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

Anonymous authors
Paper under double-blind review

ABSTRACT

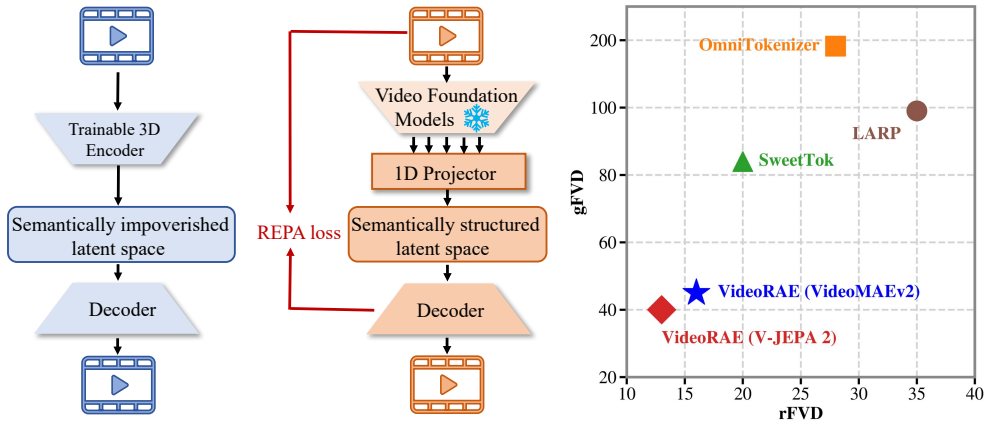
Video generation models typically rely on 3D-VAEs trained for pixel-level reconstruction, whose latent spaces may underrepresent semantic structure. We introduce VideoRAE, a representation autoencoder that converts features from a frozen video foundation model into compact, reconstruction-capable latents for video generation. Lightweight projectors compress multi-scale hierarchical features into compact 1D sequences or structured 3D grids, supporting continuous latents for Diffusion Transformers and discrete tokens for autoregressive models through multi-codebook high-dimensional quantization. During decoding, a local–global representation alignment objective transfers semantic structure from the frozen encoder and removes the need for KL regularization. Comprehensive experiments show that VideoRAE achieves strong reconstruction in both continuous and discrete regimes. On UCF-101, autoregressive and diffusion generators built on VideoRAE achieve class-conditional gFVD scores of 40 and 93, respectively, while converging approximately $5\times$ faster than autoencoder baselines. In a 4B-scale text-to-video study with randomly initialized diffusion backbones, frozen VideoRAE-3D achieves higher VBench total scores than Wan2.1-VAE at all 11 evaluated checkpoints, reaching 69.07 versus 68.24 at 200k training steps. These results establish frozen video foundation representations as compact, versatile, and generation-friendly video latents. Code and models will be released.

1 INTRODUCTION

Generative modeling has shifted from pixel-space (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, VAEs (Cheng & Yuan, 2025; Wan et al., 2025; Yang et al., 2025) and discrete tokenizers (Wang et al., 2024b; Hong et al., 2022) compress raw clips into compact spatio-temporal tokens to reduce computational complexity. However, driven primarily by pixel-level reconstruction and adversarial losses, these tokenizers are biased toward local appearance, underemphasizing high-level semantics and long-range 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 VFMTok (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 video tokenizers (Guo et al., 2025; Zhang et al., 2025; Xiong et al., 2026) merely incorporate semantic alignment as auxiliary supervision within conventional pixel-driven VAEs. Therefore, whether frozen VFM representations can be transformed into compact latent spaces that support both high-fidelity reconstruction and downstream video generation remains underexplored.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) Previous video tokenizer (b) VideoRAE (c) Video reconstruction and generation on UCF-101 dataset

Figure 1: Conceptual comparison between traditional video tokenizers and VideoRAE. Traditional 3D-VAEs are trained from scratch and driven solely by pixel-level MSE. 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.

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 4B-scale text-to-video study with randomly initialized diffusion backbones and frozen video representations. VideoRAE-3D achieves higher VBench total scores than Wan2.1-VAE at every evaluated checkpoint through 200k steps under a shared training framework and matched batch size. 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 semantic-driven 1D and 3D projectors 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 show that VideoRAE achieves competitive reconstruction quality and provides a structured semantic latent space that accelerates downstream generative training and improves generation quality.

2 Related Works

2.1 Video Autoencoders and Tokenizers

Video VAEs and tokenizers compress high-dimensional videos into compact latent spaces, essential for efficient generative modeling. For continuous spaces, recent methods enhance robustness via variable frame rates (VFRTok (Zhong et al., 2026)), decouple structure and dynamics (VidTwin (Wang et al., 2025)), or boost efficiency with lightweight wavelet designs (LeanVAE (Cheng & Yuan, 2025)). In discrete tokenization, LARP (Wang et al., 2024a) pioneers a 1D architecture with an autoregressive prior, while SweetTok (Tan et al., 2025) mitigates quantization loss via dual-stream codebooks. Despite their success, these models are typically trained from scratch, relying heavily on pixel-level reconstruction and adversarial losses. Unlike these pixel-driven approaches, VideoRAE explores a semantic-driven paradigm by leveraging frozen VFMs, 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 transferable representations from diverse data through scalable self-supervised pre-training. In the video domain, VideoMAEv2 (Wang et al., 2023) scales masked autoencoding to billion-parameter levels via dual-masking, developing robust spatio-temporal reasoning from sparse inputs. Similarly, V-JEPA 2 (Assran et al., 2025) utilizes a joint-embedding predictive architecture to predict masked latent features, deliberately discarding low-level noise to focus on high-level semantics. While these VFMs exhibit exceptional understanding capabilities, their highly abstract nature has historically deterred their direct application in pixel reconstruction. Consequently, whether such abstract representations can be directly utilized for high-fidelity video generation remains an open question.

2.3 Representation Learning and Generation

Traditionally, representations for visual understanding (Tschannen et al., 2025; Wang et al., 2023; Siméoni et al., 2025; Oquab et al., 2023) and 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; Xia et al., 2026) evolved along distinct trajectories. Understanding representations extract high-level semantics by discarding low-level noise, whereas generation latents prioritize high-fidelity textures. However, recent advancements in representation alignment are rapidly blurring this boundary. In the image domain, RAE (Zheng et al., 2025b) and VFMTok (Zheng et al., 2025a) demonstrate that frozen VFM features serve as highly expressive latent spaces for generation. Furthermore, MAETok (Chen et al., 2025) bridges self-supervised features with diffusion generation, while UniTok (Ma et al., 2025) unifies generation and understanding. These efforts reveal that operating in a highly semantic latent space significantly improves generation quality and training efficiency. Despite this, in video autoencoding, semantic features remain largely relegated to auxiliary supervision signals. Here, we introduce VideoRAE to demonstrate that video foundation model features inherently provide 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 (VFMs). 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. A lightweight 1D projector then fuses and downsamples these features into highly compact tokens. Subsequently (Sec. 3.2), these tokens are mapped into either continuous or discrete representations to support Diffusion Transformers or autoregressive models, respectively. Finally (Sec. 3.3), a decoder reconstructs the original pixels from these latents. This process is guided by a Representation Alignment (REPA) module, which explicitly aligns semantics to preserve the latent topological structure, thereby completely replacing the traditional KL-divergence constraint.

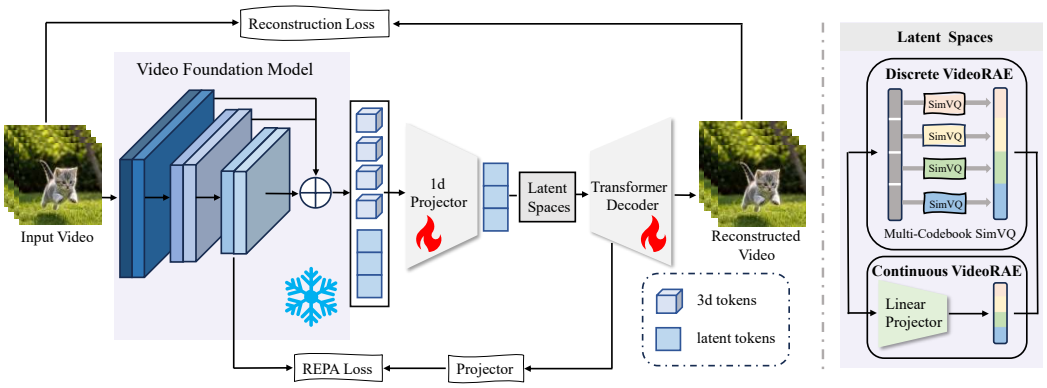


Figure 2: **Overall architecture of VideoRAE.** Given an input video, multi-scale hierarchical features are first extracted from a frozen VFM and fused, followed by a 1D self-attention projector that dynamically compresses them into compact base tokens. These tokens are then formatted into either a continuous 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.

3.1 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:

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

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 projector, denoted as $\mathcal{E}_{\text{proj}}$, to dynamically condense the long spatio-temporal sequence into a highly compact sequence of base tokens:

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

where N_{latent} is the length of the compressed latent sequence. By default, $\mathcal{E}_{\text{proj}}$ acts as a 1D projector that treats the VFM features as a flattened 1D sequence. Specifically, it concatenates a fixed number (N_{latent}) of learnable tokens to the input and employs self-attention mechanisms to aggregate global spatio-temporal information into these tokens. The updated learnable tokens are then extracted to form the compressed latent sequence Z_{base}. Additionally, we explore a 3D projector variant for continuous VideoRAE. It preserves spatio-temporal structure via 3D convolutions for local compression and self-attention for global context. The 1D projector remains the default unless otherwise specified. The large-scale T2V study uses the 3D variant: a 16-frame clip at 256×256 resolution is mapped to a $64 \times 4 \times 16 \times 16$ latent grid (channels, time, height, width). Flattening this grid yields 1024 visual tokens of dimension 64, without additional spatial patching. This preserves an explicit spatio-temporal interface for the downstream diffusion backbone.

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. Generative architectures like Diffusion Transformers (DiTs) prefer a smooth, continuous latent space. Traditional 3D-VAEs rely heavily on KL-divergence penalties to explicitly regularize latents into a Gaussian distribution. However, this rigid

constraint restricts expressive capacity and causes severe optimization conflicts with pixel reconstruction. In contrast, VideoRAE completely bypasses KL-divergence. By inheriting the structured semantic topology from the VFM encoder and implicitly regularizing representations via our Representation Alignment (REPA) module, the latent space intrinsically remains on a well-behaved semantic manifold. Consequently, without imposing artificial distribution assumptions, a simple linear layer projects Z_{base} to yield a highly expressive, 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: $Z_{\text{base}} = [Z_1, Z_2, \ldots, Z_K]$, 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, unlike standard VQ, we initialize a frozen base codebook E_k and map 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 $\mathcal{C}_k$:

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

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

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 VFM's semantic integrity. Since VFM features are inherently complex, forcing them into a low-dimensional bottleneck inevitably discards critical context and physical dynamics. By sustaining a high overall dimensionality, our multi-codebook mechanism readily accommodates this rich information, effectively preventing codebook collapse while providing ample semantic guidance for the subsequent decoding process under REPA constraints.

3.3 Decoder and Representation Alignment

The Transformer-based decoder $\mathcal{D}_{\text{sem}}$ reconstructs original video pixels from the highly compressed latents. It concatenates the 1D latent sequence with 1024 learnable 3D tokens. After processing these through stacked self-attention blocks, the 3D tokens are separated and mapped back into 3D feature maps via a depatchify operation for pixel reconstruction.

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

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

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

$$\mathcal{L}_{\text{REPA}} = \mathcal{L}_{\text{local}} + \lambda \mathcal{L}_{\text{global}} \tag{5}$$

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:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{adv}} + \lambda_{\text{repa}}\mathcal{L}_{\text{REPA}} \; (+\mathcal{L}_{\text{vq}}) \tag{6}$$

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.

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 16×256×256 video clips for both training and evaluation; the main text-to-video (T2V) study uses a large-scale in-house video–text corpus, while the earlier 2B study on VideoUFO (Wang & Yang, 2026) is retained in Appendix H. 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. We employ frozen V-JEPA 2 (Assran et al., 2025), V-JEPA 2.1 (Mur-Labadia et al., 2026), and VideoMAEv2 (Wang et al., 2023) as teacher models to extract multi-level features. Our projector compresses these features into a sequence of 512 or 1024 latent tokens. For the continuous space, we configure a 32 or 64-dimensional bottleneck, and additionally explore a 3D variant to align with mainstream 3D VAEs. For the discrete space, we apply Multi-Codebook SimVQ ($K = 4$, $V = 4096$). To evaluate generation performance, we train standard DiTs on the continuous latents and LLaMA-style AR models (Sun et al., 2024) on the discrete latents. Complete training configurations, downsampling details, and hardware specifics are provided in the Appendix.

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

Method	#rToken	PSNR↑	rFVD↓
TATS	1024	-	162
OmniTokenizer	5120	29.25	28
LARP-L-Long	1024	27.88	35
SweetTok	1280	29.27	20
VideoRAE(VideoMAEv2)	1024	**29.94**	<u>16</u>
VideoRAE(V-JEPA 2)	1024	<u>29.39</u>	**13**

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

Method	Config	PSNR↑	rFVD↓
CV-VAE	4096×4	30.11	57
LTX-VAE	128×128	**32.02**	35
LTX2	128×128	28.88	15
LeanVAE	4096×4	30.99	**10**
VideoRAE(VideoMAEv2)	512×32	<u>31.23</u>	13
VideoRAE(V-JEPA 2)	512×32	30.40	<u>14</u>

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. In the main text, we only present a portion of the reconstruction comparison results on UCF-101; the complete experimental results (including the 3D projector variants) and detailed analysis are provided in the appendix. As shown in Table 1, discrete VideoRAE establishes new state-of-the-art benchmarks. VideoRAE(V-JEPA 2) achieves an exceptional rFVD of 13, outperforming baselines like LARP (Wang et al., 2024a) and SweetTok (Tan et al., 2025), while VideoRAE(VideoMAEv2) secures the highest PSNR. For continuous representations (Table 2), VideoRAE effectively balances spatial recovery and perceptual quality. While LTX-VAE (HaCohen et al., 2024) achieves high PSNR but sub-optimal spatio-temporal consistency (rFVD 35), and LeanVAE (Cheng & Yuan, 2025) relies on massive latent resolutions

for low rFVD, VideoRAE delivers top-tier perceptual quality with competitive PSNR using a highly compact 512×32 bottleneck. Overall, these results validate our core hypothesis: frozen VFM representations do not hinder reconstruction, but rather surpass traditional pixel-driven autoencoders in yielding exceptional perceptual quality.

Table 3: Class-conditional generation evaluation (Discrete AR) on UCF-101 (10K samples). VideoRAE(VideoMAEv2) is evaluated at 224×224 resolution, while all other methods are at 256×256.

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 (Continuous DiT) on UCF-101 (2K samples). VideoRAE(VideoMAEv2) uses 224×224 resolution; others use 256×256.

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**

Discrete-token class-conditional generation. We train autoregressive (AR) models on our tokenized latents to evaluate generative capability. As shown in Table 3, VideoRAE(V-JEPA 2) achieves a state-of-the-art gFVD of 40 on UCF-101. Compared to the previously leading method, SweetTok(Tan et al., 2025) (gFVD 65, 1.9B parameters, 1280 tokens), VideoRAE secures a decisive advantage using fewer tokens and a smaller generator. This success demonstrates that discrete tokens, empowered by our Multi-Codebook SimVQ and semantic alignment strategies, effectively preserve rich VFM semantics, rendering downstream AR prediction significantly more efficient and accurate.

Continuous-space class-conditional generation. We adopt Diffusion Transformers to evaluate the continuous latent space. As presented in Table 4, VideoRAE exhibits strong generative capabilities. While recent continuous VAEs like LeanVAE(Cheng & Yuan, 2025) and LTX-VAE(HaCohen et al., 2024) improve the baseline gFVD to 164 and 161, VideoRAE further reduces it to 99 (VideoMAEv2) and 93 (V-JEPA 2). Additionally, we extended industry-grade models (WAN2.1-VAE(Wan et al., 2025) and CogVideoX(Yang et al., 2025)) to UCF-101 class-to-video generation. Although they outperform existing methods like Deco-VAE(Yin et al., 2025), VideoRAE and its 3D variants still deliver distinct superiority in generation performance. This substantial improvement indicates that VideoRAE's KL-free, semantic-aligned continuous latent space provides a structurally sound and generation-friendly manifold for diffusion models, facilitating high-fidelity video synthesis.

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. However, 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 generative tasks. We attribute this discrepancy to their distinct pre-training paradigms. VideoMAEv2 is driven by masked pixel reconstruction, heavily focusing on fine-grained local motion details. In contrast, V-JEPA 2 utilizes latent-space prediction, aligning its representations more with high-level semantics and macroscopic spatio-temporal dynamics. Although this deep semantic abstraction slightly compromises absolute pixel-level 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 high-quality video generation, endowing the latent space with stronger semantic perception is substantially more beneficial than pursuing rigid low-level pixel fidelity.

Accelerated Convergence in Generative Training. To illustrate our semantic latent space's efficiency, we compare VideoRAE's convergence speed against the discrete baseline LARP. As shown in Figure 3, an AR model trained on LARP requires 2000 epochs to reach a gFVD of 136. Con-

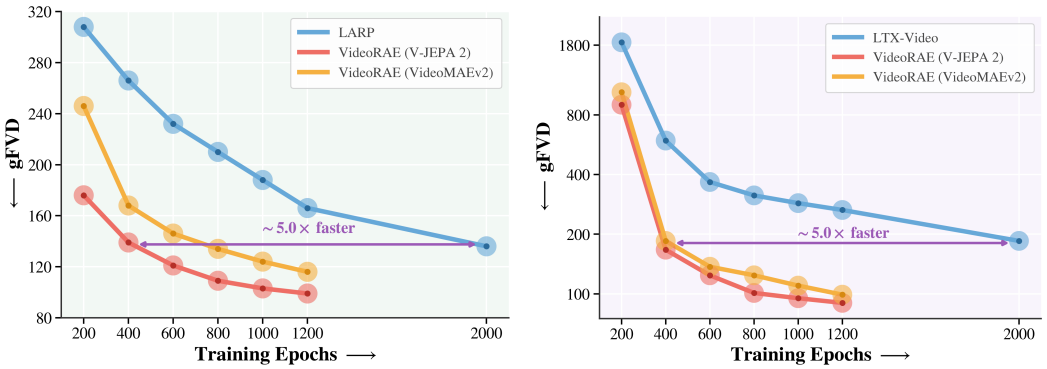


Figure 3: Training convergence speed comparison in AR (left) and DiT (right) generative modeling.

versely, VideoRAE achieves a comparable gFVD of 139 in merely 400 epochs, yielding an impressive $\sim 5.0\times$ speedup. A similar acceleration trend appears in the continuous space compared to LTX-VAE. By alleviating the burden of learning low-level spatio-temporal dynamics from scratch, VideoRAE enables generators to focus on high-level sequence transitions. This confirms that a semantically aligned latent space is pivotal for drastically reducing generative training costs.

Table 5: VBench results (200k steps, 4B MMDiT from scratch). Scores are percentages (↑). Selected dimensions are shown; full 16 dimensions are in Appendix G.1.

Representation	Total Score	Quality Score	Semantic Score	Dynamic Degree	Human Action	Color	Spatial Relation	Scene
Wan2.1-VAE	68.24	76.47	35.32	46.67	30.80	74.08	9.39	9.35
VideoRAE-3D (Ours)	**69.07**	**77.14**	**36.78**	**79.17**	**35.00**	**78.46**	**12.58**	**10.67**

Text-to-Video Generation on VBench. We compare frozen VideoRAE-3D and Wan2.1-VAE (Wan et al., 2025) using randomly initialized, identically configured MMDiT backbones within OpenSora2 (Zheng et al., 2025c). Each ∼3.64B parameter backbone features a hidden size of 2048, 16 heads, 14 dual-stream, and 28 single-stream blocks. Both employ latent flow matching alongside frozen T5 and CLIP encoders. Training runs for 200k steps (batch size 128) on an internal video-text corpus at 256×256 and 8 FPS. To respect their native interfaces, we use 16 frames (1024 tokens) for VideoRAE-3D and 17 frames (1280 tokens) for Wan2.1-VAE. Thus, the comparison matches training steps and batch sizes, though FLOPs differ. Full settings are in Appendix G.

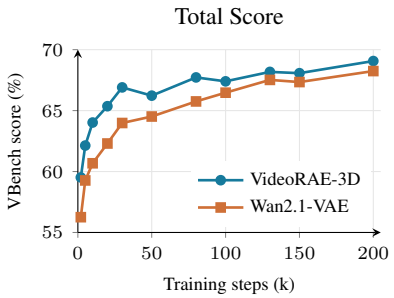


Figure 4: VBench total score during 4B-scale T2V training.

Figure 4 shows higher total scores for VideoRAE-3D at all 11 evaluated checkpoints. At 200k steps (Table 5), it reaches 69.07 versus 68.24 total score, 77.14 versus 76.47 quality score, and 36.78 versus 35.32 semantic score. The strongest gains include dynamic degree (79.17 versus 46.67), human action (35.00 versus 30.80), color (78.46 versus 74.08), and spatial relationships (12.58 versus 9.39). The full dimensional breakdown is reported in Appendix G.1. These results support stronger aggregate performance within the evaluated training budget; they do not establish convergence, uniform gains across dimensions, or a wall-clock speedup. The earlier 2B study is retained in Appendix H as supplementary evidence under a separate configuration.

Separately, we investigate whether VideoRAE can be integrated into an existing large-scale text-to-video model. Initializing from an 11B OpenSora2 checkpoint, we replace its autoencoder with VideoRAE-3D and fine-tune on 5M video-text pairs. VideoRAE-3D's grid-structured latent interface ensures seamless integration with the pretrained diffusion backbone. Qualitative results are shown in Figure 5. These high-quality generations demonstrate VideoRAE's effectiveness for large-scale text-to-video synthesis, complementing our from-scratch 4B-scale VBench evaluation.

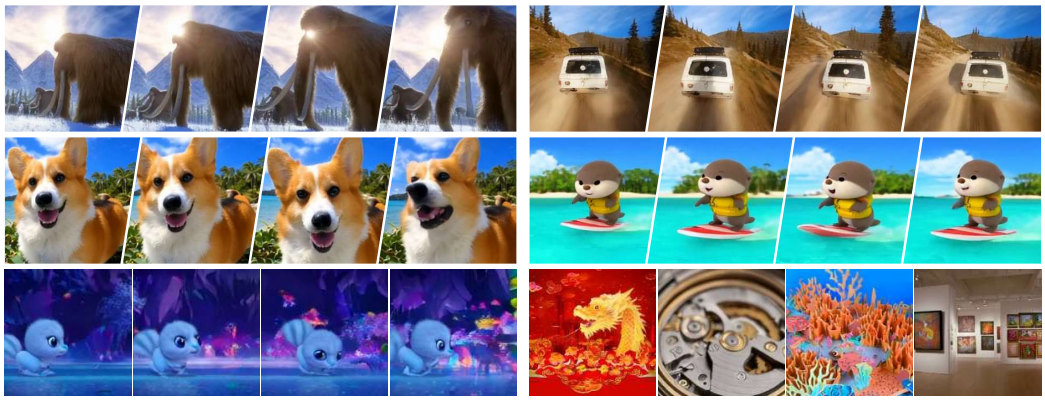


Figure 5: Text-to-video samples generated by an 11B OpenSora2 model adapted to VideoRAE-3D.

4.2 Ablation Study

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 6(a), the integration of REPA consistently yields significant performance improvements across rFVD and gFVD 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. This demonstrates that aligning the decoder's features with the VFM teacher at both local and global scales successfully regularizes the latent space topology, rendering this high-quality latent distribution more conducive to downstream generation tasks.

Multi-Scale Feature Aggregation. We investigate aggregating features from different levels of the VFM teacher. As Table 6(b) shows, relying solely on deep layers yields sub-optimal reconstruction, as they capture high-level semantics but discard fine spatial details. Progressively incorporating shallower layers steadily improves reconstruction metrics. Ultimately, utilizing Layers 8–24 achieves the optimal trade-off between reconstruction and generation. This highlights the necessity of multi-scale aggregation to simultaneously capture macro-semantics and micro-textures.

Other Ablations. We also investigate the impact of quantization and feature fusion strategies. Multi-codebook SimVQ effectively expands capacity without optimization issues, while element-wise addition ensures highly efficient fusion. Detailed tables and analyses are in the Appendix.

Table 6: Ablation study on Representation Alignment and Multi-scale Feature Aggregation.

(a) Effect of REPA

Latent Space	REPA	PSNR↑	rFVD↓	gFVD↓
Continuous	w/o	**30.57**	18	105
Continuous	w/	30.40	**14**	**93**
Discrete	w/o	**29.71**	15	67
Discrete	w/	29.39	**13**	**40**

(b) Multi-scale Aggregation

Aggregated Layers	PSNR↑	rFVD↓	gFVD↓
Layer 24	26.59	22	51
Layers 16–24	28.35	18	41
Layers 12–24	29.11	15	42
Layers 8–24	**29.39**	**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 (VFMs) 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.

AI USE STATEMENT

We used large language models to assist with translation, drafting and revising prose, condensing text, and LaTeX formatting. The authors reviewed the AI-assisted content for accuracy and consistency with the underlying experiments and sources and take full responsibility for the final manuscript.

REPRODUCIBILITY STATEMENT

We describe the VideoRAE architecture and training objectives in Section 3, and provide implementation details for the tokenizer, autoregressive generator, and diffusion models in Appendix A. Dataset and evaluation settings are described in Section 4, with baseline descriptions in Appendix B and additional experimental results in the appendices. For the main 4B-scale text-to-video study, Appendix G documents data preprocessing, latent interfaces, backbone configurations, optimization settings, random seeds, and the VBench evaluation protocol; Appendix G.1 reports all 16 evaluation dimensions. The separate 2B study is described in Appendix H. We will release code and model checkpoints to facilitate independent evaluation and further research.

REFERENCES

Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, 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 pretraining 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.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: 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, Feiwu 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. Vidtwin: Video vae with decoupled structure and dynamics. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 22922–22932, 2025.

Tianze Xia, Lijun Zhou, Kaixin Xiong, Jingfeng Yao, Yu Zhu, Zhenxin Zhu, Bing Wang, Guang Chen, Hangjun Ye, Wenyu Liu, et al. Reworld: Learning better representations for world action models. *arXiv preprint arXiv:2606.27504*, 2026.

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.

A Additional Implementation Details

A.1 Implementation Details of the VideoRAE Tokenizer

VideoRAE leverages a pretrained and strictly frozen visual foundation model (VFM)—specifically VideoMAEv2, V-JEPA 2, or V-JEPA 2.1—as the core feature extractor. We extract hierarchical intermediate features from multiple layers of the corresponding VFM and aggregate them using element-wise addition. The resulting spatio-temporal feature sequence is subsequently compressed by a lightweight 1D self-attention projector. This module utilizes learnable latent tokens to aggregate global context, producing a highly compact sequence of either 512 or 1024 latent tokens. Additionally, to align with mainstream 3D VAEs, we implement a 3D projector variant. This variant first performs local spatio-temporal compression via 3D convolutions and then models global context with self-attention, yielding a latent shape of $16 \times 16 \times 4 \times 64$. For the continuous VideoRAE, these compressed tokens are linearly projected to a bottleneck of 32 or 64 channels, omitting any KL-divergence regularization term. For the discrete variant, we adopt the Multi-Codebook SimVQ strategy, configuring $K = 4$ subspaces with a vocabulary size of $V = 4096$ per sub-codebook. The decoder is built upon a Transformer-based architecture, reconstructing pixel-level video frames by concatenating the latent representations with 1024 learnable 3D tokens. The overall training objective is a weighted sum of several components: an ℓ_1 reconstruction loss (weight 1.0), an LPIPS perceptual loss (weight 1.0), an adversarial loss (weight 0.3), and our proposed Representation Alignment (REPA) loss (weight 0.2). The REPA objective enforces both local and global semantic consistency between the decoder's intermediate features and the VFM teacher. For discrete models, we additionally incorporate a codebook loss and a commitment loss, with weights set to 1.0 and 0.25, respectively.

For the autoencoder training phase, we utilize video clips with a spatial resolution of 256×256 and a temporal length of 16 frames, sourced from the combination of UCF-101 and Kinetics-600 datasets. During the data loading process, random horizontal flipping is applied as the data augmentation technique. In most of our experimental settings, the VideoRAE model is optimized for 75 epochs (amounting to approximately 500k training steps) using a global batch size of 64. We employ the Adam optimizer with a base learning rate of 1×10^{-4}, $\beta_1 = 0.9$, and $\beta_2 = 0.95$. The learning rate is regulated by a cosine annealing schedule paired with an initial warm-up phase. All autoencoder models are trained on NVIDIA H100 GPUs.

A.2 Implementation Details of the AR Generative Model

We use Llama-like Transformer architectures as the autoregressive (AR) generative models for discrete VideoRAE latents. Since the discrete VideoRAE representation is obtained with the Multi-Codebook SimVQ strategy, each latent token is represented by multiple codebook indices. Accordingly, our AR Transformer adopts a multi-head prediction structure, where each prediction head models one of the $K = 4$ codebooks. The predictions from all heads are jointly used to reconstruct the high-dimensional discrete latent representation. Following our general Transformer implementation, we use absolute learned positional embeddings. The token dropout, residual dropout, and feed-forward dropout probabilities are all set to 0.1.

Our default AR generative model contains approximately 1.3B parameters and is trained on the training split of UCF-101 for 2000 epochs with a batch size of 64. We use the AdamW optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.95$, a weight decay of 0.05, and a base learning rate of 6×10^{-4}, following a warm-up cosine learning-rate schedule. During inference, we apply classifier-free guidance (CFG) with a scale of 1.25 and do not use top-k or top-p sampling. For class-conditional generation, the model autoregressively predicts the multi-codebook discrete latent tokens conditioned on the class label.

A.3 Implementation Details of the Diffusion Model

For class-to-video generation, we use diffusion models that match the structure of the corresponding VideoRAE latent space. Specifically, the continuous VideoRAE with 1D latent tokens is modeled with a diffusion Transformer operating on 1D token sequences, whereas the VideoRAE-3D variant is modeled with a Latte-style spatio-temporal diffusion Transformer. Both models are trained on UCF-

101 for 3000 epochs with a base learning rate of 1×10^{-4}. During inference, we use classifier-free guidance with a scale of 2.0.

For open-domain T2V generation, the main 4B-scale from-scratch study is detailed in Appendix G; the earlier 2B study is reported separately in Appendix H. In the independent large-scale adaptation experiment, we replace the original autoencoder of an 11B-parameter OpenSora2 model with VideoRAE-3D and fine-tune on approximately 5M video–text pairs.

B Detailed Description of 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. The baselines are categorized by downstream tasks as follows:

- **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.** For the main VBench study, we compare frozen VideoRAE-3D with frozen Wan2.1-VAE (Wan et al., 2025), training the same 4B-scale MMDiT configuration from random initialization for each representation. The sequence lengths and native frame counts differ as documented in Appendix G.

C Reconstruction Performance of VideoRAE

C.1 Detailed Reconstruction performance of Discrete VideoRAE

Table 7 presents the complete reconstruction results of discrete video tokenizers on UCF-101 and TokenBench, supplementing the results reported in the main text. VideoRAE achieves strong reconstruction performance on both datasets: the VideoMAEv2-based variant obtains the highest PSNR among the compared methods, while the V-JEPA 2-based variant achieves the lowest rFVD. These results demonstrate that representations from frozen video foundation models retain rich information for video reconstruction. Despite compression and discretization, these representations support the recovery of spatial details and overall scene structure, providing an effective foundation for high-fidelity video tokenization.

Figure 6 further supports these quantitative results. Compared with LARP, VideoRAE reconstructs clearer structural details, as illustrated by the enlarged basketball backboard regions, while preserving the scene layout and subject appearance across the displayed frames.

C.2 Detailed Reconstruction performance of Continuous VideoRAE

Table 8 presents the complete reconstruction results of continuous video autoencoders on UCF-101 and TokenBench, including additional baselines and the 3D projector variants. VideoRAE

Table 7: Reconstruction performance of discrete tokenizers on UCF-101 and TokenBench.

Method	#rToken	UCF-101			TokenBench-256×256		
		PSNR↑	LPIPS↓	rFVD↓	PSNR↑	LPIPS↓	rFVD↓
AR generative models with discrete video tokenizers							
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**	0.10	16	**30.69**	**0.09**	33
VideoRAE(V-JEPA 2)	1024	29.39	0.10	**13**	29.93	0.10	**28**



Figure 6: Visualization comparison of the Discrete-token reconstruction on the UCF-101 dataset.

delivers highly competitive performance, balancing spatial detail recovery and perceptual quality. Compared with LTX-VAE, VideoRAE achieves substantially lower rFVD on both datasets using a compact 512×32 bottleneck. Compared with LeanVAE, the VideoMAEv2-based variant achieves higher PSNR and lower LPIPS with substantially fewer latent positions, while maintaining competitive rFVD. Even against industry-grade autoencoders such as WAN2.1-VAE and CogVideoX-VAE, the larger VideoRAE configurations remain competitive in rFVD, despite being trained only on Kinetics-600 and UCF-101. Moreover, the 3D projector variants achieve reconstruction quality close to their 1D counterparts, demonstrating the effectiveness of VFM-based representations across different latent architectures.

Figure 7 provides further qualitative support: VideoRAE preserves the overall scene structure and recovers clearer background lettering than CV-VAE in the displayed billiards sequence. Together, these results demonstrate that frozen video foundation representations support high-quality reconstruction in both continuous and discrete latent spaces, with competitive spatial fidelity and perceptual quality.

Table 8: Reconstruction performance of continuous autoencoders on UCF-101 and TokenBench.

Method	Config	UCF-101			TokenBench-256×256		
		PSNR↑	LPIPS↓	rFVD↓	PSNR↑	LPIPS↓	rFVD↓
Diffusion generative models with continuous video tokenizers							
CV-VAE	32×32×4×4	30.11	0.12	57	30.80	0.11	61
LTX-VAE	8×8×2×128	**32.02**	0.10	35	**33.74**	**0.08**	32
LTX2	8×8×2×128	28.88	0.12	15	29.36	0.11	36
LeanVAE	32×32×4×4	30.99	0.11	**10**	31.79	0.10	**23**
VideoRAE(VideoMAEv2)	512×32	31.23	**0.08**	13	32.06	**0.08**	25
VideoRAE(V-JEPA 2)	512×32	30.40	0.09	14	31.19	0.09	25
CogvideoX-VAE	32×32×4×16	**36.22**	**0.05**	7	36.35	**0.04**	9
WAN2.1-VAE	32×32×4×16	35.70	**0.05**	**4**	**37.02**	**0.04**	**6**
VideoRAE(VideoMAEv2)	1024×64	33.64	0.06	5	34.25	0.07	11
VideoRAE(V-JEPA 2)	1024×64	32.14	0.07	7	32.56	0.08	13
VideoRAE-3D(VideoMAEv2)	16×16×4×64	33.50	0.06	5	34.07	0.07	12
VideoRAE-3D(V-JEPA 2)	16×16×4×64	31.91	0.07	7	32.48	0.08	15
VideoRAE-3D(V-JEPA 2.1)	16×16×4×64	31.88	0.07	7	32.19	0.07	16

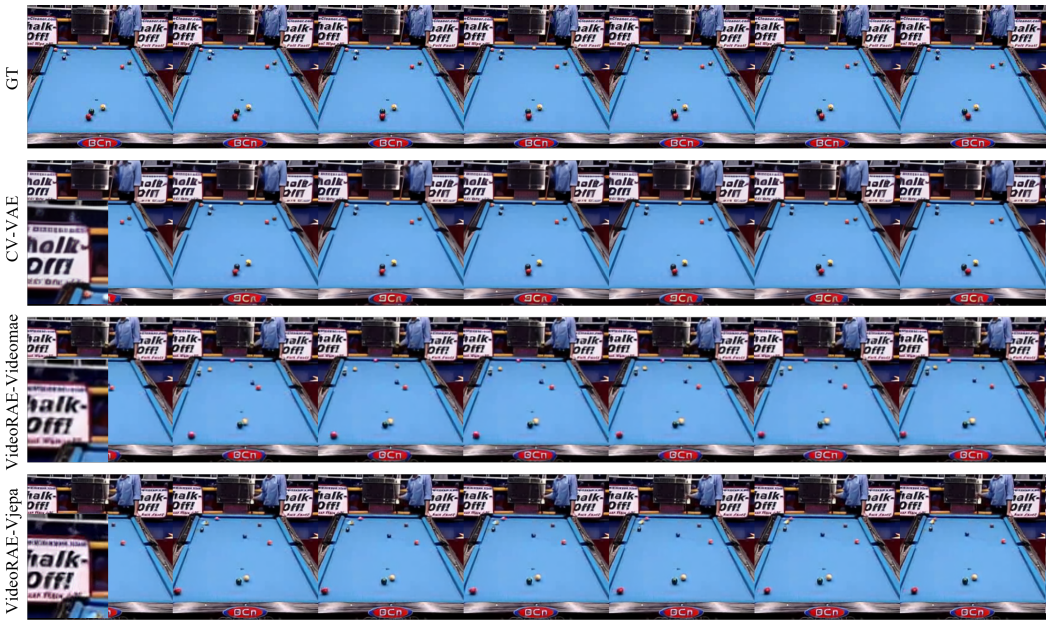


Figure 7: Visualization comparison of the Continuous-space reconstruction on the UCF-101 dataset.

D CLASS-TO-VIDEO GENERATION PERFORMANCE OF VIDEORAE

D.1 CLASS-TO-VIDEO GENERATION PERFORMANCE IN DISCRETE-SPACE

We provide qualitative results of discrete class-to-video generation on UCF-101 in Figure 8. Compared with OmniTokenizer, VideoRAE produces videos with better visual quality, clearer spatial details, and more coherent scene structure, further supporting the quantitative results reported in the main text.

Table 9: Class-conditional generation evaluation (Discrete AR) on UCF-101, evaluated on 10K generated samples.

Tokenizer	**Resolution**	**#Tokens**	**#Params**	**gFVD** ↓
CogVideo (Hong et al., 2022)	256×256	6800	9.4B	626
TATS (Ge et al., 2022)	256×256	4096	321M	332
Video-LaVIT (Jin et al., 2024)	256×256	512	7B	280
OmniTokenizer (Wang et al., 2024b)	256×256	5120	650M	191
LARP-L (Wang et al., 2024a)	256×256	1024	632M	99
SweetTok (Tan et al., 2025)	256×256	1280	650M	84
SweetTok (Tan et al., 2025)	256×256	1280	1.9B	65
VideoRAE(VideoMAEv2)	224×224	1024	1.3B	<u>45</u>
VideoRAE(V-JEPA 2)	256×256	1024	1.3B	**40**

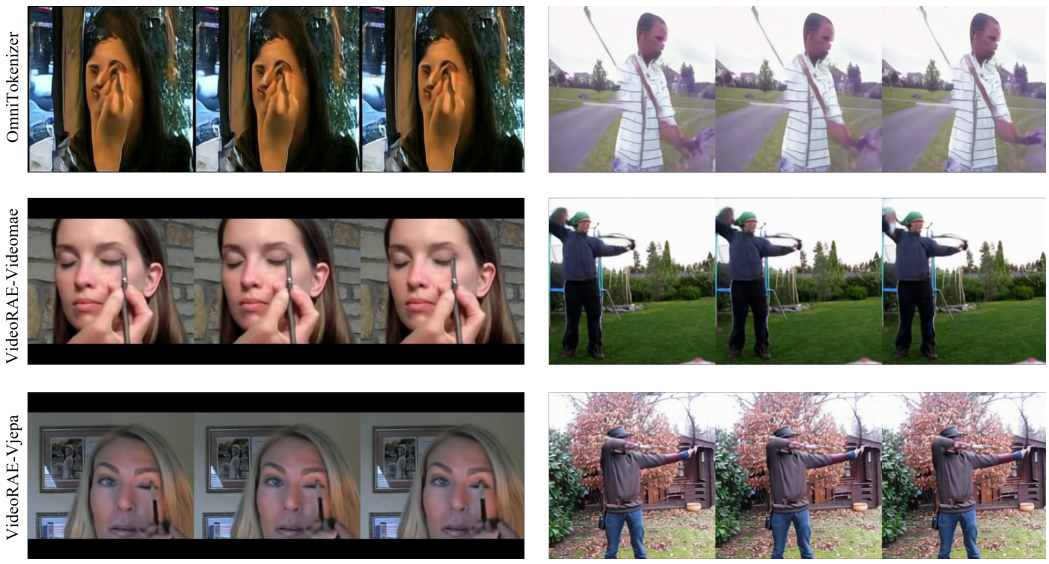


Figure 8: Visualization of the discrete-token generation results on the UCF-101 dataset.

D.2 CLASS-TO-VIDEO GENERATION PERFORMANCE IN CONTINUOUS-SPACE

Tables 10 and 11 present the complete continuous class-to-video generation results of VideoRAE and its 3D projector variants on UCF-101. VideoRAE with V-JEPA 2 achieves gFVD scores of 93 and 94 under the 512×32 and 1024×64 configurations, respectively, outperforming the compared VAE-based baselines. The 3D variants also deliver strong performance, with VideoRAE-3D using V-JEPA 2.1 achieving a gFVD of 93. These results demonstrate that frozen video foundation representations provide effective latent spaces for diffusion-based video generation across both 1D and 3D architectures. Figure 9 further provides a qualitative comparison with LTX-VAE, showing that VideoRAE generates videos with clearer spatial details and better overall visual quality, consistent with the quantitative improvements.

Table 10: Class-conditional generation evaluation (Continuous DiT) on UCF-101.

Method	Resolution	Configs	gFVD ↓
Latte (Ma et al., 2024)	256×256	4096×4	478
LeanVAE (Cheng & Yuan, 2025)	256×256	4096×4	164
LTX-VAE (HaCohen et al., 2024)	256×256	128×128	161
VideoRAE(VideoMAEv2)	224×224	512×32	<u>99</u>
VideoRAE(V-JEPA 2)	256×256	512×32	**93**
WF-VAE (Li et al., 2025)	256×256	4096×16	371
LeanVAE (Cheng & Yuan, 2025)	256×256	4096×16	175
DeCo-VAE (Yin et al., 2025)	256×256	4096×16	166
CogvideoX-VAE (Yang et al., 2025)	256×256	4096×16	151
WAN2.1-VAE (Wan et al., 2025)	256×256	4096×16	126
VideoRAE(VideoMAEv2)	224×224	1024×64	<u>103</u>
VideoRAE(V-JEPA 2)	256×256	1024×64	**94**

Table 11: Class-conditional generation evaluation (3D projector variants) on UCF-101.

Method	Resolution	Configs	gFVD ↓
VideoRAE-3D(VideoMAEv2)	224×224	1024×64	104
VideoRAE-3D(V-JEPA 2)	256×256	1024×64	<u>96</u>
VideoRAE-3D(V-JEPA 2.1)	256×256	1024×64	**93**

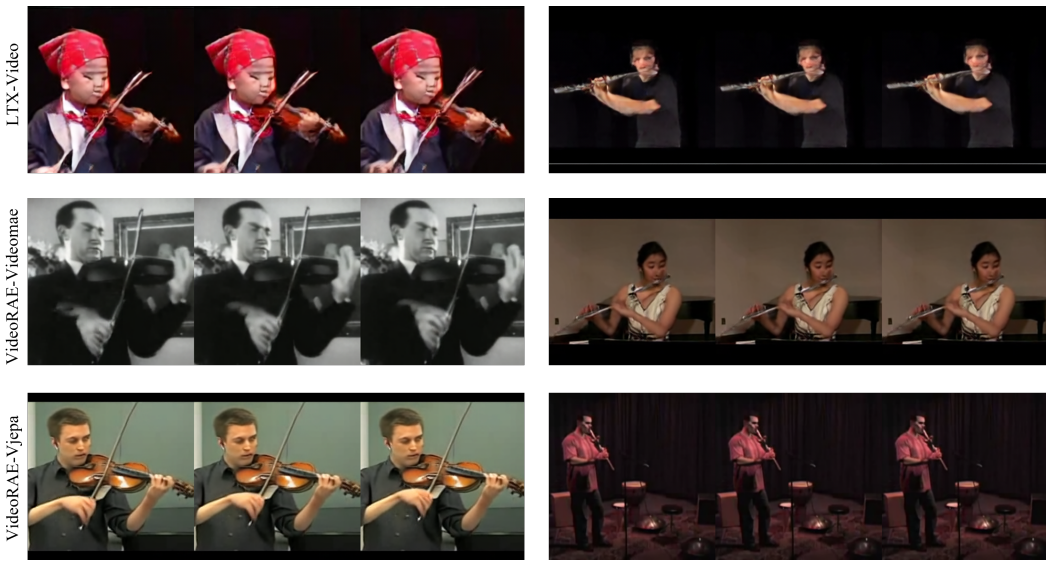


Figure 9: Visualization of the continuous-space generation results on the UCF-101 dataset.

E More Text to Video Examples of VideoRAE

Figure 10 presents additional text-to-video generation examples of VideoRAE. The generated videos exhibit high visual quality and consistent scene structure across the displayed frames, while capturing diverse content described by the text prompts. These examples further demonstrate the effectiveness of VideoRAE as a latent representation for text-conditioned video generation.

Prompt: A beautiful silhouette animation shows a wolf howling at the moon, feeling lonely, until it finds its pack.



Prompt: A yellow tram moves through a busy European square during light snowfall. Pedestrians cross behind it, warm windows glow against the blue evening, and the camera slowly pans to follow the tram past ornate historic buildings.



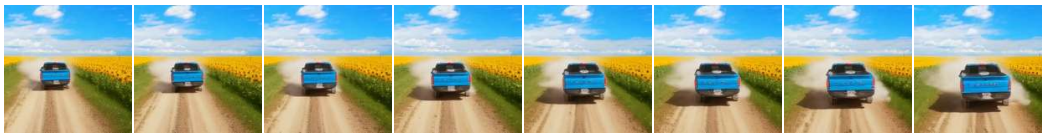
Prompt: A tiny astronaut walks across the frosted surface of a chocolate cake as if exploring an alien planet. Crumbs rise beneath each boot, birthday candles glow like distant towers, and the macro camera follows at ground level with cinematic shallow depth of field.



Prompt: A bright orange kayak moves through gentle rapids between mossy rocks. The paddler plants the blade and turns around a boulder while the camera follows just above the water, catching splashes and believable changes in speed.



Prompt: A small blue pickup truck drives down a dusty road between wide sunflower fields. Dust curls behind its tires while the camera leads the vehicle from a stable low angle and summer clouds drift across a vivid open sky.



Prompt: A wooden sailboat cuts through dark blue water under a steady ocean breeze. Its white sails fill and ropes pull taut while an aerial camera circles slowly, revealing a rugged island and a trail of sparkling wake behind the boat.



Prompt: A cook pours bright tomato sauce over handmade pasta in a black pan and tosses it once. Steam rises into warm window light while basil leaves and grated cheese shift naturally across the glossy noodles.



Figure 10: Additional text-to-video generation examples of VideoRAE.

F Detailed Ablation Study

F.1 Quantization Strategies.

We compare four quantization methods for VideoRAE: VQ, SimVQ (Zhu et al., 2025), MCQ, and Multi-Codebook SimVQ. As shown in Table 12, standard VQ yields the poorest reconstruction and generation performance, suggesting that the tested VQ configuration is insufficient to preserve the information in VFM representations. SimVQ, with a quantization dimension of 128, improves both

reconstruction and generation quality, while MCQ provides further gains over the single-codebook configuration. We attribute the latter improvement to the increased representational capacity provided by multiple codebooks, without requiring a single large codebook. Combining the multi-codebook design with SimVQ achieves the best overall results among the evaluated configurations, with an rFVD of 13 and a gFVD of 40. These results support the effectiveness of Multi-Codebook SimVQ in preserving reconstruction quality and providing useful discrete representations for downstream autoregressive generation.

Table 12: Ablation study on quantization strategies.

Quantization Method	**PSNR↑**	**SSIM↑**	**LPIPS↓**	**rFVD↓**	**gFVD↓**
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**

F.2 Representation Alignment

To evaluate the effectiveness of our Representation Alignment (REPA) module, we train both continuous and discrete models with and without the REPA objective. As shown in Table 13, incorporating REPA significantly improves semantic fidelity and generation capabilities in both latent spaces, despite a marginal and expected drop in rigid pixel-level metrics like PSNR. Specifically, REPA reduces the gFVD from 105 to 93 in the continuous setting, and from 67 to 40 in the discrete setting, while also improving perceptual reconstruction (rFVD). Furthermore, we explore the traditional KL penalty as an alternative regularizer in the continuous setting (*Continuous-kl*). Compared to the unregularized baseline, applying KL severely degrades reconstruction quality (e.g., PSNR drops from 30.57 to 27.45, rFVD worsens to 31) and offers limited generation improvement (gFVD 101). This demonstrates that explicitly aligning decoder features with the VFM teacher inherently acts as a superior semantic constraint. It successfully regularizes the latent space topology without relying on any KL penalty, maintaining high reconstruction quality while rendering a distribution highly conducive to downstream generation.

Table 13: Ablation study on the effect of REPA for semantic alignment.

Latent Space	**REPA**	**PSNR↑**	**SSIM↑**	**LPIPS↓**	**rFVD↓**	**gFVD↓**
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	0.09	14	93
Discrete	w/o	29.71	0.90	0.10	15	67
Discrete	w/	29.39	0.91	0.10	13	40

F.3 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 14, 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.

Table 14: Ablation study on multi-scale feature aggregation.

Aggregated Layers	PSNR↑	SSIM↑	LPIPS↓	rFVD↓	gFVD↓
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**

F.4 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 15. 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 15: Ablation study on feature fusion strategies.

Fusion Strategy	PSNR↑	SSIM↑	LPIPS↓	rFVD↓	gFVD↓
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

G LARGE-SCALE TEXT-TO-VIDEO STUDY

Scope and data. Only the diffusion backbone is trained from scratch; the video tokenizers and both text encoders use pretrained weights and remain frozen. We use a large-scale in-house video–text corpus. The 200k-step budget with global batch size 128 corresponds to 25.6M sample presentations, which is not the number of unique videos. Clips are sampled from random temporal windows at 8 FPS, resized while preserving aspect ratio, center-cropped to 256×256, and normalized to $[-1, 1]$. Invalid videos, clips that are too short, sources below 8 FPS, and records without valid captions are excluded. Data shards and records within each shard are shuffled. Captions may include quality and frame-rate conditioning.

Table 16: Representation interfaces in the main T2V study. Latent shapes are ordered as channels $\times$ time $\times$ height $\times$ width.

	VideoRAE-3D	Wan2.1-VAE
Tokenizer status	Frozen	Frozen
Temporal interface	Non-causal	Causal
Input frames / FPS	16 / 8	17 / 8
Input resolution	256×256	256×256
Latent shape	$64 \times 4 \times 16 \times 16$	$16 \times 5 \times 32 \times 32$
Spatial patch size	1×1	2×2
Visual token dimension	64	64
Visual sequence length	1024	1280

Backbone and optimization. Both runs use a FLUX-style MMDiT in the OpenSora2 training framework, with approximately 3.64B trainable parameters, hidden size 2048, 16 attention heads, 14 dual-stream blocks, 28 single-stream blocks, and MLP ratio 4. Frozen T5-v1.1-XXL and CLIP ViT-L/14 text encoders use maximum sequence lengths of 512 and 77, respectively. We train with latent flow matching and velocity-prediction MSE. Time is sampled from a logit-normal distribution, $t = \mathrm{sigmoid}(z)$ with $z \sim \mathcal{N}(0, 1)$, and transformed with flow shift 2.0. The optimizer is HybridAdam in AdamW mode with learning rate 5×10^{-5}, weight decay 0, epsilon 10^{-15}, and 1000 warmup steps without cosine decay. Gradient clipping is 1.0. Training uses 32 GPUs, batch size 4 per GPU, and one gradient-accumulation step (global batch 128). The DiT uses BF16 and the video tokenizer FP32; ZeRO-2 and gradient checkpointing are enabled. We use no EMA, set the random seed to 42, and apply condition dropout of 0.31622777 to each text encoder. VideoRAE-3D has a 20% shorter visual sequence, but neither matched-FLOP nor wall-clock efficiency is measured here.

VBench evaluation. At each recorded checkpoint, we sample 256×256 videos at 8 FPS using 50 sampling steps, CFG scale 3.0, flow shift 2.0, and per-sample seeds with base seed 42. Frame counts match the native representation interfaces in Table 16. For the usual dimensions, five videos are generated per prompt. For temporal flickering, 25 candidates per prompt undergo static-video filtering, and at most five retained videos are scored. We report all 16 VBench dimensions and the supplied normalized aggregate scores; total score is $0.8\,Q + 0.2\,S$. Curves use the 11 checkpoints recorded in the CSV: 2k, 5k, 10k, 20k, 30k, 50k, 80k, 100k, 130k, 150k, and 200k. All values are multiplied by 100 for presentation. The results describe one training run per representation and do not estimate variability across independent training seeds.

G.1 COMPLETE VBENCH RESULTS AT 200K

Table 17: Complete VBench results at 200k steps. Higher is better for every column. Δ is VideoRAE-3D minus Wan2.1-VAE in percentage points, computed before rounding. Bold marks the higher score.

Metric	Wan2.1-VAE	VideoRAE-3D	Δ
Total score	68.24	**69.07**	+0.82
Quality score	76.47	**77.14**	+0.66
Semantic score	35.32	**36.78**	+1.46
Subject consistency	**92.02**	89.45	-2.57
Background consistency	**95.97**	95.64	-0.34
Temporal flickering	**99.37**	99.36	-0.02
Motion smoothness	**98.47**	97.15	-1.31
Dynamic degree	46.67	**79.17**	+32.50
Aesthetic quality	41.61	**42.80**	+1.19
Imaging quality	**53.03**	47.91	-5.11
Object class	**28.31**	23.78	-4.53
Multiple objects	5.43	**6.94**	+1.51
Human action	30.80	**35.00**	+4.20
Color	74.08	**78.46**	+4.37
Spatial relationship	9.39	**12.58**	+3.19
Scene	9.35	**10.67**	+1.32
Temporal style	15.43	**16.06**	+0.64
Appearance style	**20.11**	19.71	-0.40
Overall consistency	16.66	**17.56**	+0.90

H EARLIER 2B TEXT-TO-VIDEO STUDY

The earlier study compares VideoRAE-1D with LTX-VAE using a 2B-parameter MMDiT with hidden size 2048, 16 heads, 8 dual-stream blocks, and 16 single-stream blocks, conditioned on T5 and CLIP embeddings. It was trained on VideoUFO (Wang & Yang, 2026) for 800k steps with global batch size 256. We retain its previously reported scores and curve as supplementary results. Because its data, architecture, training budget, and evaluation setup differ from the main 4B study, the absolute scores should not be used as a cross-study scaling comparison.

Table 18: VBench results comparing VideoRAE-1D with LTX-VAE in the earlier 2B text-to-video study.

Method	Total Score ↑	Quality Score ↑	Semantic Score ↑	Subject Cons. ↑	Background Cons. ↑	Imaging Qual. ↑	Object Class ↑	Human Action ↑	Scene ↑
LTX-VAE	69.35	74.03	50.67	84.14	95.51	43.54	58.51	54.50	23.29
VideoRAE-1D (Ours)	**71.26**	**76.09**	**51.92**	**91.27**	**97.02**	**46.02**	**59.07**	**60.80**	**27.57**

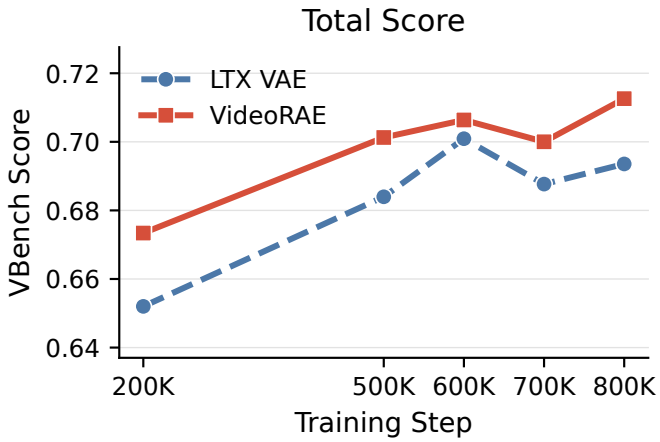


Figure 11: VBench total-score curves for VideoRAE-1D and LTX-VAE in the earlier 2B text-to-video study.