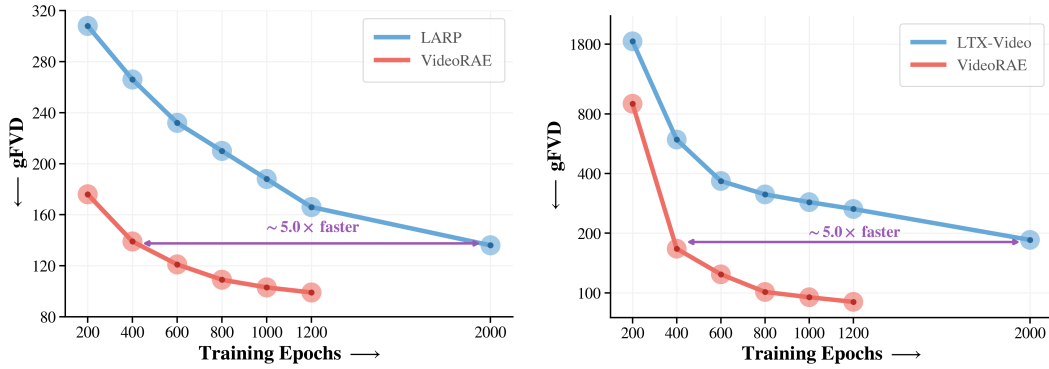


Figure 7: Training convergence speed comparison in AR (left) and DiT (right) generative modeling.

Accelerated Convergence in Generative Training. To further illustrate the efficiency of our semantic-driven latent space, we first compare the generative training convergence speed between VideoRAE and the strong baseline LARP under the discrete tokenizer setting, with 2,000 generated samples evaluated for each model. As illustrated in Figure 7, the autoregressive generative model trained on the traditional pixel-driven LARP tokenizer requires 2000 epochs to achieve a gFVD score of 136. In stark contrast, the model trained on VideoRAE’s representations reaches a comparable

gFVD of 139 at merely 400 epochs, demonstrating an impressive $\sim 5.0\times$ acceleration in training convergence. Similarly, within the continuous latent space, VideoRAE exhibits a comparable trend of significantly accelerating model convergence when evaluated against LTX-VAE. By alleviating the immense burden on the generator of learning complex spatio-temporal dynamics from scratch, VideoRAE enables downstream generative models to focus entirely on learning high-level sequence transition logic. This fundamentally validates our claim that a semantically aligned latent space is pivotal for reducing generative training costs.

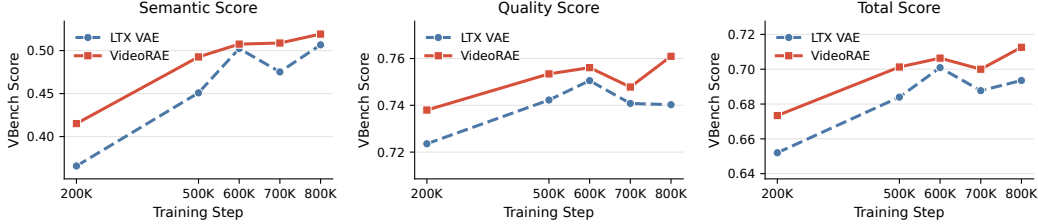


Figure 8: Convergence comparison between VideoRAE and LTX-VAE on VBench.

Table 5: Text-to-video generation evaluation on VBench. VideoRAE achieves stronger overall performance than LTX-VAE, with clear gains in both visual consistency and semantic alignment.

Method	Total Score $\uparrow$	Quality Score $\uparrow$	Semantic Score $\uparrow$	Subject Cons. $\uparrow$	Background Cons. $\uparrow$	Imaging Qual. $\uparrow$	Object Class $\uparrow$	Human Action $\uparrow$	Scene $\uparrow$
LTX-VAE	69.35	74.03	50.67	84.14	95.51	43.54	58.51	54.50	23.29
VideoRAE (Ours)	71.26	76.09	51.92	91.27	97.02	46.02	59.07	60.80	27.57

Text-to-Video Generation on VBench. To further evaluate the generative capacity of VideoRAE in complex open-domain scenarios, we extend our study to text-to-video generation. We build our T2V system on top of the OpenSora2 (Zheng et al., 2025c) training framework and replace its original video autoencoder with either LTX-VAE (HaCohen et al., 2024) or our trained VideoRAE, while keeping the diffusion backbone and training recipe comparable. Specifically, we train a 2B-scale MMDiT-style diffusion transformer, consisting of dual-stream text-video blocks followed by single-stream multimodal blocks, with hidden size 2048, 16 attention heads, 8 dual-stream blocks, and 16 single-stream blocks. The model is conditioned on T5 and CLIP text embeddings and trained to predict the flow-matching target in the latent space of each video tokenizer. We train both systems on VideoUFO (Wang & Yang, 2026) at 256×256 resolution and 12 FPS, using 17-frame clips for LTX-VAE and 16-frame clips for VideoRAE, matching the native temporal interface of each autoencoder. Both models are trained with a global batch size of 256 for 800K optimization steps.

During training, we periodically evaluate intermediate checkpoints on VBench (Huang et al., 2024) and report the evolution of Total Score, Quality Score, and Semantic Score in Figure 8. As shown in the figure, the VideoRAE-based generator converges faster than the LTX-VAE counterpart and consistently achieves better scores throughout training. After convergence, we further report detailed VBench metrics in Table 5. These results are consistent with our C2V observations and suggest that VideoRAE provides a more effective latent space for generative modeling. By using a visual foundation model as the encoder, VideoRAE produces latents that are both semantically enriched and highly compact, allowing the diffusion transformer to model high-level video content more efficiently. This improved representation reduces the burden on the diffusion model to rediscover semantic structure from low-level reconstruction latents, leading to faster convergence and stronger final generation quality.

4.2 ABLATION STUDY

Quantization Strategies. We ablate the impact of four quantization methods on the VideoRAE model: VQ, SimVQ (Zhu et al., 2025), MCQ, and Multi-codebook SimVQ. The traditional Vector Quantization yields the poorest reconstruction and generation performance. We attribute this to codebook collapse and non-convergence, which are caused by the excessively low quantization dimensionality when dealing with highly semantic latent spaces. In contrast, SimVQ scales the

quantization dimension to 128, bringing a noticeable improvement. When transitioning to MCQ, it achieves a significant performance boost compared to the single-codebook counterpart. This improvement, in our view, stems from MCQ’s ability to effectively expand the codebook capacity while avoiding the optimization difficulties associated with large codebooks. Building upon this, Multi-codebook SimVQ further extends the quantization dimension to 512. Such high-dimensional quantization effectively mitigates semantic loss, thereby delivering superior generation capabilities.

Table 6: Ablation study on quantization strategies.

Quantization Method	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	rFVD $\downarrow$	gFVD $\downarrow$
VQ	25.50	0.82	0.14	35	95
SimVQ	28.32	0.87	0.12	24	59
MCQ	29.10	0.90	0.10	16	51
Multi-codebook SimVQ	29.39	0.91	0.10	13	40

Representation Alignment. To evaluate the effectiveness of our proposed Representation Alignment (REPA) module, we conduct ablation experiments by training both continuous and discrete models with and without the REPA objective. As shown in Table 7, the integration of REPA consistently yields significant performance improvements across most evaluation metrics in both latent spaces. Specifically, incorporating REPA reduces the gFVD from 105 to 93 in the continuous setting, and from 67 to 40 in the discrete setting, while also substantially improving perceptual metrics. Furthermore, we observe that applying a KL penalty in the continuous setting compromises both reconstruction and generation quality to some extent. This demonstrates that aligning the decoder’s features with the VFM teacher at both local and global scales not only improves the model’s perceptual quality, but also acts as a strong semantic constraint that successfully regularizes the latent space topology without requiring any KL penalty, rendering this high-quality latent distribution more conducive to downstream generation tasks.

Table 7: Ablation study on the effect of REPA for semantic alignment.

Latent Space	REPA	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	rFVD $\downarrow$	gFVD $\downarrow$
Continuous	w/o	30.57	0.91	0.10	18	105
Continuous-kl	w/o	27.45	0.81	0.12	31	101
Continuous	w/	30.40	0.91	0.10	14	93
Discrete	w/o	29.71	0.90	0.10	15	67
Discrete	w/	29.39	0.91	0.10	13	40

Multi-Scale Feature Aggregation. We investigate the impact of aggregating features from different levels of the visual foundation model (VFM) teacher. As shown in Table 8, relying solely on deep layers yields sub-optimal reconstruction results. In terms of the perceptual quality metric, rFVD, whether shallow features are incorporated does not lead to a significant discrepancy. However, since deep features primarily capture high-level semantics while discarding fine-grained spatial details, they exhibit an insufficient capacity to restore high-frequency details, resulting in a severely compromised PSNR. By progressively incorporating shallower layers, all reconstruction metrics steadily improve. Ultimately, utilizing multi-level features (Layers 8–24) achieves the best trade-off between reconstruction and generation performance, yielding a PSNR of 29.39 and a gFVD of 40. This highlights the necessity of multi-scale feature aggregation to simultaneously capture both macro-semantics and micro-level texture details.

Feature Fusion Strategies. We compare four feature fusion strategies for aggregating multi-layer VFM representations: Element-wise Addition, Decoupled Fusion, Gated Fusion, and AdaLN. The results are summarized in Table 9. Empirically, the more complex Decoupled Fusion strategy and the simple Element-wise Addition perform on par with each other in terms of both reconstruction and generation quality, both slightly outperforming Gated Fusion and AdaLN. Given the advantages of simplicity, high computational efficiency, and zero additional parameter overhead, we select Element-wise Addition as our default feature fusion strategy.

Table 8: Ablation study on multi-scale feature aggregation.

Aggregated Layers	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	rFVD $\downarrow$	gFVD $\downarrow$
Layer 24	26.59	0.85	0.14	22	51
Layers 16–24	28.35	0.87	0.12	18	41
Layers 12–24	29.11	0.90	0.11	15	42
Layers 8–24	29.39	0.91	0.10	13	40

Table 9: Ablation study on feature fusion strategies.

Fusion Strategy	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	rFVD $\downarrow$	gFVD $\downarrow$
Decoupled Fusion	29.43	0.91	0.10	13	43
Semantic-Guided AdaLN Fusion	29.03	0.90	0.11	18	51
Gated Fusion	29.10	0.90	0.11	16	49
Element-wise Add(Ours)	29.39	0.91	0.10	13	40

5 CONCLUSION

In this paper, we introduced VideoRAE, a semantic-driven video autoencoder that bridges the gap between visual understanding and generation. By challenging the prevailing assumption that frozen Video Foundation Models (VFM) are unsuitable for pixel-level reconstruction, we demonstrated that VFM representations can serve as highly robust latent spaces for generative modeling. Designed as a unified framework, VideoRAE seamlessly supports both continuous and discrete generative paradigms. Through our Representation Alignment (REPA) objective, the latent manifold is intrinsically regularized, completely obviating the need for traditional distribution constraints. Extensive experiments confirm that VideoRAE not only achieves superior reconstruction quality but significantly accelerates the convergence of downstream generative models. Most importantly, our findings reveal that a semantically structured latent space is far more critical for video generation than rigid pixel-level fidelity. By establishing VFMs as the new default latent foundation, VideoRAE paves a promising pathway toward unifying representation learning and generative modeling.

REFERENCES

- Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chatopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025.
- Mahmoud Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.
- Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about kinetics-600. *arXiv preprint arXiv:1808.01340*, 2018.
- Hao Chen, Yujin Han, Fangyi Chen, Xiang Li, Yidong Wang, Jindong Wang, Ze Wang, Zicheng Liu, Difan Zou, and Bhiksha Raj. Masked autoencoders are effective tokenizers for diffusion models. In *Forty-second International Conference on Machine Learning*, 2025.
- 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*, pp. 15692–15702, 2025.

- Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12873–12883, 2021.
- Songwei Ge, Thomas Hayes, Harry Yang, Xi Yin, Guan Pang, David Jacobs, Jia-Bin Huang, and Devi Parikh. Long video generation with time-agnostic vqgan and time-sensitive transformer. *arXiv preprint arXiv:2204.03638*, 2022.
- Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.
- Pengbo Guo, Junke Wang, Zhen Xing, Chengxu Liu, Daoguo Dong, Xueming Qian, and Zuxuan Wu. Dera: Decoupled representation alignment for video tokenization. *arXiv preprint arXiv:2512.04483*, 2025.
- Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, et al. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024.
- Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, et al. Ltx-2: Efficient joint audio-visual foundation model. *arXiv preprint arXiv:2601.03233*, 2026.
- Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pre-training for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022.
- Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21807–21818, 2024.
- Yang Jin, Zhicheng Sun, Kun Xu, Liwei Chen, Hao Jiang, Quzhe Huang, Chengru Song, Yuliang Liu, Di Zhang, Yang Song, et al. Video-lavit: Unified video-language pre-training with decoupled visual-motional tokenization. *arXiv preprint arXiv:2402.03161*, 2024.
- Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014.
- 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*, pp. 17778–17788, 2025.
- Chuofan Ma, Yi Jiang, Junfeng Wu, Jihan Yang, Xin Yu, Zehuan Yuan, Bingyue Peng, and Xiaojuan Qi. Unitok: A unified tokenizer for visual generation and understanding. *arXiv preprint arXiv:2502.20321*, 2025.
- Xin Ma, Yaohui Wang, Xinyuan Chen, Gengyun Jia, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation. *arXiv preprint arXiv:2401.03048*, 2024.
- Lorenzo Mur-Labadia, Matthew Muckley, Amir Bar, Mahmoud Assran, Koustuv Sinha, Michael Rabbat, Yann LeCun, Nicolas Ballas, and Adrien Bardes. V-jepa 2.1: Unlocking dense features in video self-supervised learning. *arXiv preprint arXiv:2603.14482*, 2026.
- 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.
- William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

- Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sd-xl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.
- 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*, pp. 10684–10695, 2022.
- Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.
- 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.
- Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024.
- Zhentao Tan, Ben Xue, Jian Jia, Junhao Wang, Wencai Ye, Shaoyun Shi, Mingjie Sun, Wenjin Wu, Quan Chen, and Peng Jiang. Sweettok: Semantic-aware spatial-temporal tokenizer for compact video discretization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23541–23550, 2025.
- Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.
- Aäron Van Den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. In *International conference on machine learning*, pp. 1747–1756. PMLR, 2016.
- Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.
- Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Fei Wu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.
- Hanyu Wang, Saksham Suri, Yixuan Ren, Hao Chen, and Abhinav Shrivastava. Larp: Tokenizing videos with a learned autoregressive generative prior. *arXiv preprint arXiv:2410.21264*, 2024a.
- Junke Wang, Yi Jiang, Zehuan Yuan, Binyue Peng, Zuxuan Wu, and Yu-Gang Jiang. Omnitokenizer: A joint image-video tokenizer for visual generation. *Advances in Neural Information Processing Systems*, 37:28281–28295, 2024b.
- Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14549–14560, 2023.
- Wenhao Wang and Yi Yang. Videoufo: A million-scale user-focused dataset for text-to-video generation. *Advances in Neural Information Processing Systems*, 38, 2026.
- Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, et al. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022.
- 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*, pp. 22922–22932, 2025.

- Tianwei Xiong, Jun Hao Liew, Zilong Huang, Zhijie Lin, Jiashi Feng, and Xihui Liu. Evatok: Adaptive length video tokenization for efficient visual autoregressive generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23249–23259, 2026.
- 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. In *International Conference on Learning Representations*, volume 2025, pp. 83048–83077, 2025.
- Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.
- 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.
- Xiangdong Zhang, Jiaqi Liao, Shaofeng Zhang, Fanqing Meng, Xiangpeng Wan, Junchi Yan, and Yu Cheng. Videorepa: Learning physics for video generation through relational alignment with foundation models. *arXiv preprint arXiv:2505.23656*, 2025.
- 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.
- Anlin Zheng, Xin Wen, Xuanyang Zhang, Chuofan Ma, Tiancai Wang, Gang Yu, Xiangyu Zhang, and Xiaojuan Qi. Vision foundation models as effective visual tokenizers for autoregressive image generation. *arXiv preprint arXiv:2507.08441*, 2025a.
- Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. *arXiv preprint arXiv:2510.11690*, 2025b.
- Zangwei Zheng, Xiangyu Peng, Yuxuan Lou, Chenhui Shen, Tom Young, Xinying Guo, Binluo Wang, Hang Xu, Hongxin Liu, Mingyan Jiang, et al. Open-sora 2.0: Training a commercial-level video generation model in \$200 k. *arXiv preprint arXiv:2503.09642*, 2025c.
- Tianxiong Zhong, Xingye Tian, Boyuan Jiang, Xuebo Wang, Xin Tao, Pengfei Wan, and Zhiwei Zhang. Vfrtok: Variable frame rates video tokenizer with duration-proportional information assumption. *Advances in Neural Information Processing Systems*, 38:3325–3345, 2026.
- Yongxin Zhu, Bocheng Li, Yifei Xin, Zhihua Xia, and Linli Xu. Addressing representation collapse in vector quantized models with one linear layer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22968–22977, 2025.