OneVAE: Joint Discrete and Continuous Optimization Helps Discrete Video VAE Train Better

Yupeng Zhou[1], Zhen Li[1], Yuming Chen[1], Ziheng Ouyang[1], Ruoyi Du[4], Daquan Zhou[5], Bin Fu[4], Yihao Liu[4], Peng Gao[4], Ming-Ming Cheng[1,2,3], and Qibin Hou[1,2,3⋆]

[1] VCIP, CS, Nankai University, China
[2] NKIARI, Shenzhen Futian, China
[3] AAIS, Nankai University, China
[4] Shanghai AI Laboratory, China
[5] Peking University, China

Abstract. Encoding videos into discrete tokens can enable concise and unified multi-modal LLMs. Previous discrete video VAEs suffer from unstable training, long training time, and degraded reconstruction quality. We revisit the relationship between continuous and discrete VAEs and find that bridging discrete and continuous representations improves discrete token learning. Based on this insight, we propose a unified progressive training framework that (i) jointly optimizes continuous and discrete reconstructions within a single network, and (ii) progressively derives a family of VAEs at different compression ratios, leading to faster convergence and better final performance. Furthermore, leveraging this unified training, a single VAE can achieve competitive performance for both continuous and discrete representations. Meanwhile, we propose two architectural improvements to further boost the performance of discrete VAEs. First, inspired by the use of enlarging continuous latent dimension to boost reconstruction, we propose multi-token quantization, improving PSNR by nearly 1 dB at the same token compression ratio. Second, we introduce first-frame enhancement, which uses a lower-compression first frame as an anchor in causal VAEs, alleviates the limited-context issue of the first frame and significantly improves reconstruction under high-compression settings (e.g., $4 \times 16 \times 16$). The code is available at `https://github.com/HVision-NKU/OneVAE`.

Keywords: Discrete VAE · Video VAE · Unified VAE

1 Introduction

Auto-regressive language LLMs [1] have shown strong performance, sparking interest in extending their capabilities to image and video generation. A key challenge for visual generation with LLMs is tokenization: unlike text, visual

⋆ Corresponding author.

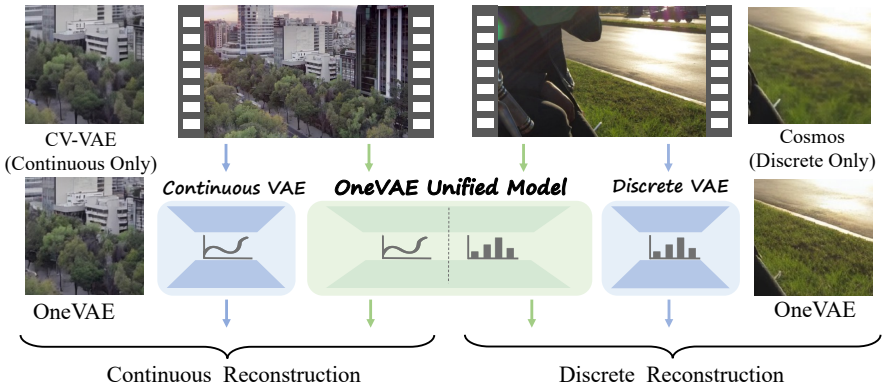

Fig. 1: Illustration of OneVAE, which encodes both discrete and continuous representations within a single model while achieving competitive performance, in contrast to traditional discrete-only [55] and continuous-only [94] VAEs.

signals (e.g., RGB pixels) are high-dimensional and complex, so it is difficult to represent them as a compact sequence of discrete tokens for generation modeling. Looking back at diffusion-based generation works, the pioneering work Stable Diffusion [56] introduces a continuous VAE to encode visual content into a compact continuous latent representation. However, continuous representation does not align well with LLMs for pure next-token prediction, as it is difficult to directly predict continuous tokens for LLM [36]. Therefore, to support the discrete modeling required by next-token prediction, prior autoregressive generative models [13, 62, 64] utilized VQ-VAEs [18] to encode visual content into discrete tokens, align with text tokens, and enable a unified next-token prediction framework. Extending such discrete tokenizers from images to videos is more demanding, since video tokenization must additionally compress the temporal dimension while preserving coherent motion. Early discrete video VAEs also rely on VQ-based quantization [74, 76, 83]. To mitigate VQ codebook collapse, Cosmos Tokenizer [55] adopts FSQ [46] for discretization. Overall, existing discrete VAEs remain imperfect: they suffer from degraded reconstruction quality and prolonged, unstable training, especially at high compression ratios, and still lag behind their continuous counterparts.

This discrepancy motivates us to further explore ways to enhance the performance of discrete VAEs in this paper. We first rethink the distinctions between continuous and discrete encoding. Based on how the features are processed, we categorize the quantization methods into two types: (1) codebook lookup-based ones, which aim to find the vector with the closest distance in the codebook; (2) rounding operation-based ones, which directly round on the mapped features. Unlike VQ [18], which uses codebook distance, FSQ [46] maps continuous features into a bounded range and rounds them, making it closer to continuous encoding. We verified this by training a lightweight quantizer in the latent space of continuous VAEs, optimizing only its parameters. As shown in Tab. 1, FSQ

outperforms VQ (PSNR 26.46 vs. 22.82). This confirms FSQ's closer relation to continuous VAEs, motivating us to explore whether continuous representations can improve discrete VAE training, rather than treating them as independent as in prior work [55].

Building on this insight, we propose OneVAE, a unified VAE training framework that couples continuous and discrete learning within a single model. Specifically, we introduce a joint discrete-continuous dual-path training strategy that interleaves continuous reconstruction with FSQ-based discrete reconstruction during optimization. The continuous path serves as a stable learning signal, improving the robustness of discrete tokenization training and leading to better reconstructions. Moreover, we develop a unified progressive training pipeline that learns tokenizers at multiple compression ratios by warm-starting from a low-compression continuous VAE and progressively extending it to higher-compression and discrete settings, instead of training each configuration from scratch. As shown in Fig. 1, the resulting unified VAE supports both continuous and discrete tokenization, bridging diffusion-style continuous latents and LLM-style discrete next-token prediction. Experiments validate the effectiveness of our framework across both representations, and ablations further show consistent gains under key metrics.

Given the characteristics and limitations of discrete causal VAEs, we also introduce two architectural enhancements, multi-token quantization and first-frame enhancement, to further improve discrete reconstruction. Multi-token quantization is inspired by high-compression continuous VAEs, where increasing the latent dimensionality improves capacity. Instead of assigning a single token to each latent position, we allocate multiple tokens along the channel dimension (two in our experiments), forming a compositional code with higher representational power. Experimental results show that it improves PSNR by nearly 1 dB without reducing the token compression ratio. First-frame enhancement addresses a common failure case in causal VAEs: under high compression, the first frame is often reconstructed worse than subsequent frames due to limited context. We therefore use a lower compression ratio for the first frame only and treat it as an anchor for reconstructing later frames, improving both first-frame and overall video reconstruction quality.

Our contributions are summarized as follows:

- We revisit the relationship between continuous and discrete tokenization, and show that the two can be effectively bridged. Based on this finding, we propose a unified progressive training strategy that gradually transitions (i) from continuous VAEs to discrete VAEs and (ii) from low- to high-compression settings, while reusing previously learned priors to accelerate convergence and improve reconstruction quality.
- We develop a unified tokenizer that jointly learns continuous and discrete representations within a single model through a coupled training objective, achieving strong and stable reconstruction performance under both continuous and discrete tokenization.

– We propose two architectural improvements for high-compression discrete causal VAEs: multi-token quantization, which enables compositional code representations with a substantially larger effective vocabulary and stronger capacity; and first-frame enhancement, which allocates extra capacity to the first frame to alleviate the causal-context bottleneck, improving overall reconstruction quality.

2 Related Work

2.1 Generative Models

Generative models have demonstrated remarkable performance across various visual generation tasks. In the early deep learning era, Generative Adversarial Networks (GANs) [5,20,29] emerged as one of the pioneering and most influential paradigms for high-fidelity image synthesis. Recently, diffusion models [22] have demonstrated unprecedented performance in generative tasks, notably in text-to-image synthesis [53, 56, 57], through a systematic noise removal process. Beyond these canonical generation tasks, diffusion models have also been extended to various visual applications [38, 49, 68–70, 90, 98]. Subsequent advancements extended diffusion-based methodologies to video generation. Some video diffusion models directly model pixel-value distributions [34, 60], while others concentrate on modeling the distributions of latent tokens extracted by autoencoders [4, 73], rather than directly handling pixel data. As latent-space generation becomes increasingly dominant, the design of visual tokenizers has become a critical component of modern generative systems. Recent studies explore various spatiotemporal compression settings to balance efficiency and fidelity [8, 10, 21, 81, 89], incorporate diffusion-style decoders or denoising-based decoding objectives into autoencoders [14, 35, 58, 65, 81, 84, 93, 99], and improve latent spaces via post-training or latent regularization [7, 32, 33, 47, 51, 75, 77, 92]. As an alternative technical approach, autoregressive models [18, 54, 86], inspired by GPT-style architectures [52], generate visual content by predicting the next token in a sequence.

2.2 Discrete VAEs

Recent progress in video generation has been driven by autoregressive models inspired by large language models, which rely on discrete tokenizers to quantize visual inputs into codebook indices in a latent space. Notable models such as TATS [19], MAGVIT [87], and VideoGPT [83] leverage discrete token training within the VQVAE [67] framework, employing 3D VAEs to extract discrete tokens. TATS [19], for example, introduces a hierarchical sampling strategy with autoregressive and interpolation Transformers, improving both generation quality and efficiency. MAGVIT-v2 [88] refines the VQ-VAE codebook by reducing its embedding dimension to zero and incorporating Lookup-Free

Quantization (LFQ), which has been widely adopted in contemporary models. VideoGPT [83] utilizes VQ-VAE, applying 3D convolutions and axial self-attention to learn downsampled discrete latent representations of video data. Additionally, newer models, such as LaViT [28] and ElasticTok [82], generate dynamic discrete visual tokens while preserving high-level semantic information. Approaches like the Omni-tokenizer [74], a joint image and video tokenizer, and Cosmos-Tokenizer [55], which uses Finite Scalar Quantization, further enhance the discrete tokenization process. At the same time, different improvements for discretization methods have been proposed in combination with diverse model architectures, such as VQ [18], LFQ [88], CVQ [96], IBQ [59], and FSQ [46]. These strategies optimize the tokenization process, significantly contributing to advancements in visual generation. Meanwhile, a growing line of work injects semantic signals into VQ-based (and hybrid) tokenizers to unify visual understanding and reconstruction [15, 16, 25, 27, 37, 41, 43, 45, 95].

2.3 Continuous VAE

Compared to discrete tokenization, continuous tokenization [11, 63, 85, 94] typically offers higher reconstruction fidelity. In the field of image generation, Stable Diffusion [56] introduced the use of continuous VAEs in diffusion models, significantly enhancing image generation. This approach has inspired numerous subsequent works [9, 17, 50] in the field of video generation. Some studies integrate 3D convolutions or spatio-temporal attention mechanisms into the backbone network [3, 23, 26, 60, 80], creating latent spaces specifically designed for video data. Video generation models initially performed frame-by-frame decoding, but following Sora [48], a series of video diffusion models trained video VAEs to compress data along the temporal dimension [31, 39, 40, 85, 97]. Some works like Hunyuan Video [31] and Allegro VAE [39] also contributed to this paradigm shift. Open-Sora [97] and Open-SoraPlan [40] are open-source projects aimed at replicating OpenAI's Sora, offering efficient continuous VAEs. Meanwhile, CogVideoX [85] retains more information by preserving a larger number of latent channels. Additionally, Cosmos-Tokenizer [55] provides a suite of continuous video tokenizers with various compression ratios. CV-VAE [94] employs latent space regularization to align the video latent space with the latent space of existing image VAEs [56], enabling effective spatio-temporal compression. Unlike previous methods that treat discrete and continuous representations separately, we propose OneVAE, a unified framework. Through progressive training and architectural enhancements, our approach improves reconstruction quality for discrete representations while ensuring robust performance across both types.

3 Method

3.1 Preliminaries of Continuous and Discrete VAEs

Continuous and discrete VAEs differ fundamentally in how they represent latent variables. Continuous VAEs [30] model the latent space as a continuous probability distribution, where the encoder $E(x)$ maps the input x to the parameters

of a distribution—typically a Gaussian. The latent variable z is then sampled from this distribution.

$$E(x) = \mathcal{N}(\mu, \sigma^2), \quad z \sim E(x) \tag{1}$$

where $\mathcal{N}(\mu, \sigma^2)$ represents a Gaussian distribution with mean μ and variance σ^2.

Similar to continuous VAEs, discrete VAEs compress video data $V \in \mathbb{R}^{C\times T\times H\times W}$ into a low-resolution representation $z \in \mathbb{R}^{D\times \frac{T}{t}\times \frac{H}{h}\times \frac{W}{w}}$ through the encoder and then decompress back to the original representation $\hat{x}$ using the decoder:

$$z = E(x), \quad \hat{x} = D(z) \tag{2}$$

where $E(\cdot)$ represents the encoder and $D(\cdot)$ is the decoder. Discrete VAEs [18,59] typically remove the Gaussian sampling process used in continuous VAEs. To convert z into a discrete token, two primary quantization strategies exist. VQ-based quantization selects the closest vector from a predefined codebook $\mathcal{C} = \{c_i\}_{i=1}^{N}$ based on Euclidean distance:

$$q_i = \arg\min_{c_i\in\mathcal{C}} \|z_i - c_i\|_2. \tag{3}$$

Alternatively, FSQ-based quantization [46] directly constrains each dimension of z within a bounded range and applies a rounding operation:

$$q_i = \text{Round}(\text{Bound}(z_i)). \tag{4}$$

After that, the d dimensions of q_i are all integers in $(-L/2, L/2)$, and they can be easily converted into an index.

Table 1: Comparison of VQ [18] and FSQ [46] trained in the latent space of a continuous VAE, where only the intermediate quantization modules are trained.

Quant. Method	PSNR ↑	SSIM ↑	LPIPS ↓
FSQ	26.46	0.8566	0.0825
VQ	22.82	0.7310	0.2071

Intuitively, the FSQ-based approach involves only a simple mapping and rounding of continuous features, which makes it highly similar to those features. To explore this, we experiment in Tab. 1 by training quantizer modules based on FSQ and VQ into a pre-trained continuous VAE to train using discrete reconstruction loss. The results show that FSQ outperforms VQ, suggesting that the discrete features encoded by FSQ exhibit a higher similarity to the continuous features of the VAE.

3.2 Training Procedure Improvements

Joint Discrete and Continuous Dual Path Training. Observing that FSQ yields discrete latents close to the continuous VAE latent geometry (Tab. 1), we explore leveraging continuous reconstruction to strengthen discrete-token learning, and introduce OneVAE, a unified VAE framework that jointly learns continuous and discrete latents within a single model. As shown in Fig. 2(a), our method simultaneously optimizes two pathways during training: One path performs discrete reconstruction, while the other performs continuous reconstruction. Specifically, at each training step, we randomly choose whether the latent variable will be encoded discretely or continuously, and the model adjusts its encoding accordingly. The model is jointly optimized using the loss from discrete reconstruction L_d and the loss from continuous reconstruction L_c. The loss function of our joint discrete and continuous optimization is defined as follows:

$$L = \begin{cases} L_d, & \text{if } r < R_{dis} \\ L_c, & \text{if } r \geq R_{dis} \end{cases} \tag{5}$$

where R_{dis} is a ratio that controls the proportion between discrete and continuous reconstruction, and r is a random number in the range $(0, 1)$. This dual-path optimization interleaves a stable continuous reconstruction signal with discrete-token learning, retaining the learned continuous latent structure and improving the robustness of discrete VAE training. The results of the ablation experiments in Tab. 2 further validate our conclusions.

Progressive Training. Building on the above finding that joint discrete continuous optimization improves discrete reconstruction, we design a unified progressive training strategy. It is motivated by two practical considerations: (1) continuous optimization is more stable than discrete optimization, and can thus serve as a stabilizer during training; (2) since continuous and discrete VAEs are compatible, a pretrained continuous VAE provides a strong initialization for the discrete model. Specifically, we aim to train VAEs across multiple compression ratios in a unified and efficient manner, instead of training separate models for each compression setting as in conventional training pipelines. As shown in Fig. 2(b), we first train a continuous VAE with configuration $8 \times 8 \times 4$, and then progressively extend it to higher-compression settings (e.g., $16 \times 16 \times 4$) for both continuous and discrete tokenization. To obtain a discrete VAE at the same compression, we insert an FSQ quantization module into the continuous VAE to produce tokenized latents, and then continue training with the joint discrete-continuous objective described above. We refer to this pipeline as unified progressive training, where all configurations are derived from the initial $8 \times 8 \times 4$ continuous VAE. As shown in Fig. 4, this strategy accelerates convergence and improves the final reconstruction performance.

Unified VAE. Through the explorations above, we discover the interplay between discrete and continuous representations and find that the continuous VAE can significantly aid in training the discrete VAE. In our original dual-path training, we introduce a ratio R_{dis} that controls the proportion between discrete and

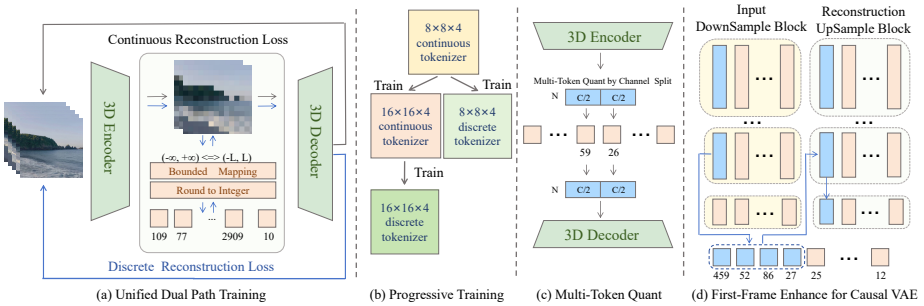

Fig. 2: Training method and architecture of the proposed model. Figure (a) presents the unified training framework that integrates both continuous and discrete paths in the latent space, improving the stability of discrete tokenizer training and raising its performance. Figure (b) illustrates the progressive training process, where all models are derived from an 8×8×4 continuous VAE. Figure (c) demonstrates the working principle of multi-token quantization, which can significantly enhance the representational capacity. Figure (d) illustrates the First-Frame Enhancement mechanism: by reducing the compression strength of the first frame to strengthen its reconstruction, it enables the causal VAE to leverage this information in subsequent frames, markedly improving the 4×16×16 discrete encoder's reconstruction quality.

continuous reconstruction loss. Since the original dual-path training is primarily designed to enhance the discrete representation, R_{dis} is set to a high value to ensure that the discrete path occupies a large proportion. A natural idea that follows is whether this dual-path joint optimization, can be extended to a unified model capable of simultaneously encoding both continuous and discrete representations. To explore the unified model, we set the ratio to 0.5 to balance the proportion, which could enable a trade-off between discrete and continuous representations, and train the unified model accordingly. Surprisingly, we find that with balanced dual path training, we obtain a unified model that performs well in both discrete and continuous encoding. As shown in Tab. 3, our unified model achieves strong performance in both discrete and continuous representation tasks, which further proves that discrete and continuous representations are not isolated, breaking the previous notion of training them separately and providing interesting insights for future VAE research.

3.3 Model Architecture Improvements

Multi-Token Quantization. As illustrated in Fig. 2 (c), we propose a multi-token quantization strategy that decomposes each latent feature vector along its channel dimension into multiple smaller sub-vectors, and quantizes each independently. Specifically, each feature channel is divided into several segments—for example, two equal parts in our implementation. Each segment is then quantized separately into discrete tokens using shared codebooks. This approach increases

Fig. 3: Visual comparison illustrating our First-Frame Enhancement strategy. The second column shows results from a causal $4 \times 16 \times 16$ video VAE, where the first frame suffers from noticeably poorer reconstruction quality than subsequent frames, and the overall performance lags behind the non-causal VAE in the third column. With our First-Frame Enhancement, the fourth column shows substantial improvements in both the first and subsequent frames. Zoom in to examine text reconstruction details.

the effective representational capacity without enlarging the total codebook size, as multiple tokens collectively encode the original vector with finer granularity. By capturing more subtle variations within each channel, multi-token quantization enables a richer and more flexible discrete representation of latent features. As a result, this leads to improved reconstruction fidelity and overall performance. Compared to conventional single-token quantization that compresses an entire channel vector into one token, our method strikes a balance between compression and expressiveness, addressing the limited expressiveness of traditional vector quantization methods under high compression ratios.

First-Frame Enhancement. Previous works on causal VAEs [55, 63, 78] have either explicitly noted or been implicitly observed (e.g., in open-source model outputs) that high-compression video VAEs tend to degrade the quality of early frames in a sequence. However, this issue has not been effectively addressed. By comparing the same VAE architecture under causal and non-causal settings in Fig. 3, we found that the inferior quality of early frames in the causal setting likely stems from their limited accessible information. As illustrated in Fig. 2(d), we propose a *First-Frame Enhancement* strategy to address this issue. The key idea is to reduce the compression strength of the first frame by allocating more tokens per spatial unit, while keeping the overall video compression ratio unchanged. This design enriches the structural and textural fidelity of the anchor frame, which in turn serves as a higher-quality reference for reconstructing subsequent frames in a causal manner. As shown in Fig. 3, compared to uniform high-ratio compression, which can severely degrade early-frame details and dis-

rupt temporal consistency, our method preserves anchor-frame fidelity without increasing the token budget. This targeted adjustment significantly improves both the first-frame reconstruction and the temporal coherence of later frames, leading to substantial performance gains in high-compression discrete VAE.

4 Experiments

4.1 Implementation Details

Training. Following previous work [94], we train our model on the WebVid-10M [2] dataset. AdamW [42] optimizer is used in the training, with hyperparameters set to $\beta_1 = 0.5$ and $\beta_2 = 0.9$, following VQGAN [59]. To reduce memory overhead, we train the model using 16-bit floating-point precision (FP16). The reconstruction objective includes pixel-level reconstruction loss and perceptual loss, where the LPIPS loss weight is set to 0.1. We maintain an exponential moving average (EMA) of model parameters with a decay rate of 0.999 to stabilize training and evaluation. Before training, video data was preprocessed into 17-frame clips, with each frame resized to 256×256 resolution. Once the training converges, we finetune the decoder on 33-frame videos. We also incorporated Temporal PatchGAN [6] to enhance the model's generative capabilities. The GAN training was initiated after a 20K-step warm-up phase, ensuring stable adversarial learning.

Evaluation. Following previous work [40, 94], evaluation is conducted on Panda 70M [12] and MCL-JCV [71] at a resolution of 33×512×512. Due to page limitations, the comparison on MCL-JCV is included in the supplementary materials. We first resize the shorter side of the original video to 512, followed by a central crop. To evaluate our model, we use multiple metrics across different aspects. For reconstruction quality, we employ PSNR, SSIM [79], and LPIPS [91]. These metrics measure pixel-wise accuracy, structural similarity, and perceptual similarity, respectively. Furthermore, FVD [66] is used to assess visual quality and temporal consistency.

4.2 Ablation Study

Ablation of Training Strategies. We conduct ablation studies to evaluate our unified progressive training strategy, which unifies continuous and discrete optimization to facilitate better discrete tokenization. As shown in Tab. 2, the baseline is trained from scratch without unified training. Enabling unified training, which uses continuous initialization together with dual-path discrete-continuous optimization, improves PSNR from 27.22 to 28.15 and SSIM from 0.8673 to 0.9039. These results validate the effectiveness of our unified training formulation. To further illustrate the effect of the proposed training recipe over the entire optimization trajectory, we report the reconstruction curves under different training strategies in Fig. 4. Specifically, unified training with FSQ achieves

Table 2: Ablation studies of unified training and model structures. The first two rows evaluate unified training, and the remaining rows ablate Multi-Token and Feature Enhancement on top of it. The results show that both the training strategy and structural designs improve reconstruction quality.

Unified Training	Multi-Token	Feature Enhancement	PSNR↑	SSIM↑	LPIPS↓
			27.22	0.8673	0.1029
✓			28.15	0.9039	0.0756
✓	✓		28.87	0.9207	0.0683
✓		✓	28.88	0.8973	0.0669
✓	✓	✓	29.35	0.9183	0.0611

a significant 5× convergence speedup compared to vanilla FSQ trained from scratch, substantially reducing the iterations needed to reach strong reconstruction quality. Moreover, unified training with FSQ attains better upper-bound performance than both unified training with SimVQ and vanilla FSQ. Overall, these results validate unified training as an effective way to make discrete VAE optimization both faster and more reliable.

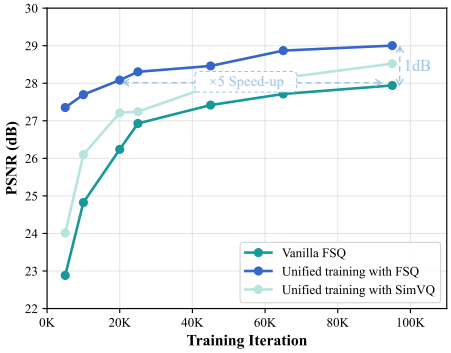

Fig. 4: PSNR versus training iterations for three methods: unified training with FSQ, unified training with SimVQ, and vanilla FSQ trained from scratch. Unified training with FSQ achieves a significant 5× convergence speedup compared to vanilla FSQ, and attains better upper-bound performance than both.

Fig. 5: Visual comparison between training with only discrete path and both paths. We visualize and compare the features before quantization. Our proposed joint discrete and continuous paths training enables the model to capture richer features in the latent space for more effective video encoding.

Ablation of Structure Improvement. Tab. 2 ablates the two proposed structure components: multi-token quantization (MT) and first-frame enhancement (FE). MT improves reconstruction performance by enlarging the effective dis-

Table 3: Quantitative comparisons of reconstruction quality on the Panda70M [12] test set. Our OneVAE achieves strong reconstruction performance across both continuous and discrete tokenizer settings. **Notably, for OneVAE, discrete and continuous results at same compression ratio are evaluated using the same model, enabled by our unified discrete-continuous training scheme.** The comparison on MCL-JCV [71] is provided in the supplementary materials, where our method also demonstrates better performance. OneVAE-MT and OneVAE-FE correspond to our proposed multi-token quantization and first-frame enhancement strategies, respectively, as detailed in Sec. 3.3.

	Method	Compress	z_dim	PSNR (↑)	SSIM (↑)	LPIPS (↓)	FVD (↓)
Continuous	CV-VAE	8×8×4	4	31.12	0.9142	0.0554	100.72
	Open-Sora-Plan		4	27.88	0.8840	0.0641	124.66
	Open-Sora-v1.2		4	31.67	0.9147	0.0648	104.48
	OneVAE		6	**32.48**	**0.9335**	**0.0430**	**59.16**
	Cosmos	16×16×4	16	30.61	0.9144	0.0739	135.68
	Cosmos	16×16×8		29.98	0.9041	0.0868	154.34
	OneVAE	16×16×4		**32.43**	**0.9318**	**0.0535**	**87.61**
Discrete	OmniTokenizer	8×8×4	-	23.73	0.8974	0.0735	206.18
	Emu-3			29.45	0.8912	0.0584	225.46
	Cosmos			30.34	0.9159	0.0524	82.86
	OneVAE			30.67	0.9228	0.0528	108.93
	OneVAE-MT	8×8×8		**31.80**	**0.9399**	**0.0450**	**78.35**
	Cosmos	8×8×8	-	27.17	0.8638	0.1409	303.37
	Cosmos	16×16×4		26.80	0.8531	0.1260	350.87
	Cosmos	16×16×8		26.18	0.8342	0.1337	396.97
	OneVAE	16×16×4		27.40	0.8690	0.1063	336.48
	OneVAE-FE	16×16×4		**28.11**	**0.8772**	**0.0814**	**213.04**

crete representation capacity without sacrificing the token compression ratio, while FE uses a lower-compression frame as an anchor in causal decoding, alleviating the limited-context issue of the first frame and improving overall reconstruction quality. These results demonstrate that our proposed MT and FE improve reconstruction quality over the baseline. When applied together, the two modules provide complementary benefits, leading to the highest quality of reconstruction.

Latent Space Visualization. To demonstrate the effectiveness of our dual-path optimization, we visualize and compare the features before quantization between the model optimized using only the discrete path and the model optimized using both the continuous and discrete paths. As shown in Fig. 5, training with both discrete and continuous paths enables the model to capture richer features in the latent space. We hypothesize that this improvement is due to the enhanced connection between the decoder and encoder in dual-path reconstruction. In discrete tokenizers, backpropagation relies on using the gradients of discrete vectors as approximations for continuous features by stopping gradients, which may introduce inaccuracies in gradient propagation. In contrast, our unified training approach likely helps the model achieve more precise gradient flow, leading to more effective video encoding.

Table 4: Video generation comparison on UCF-101. To align with previous works, we adopt the 16×128×128 video format for evaluation. Unlike prior methods, our original OneVAE tokenizer is not trained on the UCF-101 dataset or at the 128×128 resolution. Following CogVideo [24], we report FVD between generated videos and reconstructed ground-truth videos, denoted by *. For completeness, we also report FVD computed against the original ground-truth videos when available. Parameter counts are reported in millions (M). Our generator employs the identical AR architecture as LARP [72] for a fair comparison; the additional 143M parameters mainly come from the larger vocabulary embedding used by our general-purpose video tokenizer.

Method	Tokenizer Params	Generator Params	**gFVD↓**
Diffusion-based			
VideoFusion [44]	-	2000	173
HPDM [61]	-	725	66
MLM-based			
MAGVIT [87]	158	305	76
MAGVIT-v2 [88]	-	307	58
AR-based			
TATS	32	321	332
MAGVIT-AR [87]	158	464	265
MAGVIT-v2-AR [88]	-	306	109
OmniTokenizer [74]	82.2	650	191
Cogvideo [24]	-	9400	545* / 626
LARP-L [72]	173	632	43* / 57
Ours	135	775	37* / 62

4.3 Quantitative Comparison

Comparison with Tokenizers in Reconstruction. As shown in Tab. 3, we provide a quantitative comparison of our unified model, OneVAE with several continuous-only and discrete-only baseline models, including CV-VAE [94], Open-Sora [97], Open-Sora-Plan [40], OmniTokenizer [74], Cosmos Tokenizer [55] and Emu-3 [76]. Since few baselines provide a 16×16×4 setting, we adapt Cosmos Tokenizer from its 16×16×8 configuration for comparison. The models are evaluated using various metrics, including PSNR, SSIM, LPIPS, and FVD, which are commonly used to assess the quality of reconstructions and generative performance. From the table, we can observe that OneVAE consistently outperforms other discrete-only and continuous-only models in most metrics. In summary, the quantitative results demonstrate that our OneVAE not only unifies discrete and continuous representations but also achieves superior overall reconstruction performance, further corroborating that discrete and continuous representations can be unified.

Video Generation on UCF-101. While previous studies [39, 55, 94] on video VAEs mainly validate their models through reconstruction tasks, the primary role of a VAE is to support downstream generative modeling. This raises an important question: whether the improved reconstruction fidelity of our model

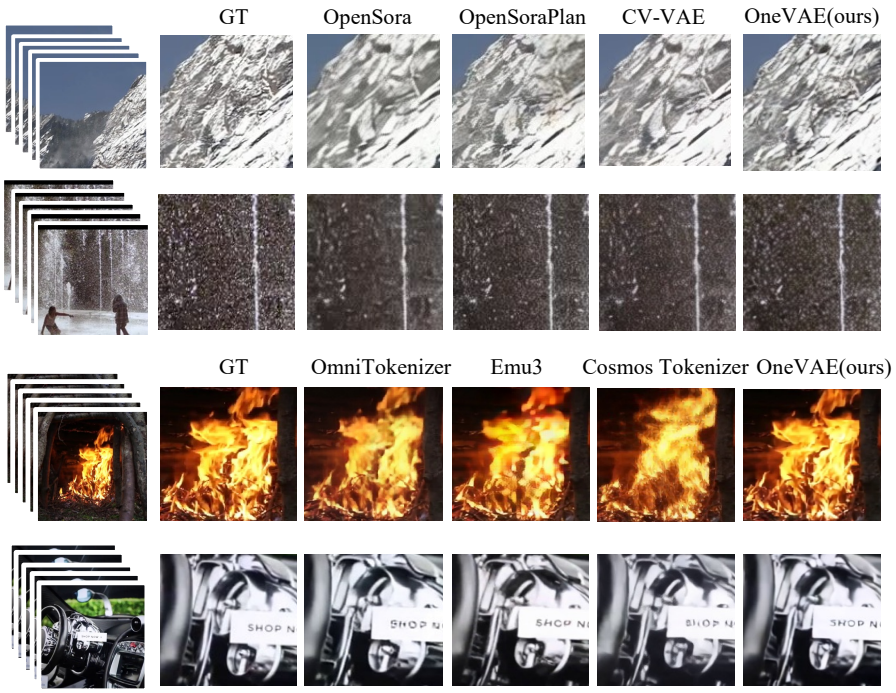

Fig. 6: We present a qualitative comparison of our proposed unified model OneVAE with continuous-only and discrete-only methods, including CV-VAE [94], Open-Sora [97], Open-Sora-Plan [40], OmniTokenizer [74], and Emu-3 [76]. These examples provide an intuitive comparison of reconstruction quality across different tokenizer designs.

makes the resulting discrete representation harder for a subsequent autoregressive (AR) model to learn. To examine this, we conduct video generation experiments on UCF-101. Following prior work, we train an AR video generation model at 128×128 resolution and use the same AR architecture as LARP [72] for a controlled comparison. Note that this comparison is not intended to claim a strictly superior generation model, since our OneVAE uses a larger codebook for general-purpose video tokenization, which introduces an additional 143M parameters in the vocabulary embedding compared with LARP. Following CogVideo [24], we report FVD between generated videos and reconstructed ground-truth videos. As shown in Tab. 4, our method achieves results comparable to LARP under the same AR architecture, suggesting that the discrete representation learned by OneVAE remains compatible with next-token prediction and does not hinder downstream AR generation.

4.4 Qualitative Comparison

As shown in Fig. 6, we present a qualitative comparison of the performance of our proposed unified VAE (OneVAE) against several discrete only and continuous only baseline models under the closest model settings. When compared to continuous only methods, it is evident that OneVAE produces accurate and high-fidelity reconstructions among all models. In the first example, our method better preserves the sharp mountain structures and fine snow textures. In the second example, OneVAE successfully restores dense visual details, demonstrating its superior capability in preserving structural integrity and texture clarity. When compared to discrete-only tokenizers, OneVAE again stands out by producing more realistic and visually appealing reconstructions. For the first video, OneVAE reconstructs richer and more detailed flames, whereas other methods produce blurry textures, especially struggling to preserve fine details. In the second comparison, OneVAE reconstructs sharper object and text regions, making the visual content more faithful to the input video. Overall, although OneVAE is a unified discrete-continuous VAE, the qualitative results highlight that OneVAE reconstructs accurate and visually coherent videos, demonstrating its strong performance in preserving details and generating high-quality outputs.

5 Conclusions

In this work, we propose OneVAE, a unified VAE framework that bridges discrete and continuous video representations. By rethinking their relationship, we propose a progressive training strategy built upon joint discrete-continuous optimization, which accelerates convergence and improves performance. We further present a unified VAE that supports both discrete and continuous representations within a single model. To further enhance video VAE performance, we propose meaningful structural improvements: multi-token quantization to increase representational capacity, and first-frame enhancement to preserve anchor-frame fidelity in causal VAEs and improve temporal consistency. These designs significantly improve reconstruction quality under high compression. Extensive experiments validate that our approach achieves faster training, better reconstruction quality, and unified modeling capability, offering a promising direction for future video generation and video VAE.

Acknowledgements

This work was funded by the National Natural Science Foundation of China under 62522607, 62495061, and 625B2093, and the Fundamental Research Funds for the Central Universities (Nankai University), Shenzhen Science and Technology Program (JCYJ20240813114237048).

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Bain, M., Nagrani, A., Varol, G., Zisserman, A.: Frozen in time: A joint video and image encoder for end-to-end retrieval. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1728–1738 (2021)
3. Bar-Tal, O., Chefer, H., Tov, O., Herrmann, C., Paiss, R., Zada, S., Ephrat, A., Hur, J., Liu, G., Raj, A., et al.: Lumiere: A space-time diffusion model for video generation. In: SIGGRAPH Asia 2024 Conference Papers. pp. 1–11 (2024)
4. Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S.W., Fidler, S., Kreis, K.: Align your latents: High-resolution video synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22563–22575 (2023)
5. Brock, A., Donahue, J., Simonyan, K.: Large scale gan training for high fidelity natural image synthesis. arXiv preprint arXiv:1809.11096 (2018)
6. Chang, Y.L., Liu, Z.Y., Lee, K.Y., Hsu, W.: Free-form video inpainting with 3d gated convolution and temporal patchgan. In: Proceedings of the IEEE/CVF international conference on computer vision (2019)
7. Chen, C., Huang, Z., Zou, C., Zhu, M., Ji, K., Liu, J., Chen, J., Chen, H., Shen, C.: Hieratok: Multi-scale visual tokenizer improves image reconstruction and generation. arXiv preprint arXiv:2509.23736 (2025)
8. Chen, H., Wang, Z., Li, X., Sun, X., Chen, F., Liu, J., Wang, J., Raj, B., Liu, Z., Barsoum, E.: Softvq-vae: Efficient 1-dimensional continuous tokenizer. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 28358–28370 (2025)
9. Chen, J., Yu, J., Ge, C., Yao, L., Xie, E., Wang, Z., Kwok, J., Luo, P., Lu, H., Li, Z.: Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis **2024**, 57611–57640 (2024)
10. Chen, J., Cai, H., Chen, J., Xie, E., Yang, S., Tang, H., Li, M., Lu, Y., Han, S.: Deep compression autoencoder for efficient high-resolution diffusion models. arXiv preprint arXiv:2410.10733 (2024)
11. Chen, L., Li, Z., Lin, B., Zhu, B., Wang, Q., Yuan, S., Zhou, X., Cheng, X., Yuan, L.: Od-vae: An omni-dimensional video compressor for improving latent video diffusion model pp. 1–6 (2025)
12. Chen, T.S., Siarohin, A., Menapace, W., Deyneka, E., Chao, H.w., Jeon, B.E., Fang, Y., Lee, H.Y., Ren, J., Yang, M.H., et al.: Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13320–13331 (2024)
13. Chen, X., Wu, Z., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., Ruan, C.: Janus-pro: Unified multimodal understanding and generation with data and model scaling. arXiv preprint arXiv:2501.17811 (2025)
14. Chen, Y., Girdhar, R., Wang, X., Rambhatla, S.S., Misra, I.: Diffusion autoencoders are scalable image tokenizers. arXiv preprint arXiv:2501.18593 (2025)
15. Chen, Z., Wang, C., Chen, X., Xu, H., Huang, R., Zhou, J., Han, J., Xu, H., Liang, X.: Semhitok: A unified image tokenizer via semantic-guided hierarchical codebook for multimodal understanding and generation. arXiv preprint arXiv:2503.06764 (2025)

16. Du, S., Guo, J., Li, B., Cui, S., Xu, Z., Luo, Y., Wei, Y., Gai, K., Wang, X., Wu, K., et al.: Vqrae: Representation quantization autoencoders for multimodal understanding, generation and reconstruction. arXiv preprint arXiv:2511.23386 (2025)
17. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: Forty-first international conference on machine learning (2024)
18. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12873–12883 (2021)
19. Ge, S., Hayes, T., Yang, H., Yin, X., Pang, G., Jacobs, D., Huang, J.B., Parikh, D.: Long video generation with time-agnostic vqgan and time-sensitive transformer. In: European Conference on Computer Vision. pp. 102–118. Springer (2022)
20. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)
21. Guo, M.H., Wang, C., Liu, W., Hu, S.M.: Fastmae: Efficient masked autoencoder with offline tokenizer. Computational Visual Media **11**(3), 483–496 (2025)
22. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
23. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. Advances in Neural Information Processing Systems **35**, 8633–8646 (2022)
24. Hong, W., Ding, M., Zheng, W., Liu, X., Tang, J.: Cogvideo: Large-scale pretraining for text-to-video generation via transformers. arXiv preprint arXiv:2205.15868 (2022)
25. Huang, Z., Zheng, D., Zou, C., Liu, R., Wang, X., Ji, K., Chai, W., Sun, J., Wang, L., Lv, Y., et al.: Ming-univision: Joint image understanding and generation with a unified continuous tokenizer. arXiv preprint arXiv:2510.06590 (2025)
26. Jiang, J., Hong, G., Zhou, L., Ma, E., Hu, H., Zhou, X., Xiang, J., Liu, F., Yu, K., Sun, H., et al.: Dive: Dit-based video generation with enhanced control. arXiv preprint arXiv:2409.01595 (2024)
27. Jiao, Y., Qiu, H., Jie, Z., Chen, S., Chen, J., Ma, L., Jiang, Y.G.: Unitoken: Harmonizing multimodal understanding and generation through unified visual encoding. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 3600–3610 (2025)
28. Jin, Y., Sun, Z., Xu, K., Chen, L., Jiang, H., Huang, Q., Song, C., Liu, Y., Zhang, D., Song, Y., et al.: Video-lavit: Unified video-language pre-training with decoupled visual-motional tokenization. arXiv preprint arXiv:2402.03161 (2024)
29. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4401–4410 (2019)
30. Kingma, D.P., Welling, M.: Auto-encoding variational bayes (2013)
31. Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., et al.: Hunyuanvideo: A systematic framework for large video generative models. arXiv preprint arXiv:2412.03603 (2024)
32. Kouzelis, T., Kakogeorgiou, I., Gidaris, S., Komodakis, N.: Eq-vae: Equivariance regularized latent space for improved generative image modeling. arXiv preprint arXiv:2502.09509 (2025)

33. Lee, H., Kim, M., Jang, S., Jeong, J., Hwang, S.J.: Enhancing variational autoencoders with smooth robust latent encoding. arXiv preprint arXiv:2504.17219 (2025)
34. Li, M., Lin, J., Meng, C., Ermon, S., Han, S., Zhu, J.Y.: Efficient spatially sparse inference for conditional gans and diffusion models. Advances in neural information processing systems **35**, 28858–28873 (2022)
35. Li, S., Zhuo, L., Wang, Y., Li, M., He, X., Wu, F., Li, H., Heng, P.A.: Proteinae: Protein diffusion autoencoders for structure encoding. arXiv preprint arXiv:2510.10634 (2025)
36. Li, T., Tian, Y., Li, H., Deng, M., He, K.: Autoregressive image generation without vector quantization. Advances in Neural Information Processing Systems **37**, 56424–56445 (2024)
37. Li, Y., Qian, R., Pan, B., Zhang, H., Huang, H., Zhang, B., Tong, J., You, H., Du, X., Gan, Z., et al.: Manzano: A simple and scalable unified multimodal model with a hybrid vision tokenizer. arXiv preprint arXiv:2509.16197 (2025)
38. Li, Z., Cao, M., Wang, X., Qi, Z., Cheng, M.M., Shan, Y.: Photomaker: Customizing realistic human photos via stacked id embedding. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8640–8650 (2024)
39. Li, Z., Lin, B., Ye, Y., Chen, L., Cheng, X., Yuan, S., Yuan, L.: Wf-vae: Enhancing video vae by wavelet-driven energy flow for latent video diffusion model pp. 17778–17788 (2025)
40. Lin, B., Ge, Y., Cheng, X., Li, Z., Zhu, B., Wang, S., He, X., Ye, Y., Yuan, S., Chen, L., et al.: Open-sora plan: Open-source large video generation model. arXiv preprint arXiv:2412.00131 (2024)
41. Lin, H., Wang, T., Ge, Y., Ge, Y., Lu, Z., Wei, Y., Zhang, Q., Sun, Z., Shan, Y.: Toklip: Marry visual tokens to clip for multimodal comprehension and generation. arXiv preprint arXiv:2505.05422 (2025)
42. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
43. Lu, J., Song, L., Xu, M., Ahn, B., Wang, Y., Chen, C., Dehghan, A., Yang, Y.: Atoken: A unified tokenizer for vision. arXiv preprint arXiv:2509.14476 (2025)
44. Luo, Z., Chen, D., Zhang, Y., Huang, Y., Wang, L., Shen, Y., Zhao, D., Zhou, J., Tan, T.: Videofusion: Decomposed diffusion models for high-quality video generation. arXiv preprint arXiv:2303.08320 (2023)
45. Ma, C., Jiang, Y., Wu, J., Yang, J., Yu, X., Yuan, Z., Peng, B., Qi, X.: Unitok: A unified tokenizer for visual generation and understanding. arXiv preprint arXiv:2502.20321 (2025)
46. Mentzer, F., Minnen, D., Agustsson, E., Tschannen, M.: Finite scalar quantization: Vq-vae made simple **2024**, 51772–51783 (2024)
47. Novitskiy, L., Vasilev, V., Kovaleva, M., Arkhipkin, V., Dimitrov, D.: Vivat: Virtuous improving vae training through artifact mitigation. arXiv preprint arXiv:2506.07863 (2025)
48. OpenAI: Video generation models as world simulators (2024), `https://openai.com/index/video-generation-models-as-world-simulators/`, last accessed 27 Jun 2026
49. Ouyang, Z., Li, Z., Hou, Q.: K-lora: Unlocking training-free fusion of any subject and style loras. In: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13041–13050. IEEE (2025)

50. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis **2024**, 1862–1874 (2024)
51. Qiu, K., Li, X., Chen, H., Kuen, J., Xu, X., Gu, J., Luo, Y., Raj, B., Lin, Z., Savvides, M.: Image tokenizer needs post-training. arXiv preprint arXiv:2509.12474 (2025)
52. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al.: Improving language understanding by generative pre-training (2018)
53. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 **1**(2), 3 (2022)
54. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: International conference on machine learning. pp. 8821–8831. Pmlr (2021)
55. Reda, F., Gu, J., Liu, X., Ge, S., Wang, T.C., Wang, H., Liu, M.Y.: Cosmos tokenizer: A suite of image and video neural tokenizers (2024), `https://github.com/NVIDIA/Cosmos-Tokenizer`, last accessed 27 Jun 2026
56. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: 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)
57. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in neural information processing systems **35**, 36479–36494 (2022)
58. Sargent, K., Hsu, K., Johnson, J., Fei-Fei, L., Wu, J.: Flow to the mode: Mode-seeking diffusion autoencoders for state-of-the-art image tokenization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19471–19481 (2025)
59. Shi, F., Luo, Z., Ge, Y., Yang, Y., Shan, Y., Wang, L.: Taming scalable visual tokenizer for autoregressive image generation. arXiv preprint arXiv:2412.02692 (2024)
60. Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., et al.: Make-a-video: Text-to-video generation without text-video data. arXiv preprint arXiv:2209.14792 (2022)
61. Skorokhodov, I., Menapace, W., Siarohin, A., Tulyakov, S.: Hierarchical patch diffusion models for high-resolution video generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7569–7579 (2024)
62. Sun, P., Jiang, Y., Chen, S., Zhang, S., Peng, B., Luo, P., Yuan, Z.: Autoregressive model beats diffusion: Llama for scalable image generation. arXiv preprint arXiv:2406.06525 (2024)
63. Tang, A., He, T., Guo, J., Cheng, X., Song, L., Bian, J.: Vidtok: A versatile and open-source video tokenizer. arXiv preprint arXiv:2412.13061 (2024)
64. Team, C.: Chameleon: Mixed-modal early-fusion foundation models. arXiv preprint arXiv:2405.09818 (2024)
65. Teng, Y., Lin, M., Liu, X., Wang, S., Yang, X., Liu, X.: Adaptive 1d video diffusion autoencoder. arXiv preprint arXiv:2602.04220 (2026)
66. Unterthiner, T., Van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: Fvd: A new metric for video generation (2019)
67. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. Advances in neural information processing systems **30** (2017)

68. Wan, Y., Jiang, P.T., Hou, Q., Zhang, H., Chen, J., Cheng, M.M., Li, B.: Controlsr: Taming diffusion models for consistent real-world image super resolution. arXiv preprint arXiv:2410.14279 (2024)
69. Wan, Y., Liu, L., Zhou, J., Zhou, Z., Zhang, X., Zhang, D., Jiao, S., Hou, Q., Cheng, M.M.: Geoworld: Unlocking the potential of geometry models to facilitate high-fidelity 3d scene generation. arXiv preprint arXiv:2511.23191 (2025)
70. Wang, C., Peng, H.Y., Liu, Y.T., Gu, J., Hu, S.M.: Diffusion models for 3d generation: A survey. Computational Visual Media **11**(1), 1–28 (2025)
71. Wang, H., Gan, W., Hu, S., Lin, J.Y., Jin, L., Song, L., Wang, P., Katsavounidis, I., Aaron, A., Kuo, C.C.J.: Mcl-jcv: a jnd-based h. 264/avc video quality assessment dataset. In: 2016 IEEE international conference on image processing (ICIP). pp. 1509–1513. IEEE (2016)
72. Wang, H., Suri, S., Ren, Y., Chen, H., Shrivastava, A.: Larp: Tokenizing videos with a learned autoregressive generative prior **2025**, 60671–60697 (2025)
73. Wang, J., Yuan, H., Chen, D., Zhang, Y., Wang, X., Zhang, S.: Modelscope text-to-video technical report. arXiv preprint arXiv:2308.06571 (2023)
74. Wang, J., Jiang, Y., Yuan, Z., Peng, B., Wu, Z., Jiang, Y.G.: Omnitokenizer: A joint image-video tokenizer for visual generation. Advances in Neural Information Processing Systems **37**, 28281–28295 (2024)
75. Wang, M., Jiang, D., Li, L., Lin, Y., Shen, G., Kong, X., Liu, Y., Dai, G., Wang, J.: Vae-repa: Variational autoencoder representation alignment for efficient diffusion training. arXiv preprint arXiv:2601.17830 (2026)
76. Wang, X., Zhang, X., Luo, Z., Sun, Q., Cui, Y., Wang, J., Zhang, F., Wang, Y., Li, Z., Yu, Q., et al.: Emu3: Next-token prediction is all you need. arXiv preprint arXiv:2409.18869 (2024)
77. Wang, Y., Lin, Z., Teng, Y., Zhu, Y., Ren, S., Feng, J., Liu, X.: Bridging continuous and discrete tokens for autoregressive visual generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 18596–18605 (2025)
78. Wang, Y., Xiong, T., Zhou, D., Lin, Z., Zhao, Y., Kang, B., Feng, J., Liu, X.: Loong: Generating minute-level long videos with autoregressive language models. arXiv preprint arXiv:2410.02757 (2024)
79. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing (2004)
80. Wu, J.Z., Ge, Y., Wang, X., Lei, S.W., Gu, Y., Shi, Y., Hsu, W., Shan, Y., Qie, X., Shou, M.Z.: Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7623–7633 (2023)
81. Wu, Y., Cai, H., Chen, J., Zhang, Z., Xie, E., Yu, J., Chen, J., Hu, J., Lu, Y., Han, S.: Dc-ar: Efficient masked autoregressive image generation with deep compression hybrid tokenizer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 18034–18045 (2025)
82. Yan, W., Mnih, V., Faust, A., Zaharia, M., Abbeel, P., Liu, H.: Elastictok: Adaptive tokenization for image and video **2025**, 38036–38056 (2025)
83. Yan, W., Zhang, Y., Abbeel, P., Srinivas, A.: Videogpt: Video generation using vq-vae and transformers. arXiv preprint arXiv:2104.10157 (2021)
84. Yang, J., Li, T., Fan, L., Tian, Y., Wang, Y.: Latent denoising makes good visual tokenizers. arXiv preprint arXiv:2507.15856 (2025)
85. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer **2025**, 83048–83077 (2025)

86. Yu, J., Xu, Y., Koh, J.Y., Luong, T., Baid, G., Wang, Z., Vasudevan, V., Ku, A., Yang, Y., Ayan, B.K., et al.: Scaling autoregressive models for content-rich text-to-image generation. arXiv preprint arXiv:2206.10789 **2**(3), 5 (2022)
87. Yu, L., Cheng, Y., Sohn, K., Lezama, J., Zhang, H., Chang, H., Hauptmann, A.G., Yang, M.H., Hao, Y., Essa, I., et al.: Magvit: Masked generative video transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10459–10469 (2023)
88. Yu, L., Lezama, J., Gundavarapu, N.B., Versari, L., Sohn, K., Minnen, D., Cheng, Y., Gupta, A., Gu, X., Hauptmann, A.G., et al.: Language model beats diffusion-tokenizer is key to visual generation. In: International Conference on Learning Representations. vol. 2024, pp. 765–783 (2024)
89. Yuan, Z., Wang, S., Shang, Y., Zhang, H., Fang, T., Xie, R., Yan, S., Dai, G., Wang, Y.: Dlfr-vae: Dynamic latent frame rate vae for video generation. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 10388–10397 (2025)
90. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3836–3847 (2023)
91. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
92. Zhang, S., Zhang, H., Zhang, Z., Ge, C., Xue, S., Liu, S., Ren, M., Kim, S.Y., Zhou, Y., Liu, Q., et al.: Both semantics and reconstruction matter: Making representation encoders ready for text-to-image generation and editing. arXiv preprint arXiv:2512.17909 (2025)
93. Zhao, L., Woo, S., Wan, Z., Li, Y., Zhang, H., Gong, B., Adam, H., Jia, X., Liu, T.: Epsilon-vae: Denoising as visual decoding. arXiv preprint arXiv:2410.04081 (2024)
94. Zhao, S., Zhang, Y., Cun, X., Yang, S., Niu, M., Li, X., Hu, W., Shan, Y.: Cv-vae: A compatible video vae for latent generative video models. Advances in Neural Information Processing Systems **37**, 12847–12871 (2024)
95. Zhao, X., Zhang, Z., Huang, Y., Mi, Y., Mu, G., Ding, S., Wang, J., Guo, R., Zhou, S.: Glotok: Global perspective tokenizer for image reconstruction and generation. arXiv preprint arXiv:2511.14184 (2025)
96. Zheng, C., Vedaldi, A.: Online clustered codebook. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22798–22807 (2023)
97. Zheng, Z., Peng, X., Yang, T., Shen, C., Li, S., Liu, H., Zhou, Y., Li, T., You, Y.: Open-sora: Democratizing efficient video production for all (2024)
98. Zhou, Y., Zhou, D., Wang, Y., Feng, J., Hou, Q.: Maskdiffusion: Boosting text-to-image consistency with conditional mask. International Journal of Computer Vision **133**(5), 2805–2824 (2025)
99. Zhuang, S., Guo, Y., Fu, C., Huang, Z., Tian, Z., Wang, F., Zhang, Y., Li, C., Wang, Y.: Wetok: Powerful discrete tokenization for high-fidelity visual reconstruction. arXiv preprint arXiv:2508.05599 (2025)