VC-VAE: Leveraging Video Codecs for Training-Efficient and High-Fidelity Video VAE

Xinxu Ge[1], Shang Chai[1], Litong Gong[1], Zitong Yu[2,3], Xin Liu[4], and Tiezheng Ge[1⋆]

[1] Taobao & Tmall Group of Alibaba, China
gexinxu.gxx@alibaba-inc.com
{chaishang.cs,gonglitong.glt,getiezheng.gtz}@taobao.com
[2] Great Bay University, China
yuzitong@gbu.edu.cn
[3] Dongguan Key Laboratory for Intelligence and Information Technology, China
[4] Shanghai Jiao Tong University, School of Psychology, China
linuxsino@gmail.com

Abstract. Video Variational Auto-Encoders (Video VAEs) compress video data from the highly redundant pixel space into a compact latent representation, playing an important role in state-of-the-art video generation models. However, existing methods typically learn inter-frame correlations implicitly, overlooking the potential of breaking down video compression into two separate parts: keyframe encoding and inter-frame dynamic encoding, which is a fundamental design of traditional video codecs. To address this, we incorporate traditional video codec standard design into the Video VAE and introduce VC-VAE, a model that explicitly separates keyframe and inter-frame dynamic compression. We start by establishing a high-fidelity static keyframe anchor through initialization from a powerful pre-trained image VAE. Then, to explicitly model dynamic relative to this anchor, we introduce the Temporal Dynamic Difference Convolution (TDC), an operator designed to learn sparse motion residuals from inter-frame differences while maintaining a separate pathway for static content. Qualitative and quantitative experiments show that our proposed VC-VAE significantly outperforms baseline models in reconstruction quality, dynamic modelling, and training efficiency.

1 Introduction

Recent advances in video generation enable the creation of high-quality videos [13, 18, 21, 24, 29]. The progress is mainly built upon the Latent Diffusion Model (LDM) architecture [23], which relies heavily on the Video VAE as its core component. The task of the Video VAE is to compress video data from the highly redundant pixel space into a compact latent representation, directly influencing the efficiency of the following diffusion process and the quality of the final generated videos.

⋆ Corresponding author.

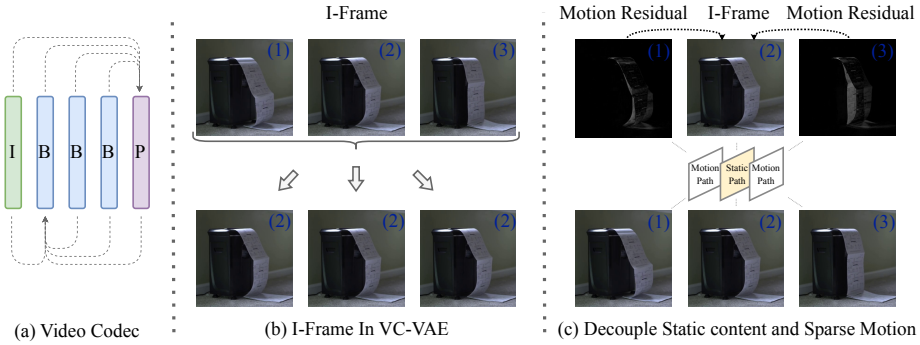

Fig. 1: (a) Inter-frame dependency in Video Codec. (b) Our model first anchors the reconstruction on a single, representative I-frame. (c) Our proposed TDC explicitly learns from the decoupled sparse motion residual from the I-frame.

Extensive research has been conducted on enhancing Video VAEs. From the signal processing viewpoint, WF-VAE [16] and Cosmos Tokenizer [1] aim to handle the spatio-temporal redundant in the frequency domain. By employing Haar Wavelet Transform [22], these methods compactly encode the structurally-vital but temporally-redundant low-frequency information, thereby freeing up more model capacity to capture the more complex, high-frequency variations. Another line of work focuses on basic architectural design. For instance, MovieGen [21] uses a non-causal architecture to improve inter-frame modeling, whereas IV-VAE [34] leverages a grouping mechanism to create local bidirectional dependencies, which enhances reconstruction stability while maintaining causality. However, leveraging the design principles of video codecs [32] to improve Video VAEs remains a sparsely explored area.

Rethinking the evolution of convolutional operators in Video VAEs uncovers an implicit parallel with the video codecs standards. As shown in Fig. 1(a), Video codecs standards [32] employ a Group-of-Pictures structure based on three types of frames: I-frames, which are independently encoded as self-contained anchor points; P-frames, which efficiently encode motion residuals relative to a previous frame; and B-frames, which further enhance compression by referencing frames in both past and future context. Early Video VAEs [37, 39] used causal convolution, establishing an inter-frame dependency similar to P-frame prediction where the current frame is reconstructed using only past and current context. Subsequent work, such as IV-VAE [34], introduced group causal convolution, which creates a B-frame-like dependency by utilizing both past and future context for more stable reconstruction. These architectures implicitly reflect the dependency structure of P and B frames; however, they treat static content and dynamic motion as an entangled learning target, overlooking the essential operational principle behind them. Video codecs leverage temporal redundancy through the I-frame prediction and motion residual mechanism, enabling the explicit decoupling of static content and dynamic motion. By isolating complex,

sparse motion from the redundant static background, this separation offers a fundamentally more efficient approach to video modeling.

Inspired by the above insights, our objective is to enhance both reconstruction quality and training efficiency by incorporating the above design principles of video codecs into Video VAE. We start by establishing an I-frame prior within the Video VAE, initializing our model from a pre-trained image VAE [15] to endow our Video VAE with a strong static content prior. As shown in Fig. 1(b), the model can generate high-fidelity, single-frame reconstructions at the initial state, providing a stable foundation for learning dynamic motion. We then introduce the Temporal Dynamic Difference Convolution (TDC), which is designed to incorporate decoupling inductive bias at the operator level. As shown in Fig. 1(c), TDC achieves this by constructing two parallel pathways: a dedicated pathway to preserve the static content from the I-frame prior, while the other learns sparse inter-frame dynamics from explicit feature differences. Extensive experiments validate the efficiency of our video codec-inspired learning paradigm. By explicitly separating the learning of dynamic sparse motion from static content, our model achieves superior reconstruction performance while reducing training overhead by 50%. In summary, the core contributions of this paper are as follows:

- We introduce VC-VAE, a novel architecture that utilizes a codec-inspired inductive bias. It explicitly decouples the video sequence into a static keyframe anchor and inter-frame dynamic variations. To the best of our knowledge, VC-VAE is the first work to successfully incorporate this spatio-temporal decoupling strategy into the Video VAE framework.
- To implement this explicit modeling of inter-frame temporal dynamics into the convolutional Video VAE architecture, we introduce the Temporal Dynamic Difference Convolution (TDC), a novel operator designed for this purpose.
- Extensive experiments on several benchmarks demonstrate the SOTA video reconstruction and generation capabilities of the proposed VC-VAE.

2 Related Work

2.1 Variational Autoencoder

The Variational Autoencoder (VAE) [12] is a foundational paradigm in generative modeling, designed to capture complex probability distributions within high-dimensional data. VAEs' development of both continuous [6, 23] and discrete latent representations [10] has established them as a versatile component in modern generative pipelines, functioning as visual compressors for latent diffusion models and as visual tokenizers for autoregressive systems. Contemporary research on continuous VAEs largely focuses on enhancing their efficiency as the first stage of latent diffusion systems, pursuing higher compression rates [6], faster throughput [42,43], and ensuring better compatibility with latent diffusion systems [14, 25, 38].

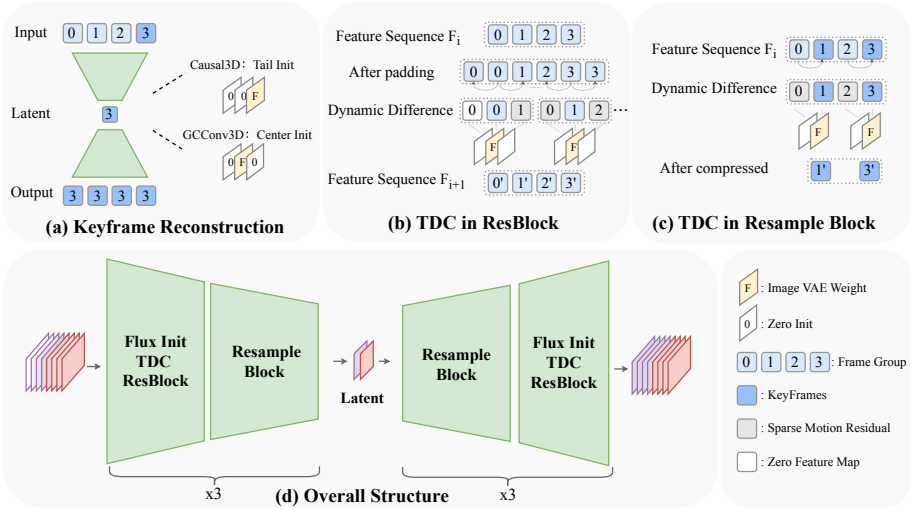

Fig. 2: (a). Initialization with the Flux VAE enables the Video VAE to reconstruct keyframes. (b)-(c). Implementation of the Temporal Dynamic Difference Convolution operator within different blocks of the Video VAE. (d). Overall structure of the proposed Video VAE.

2.2 Video VAE

Early video generation systems employ SVD-VAE [3], which does not perform temporal compression, necessitating temporal interaction layers to ensure temporal consistency. CV-VAE [41] introduced temporal compression into the Video VAE framework, while OD-VAE [7] and CogVideoX-VAE [37] incorporated the temporal causal structure [39] for temporal tiling. WF-VAE [16] explored wavelet transforms to efficiently reduce the expensive representation dimensionality of high resolutions. IV-VAE [34] proposed the dual-stream architecture and Group Causal Convolutions for robust inter-frame information interaction. To reduce the computational cost in Diffusion Transformers [20], H3AE [35] and Reducio VAE [27] aim for extremely high compression rates while maintaining the reconstruction capability. Vid-twin [30] improves the reconstruction quality of video tokenizer by compressing structural and dynamic information into separate latent components. However, existing Video VAEs merely exploit redundancy through temporal compression, without explicitly considering the inherent redundant nature of video data [17]. Inspired by video codec standard, we incorporate the principle of separating static content from dynamic motion into our Video VAE design.

3 Method

This section details our approach to integrating video codec principles into our VC-VAE. The overall model architecture is illustrated in Fig. 2. First, we in-

troduce the existing convolution operators in Sec. 3.1. Then, we build a reliable keyframe prior by initializing the model with a pre-trained image VAE in Sec. 3.2. Subsequently, to explicitly model temporal dynamics, we introduce the Temporal Dynamic Difference Convolution in Sec. 3.3. Finally, we describe additional refinements for consistent reconstruction in Sec. 3.4.

3.1 Preliminary

The core task of Video VAE is to learn a compressed yet informative latent representation for video data. Traditional video codec standards, designed explicitly for reducing the spatio-temporal redundancy in videos, offer invaluable insights. Motivated by this, our preliminary section revisits the fundamental operators of Video VAEs, such as Causal and Group Causal Convolutions, from the perspective of video codec principles.

Causal Convolution To achieve unified encoding of images and videos in a Video VAE, Causal 3D Conv [37, 39] is widely adopted, which is achieved by applying asymmetric padding only at the beginning of the temporal dimension, so the receptive field at time step t only covers current and past frames. Formally, for input sequence $\{X_0, ..., X_{T-1}\}$, the output Y_t at time step t is:

$$Y_t = \mathit{CausalConv}(X_t, X_{t-1}, \ldots, X_{t-k+1}) \tag{1}$$

where *CausalConv* represents a 3D convolution with temporal kernel size k. However, from a video codec perspective, the unidirectional dependency resembles P-frame prediction, leading to an asymmetric temporal receptive field. By only looking backwards in the temporal dimension, the model lacks future context, resulting in an information imbalance that limits its ability to model inter-frame dynamics effectively.

Group Causal Convolution Group Causal Convolution (GCConv) [34] mitigates this issue by introducing local bi-directional interactions, fostering stronger local dependencies by segmenting the video sequence into local non-causal frame groups. Formally, for an input frame group $G_i = \{X_0, ..., X_{T-1}\}$, the output group Y_i is computed as:

$$Y_i = \mathit{Conv}(\text{Concat}[G_{i-1}[-1], G_i]) \tag{2}$$

where $G_{i-1}[-1]$ is the last frame from the past group to maintain global causality. GCConv's advantage manifests in the superiority of bi-directional frames over predictive frames in video codec, as it allows the frame group to be encoded with richer, bi-directional context. The success of GCConv highlights the benefits of borrowing principles from video codecs. Motivated by this, the following section further instantiates the principle of video codec within the Video VAE architecture.

Fig. 3: Reconstruction results of different methods for a video sequence. Visual artifacts are enclosed in red bounding boxes for clarity.

Table 1: Video reconstruction results on WebVid-test and UCF-101 val. ↓ indicates lower is better, ↑ indicates higher is better.

Method	FCR	Chn	Params	WebVid-10M				UCF-101			
				FVD↓	PSNR↑	LPIPS↓	SSIM↑	FVD↓	PSNR↑	LPIPS↓	SSIM↑
CogVideoX	4*8*8	16	215M	41.70	35.23	0.04125	0.9394	50.47	37.61	0.03292	0.9605
Wanx2.1	4*8*8	16	127M	42.40	35.40	0.03724	0.9385	43.26	38.19	0.02971	0.9644
IVVAE	4*8*8	16	108M	42.80	35.21	0.04301	0.9343	40.68	38.48	0.02853	0.9660
WFVAE	4*8*8	16	222M	36.86	35.60	0.03778	0.9394	42.73	38.31	0.03036	0.9645
Hunyuan	4*8*8	16	246M	**36.23**	35.84	0.03452	0.9425	37.07	38.80	0.02757	0.9671
VC-VAE	4*8*8	16	146M	36.80	**36.04**	**0.03315**	**0.9459**	**36.30**	**38.97**	**0.02747**	**0.9684**
WFVAE	8*8*8	32	227M	44.13	35.86	0.03735	0.9396	53.11	37.95	0.03307	0.9627
VC-VAE	8*8*8	32	194M	**36.63**	**36.40**	**0.03298**	**0.9485**	**38.47**	**38.85**	**0.02854**	**0.9672**

3.2 Static Spatial Prior via Image VAE Initialization

While basic operators effectively mirror the inter-frame interaction of a video codec, an efficient encoding scheme also depends on high-quality Intra-encoded keyframes, which are self-contained, independently compressed frames serving as robust anchors for subsequent predictions. To integrate this essential feature into our Video VAE, we initialize its weights with a powerful, pre-trained image VAE, giving our model a strong spatial prior before it begins to learn temporal dynamics. We start by inflating the 2D kernels of the image VAE into 3D kernels to enable the model to act as a high-fidelity frame-wise encoder or decoder, employing Tail Initialization [7] for Causal Convolution and Center Initialization [5] for Group Causal Convolution, as shown in Fig. 2(a). To leverage the temporal redundancy inherent in video data, we configure Temporal Layers for Keyframe Selection and Broadcast. Following IVVAE [34], we utilize a convolutional layer with a temporal kernel size $k_t = 2$ and stride $s_t = 2$, denoted as $W_{down} = [C, C, k_t]$, where C represents the number of input and output channels.

We then initialize it as follows:

$$W_{\text{down}}[:, :, k_t - 1] = \mathbf{I}. \tag{3}$$

where $\mathbf{I} \in \mathbb{R}^{C \times C}$ denotes an identity matrix. For temporal upsampling, the operator is implemented by a convolution that doubles the channel from C to $2 * C$, followed by a PixelShuffle operation to expand the temporal dimension, denoted as $W_{up} = [2 * C, C]$. Then we initialize it as:

$$W_{\text{up}}[2 * C, C] = \text{concat}(\mathbf{I}, \mathbf{I}). \tag{4}$$

As shown in Fig. 2 (a), when processing an input sequence consisting of frames {0, 1, 2, 3}, the output of the encoder corresponds to the latent representation of frame {3}, which is selected as the keyframe. Meanwhile, the decoder broadcasts the latent representation of {3} across the entire sequence, resulting in {3, 3, 3, 3}. At the start of training, the model's behavior is to select a keyframe from a video clip, encoding and decoding it using the powerful inherited image VAE and subsequently broadcasting this static frame to reconstruct the entire clip. Consequently, the learning objective is simplified: instead of learning complex visual features from scratch, the model only needs to learn the temporal dynamics required to animate the static keyframe into a dynamic sequence. This drastically reduces training costs and provides a more efficient learning task.

3.3 Temporal Dynamic Difference Convolution

With a strong spatial prior established by our keyframe-centered initialization, the model's primary learning objective shifts from reconstructing entire frames to capturing sparse temporal dynamics. To explicitly enforce this learning paradigm, we introduce the Temporal Dynamic Difference Convolution (TDC), an operator designed to follow the efficient prediction-residual design principle of video codecs. Unlike standard convolutions that process entangled features, TDC operates on the explicit feature differences between frames. It creates two distinct pathways: one that preserves the static content from the keyframe prior, and a parallel one dedicated to learning a sparse motion residual. As illustrated in Fig. 2, we incorporate specialized variants of TDC into the different blocks of our Video VAE to implement this decoupled strategy.

TDC decomposes the features within its receptive field into a static anchor frame and a set of dynamic difference features, forcing the network to learn from motion explicitly. When TDC is integrated into a Group Causal Convolution Resblock, the central frame X_t of the group naturally serves as the static anchor. The TDC operator is formulated to be equivalent to a standard convolution while the kernel is explicitly re-parameterized for separate modeling, with no additional parameters or computational overhead:

$$Y_t = W_a * X_t + Wp * (X_{t-1} - X_t) + W_f * (X_{t+1} - X_t) \tag{5}$$

Here, W_a, W_p, W_f are the temporal slices of the 3D kernel. The $W_a * X_t$ term, backed by our image VAE initialization, preserves the anchor frame's spatial

Fig. 4: Visualization of the high frequency details reconstruction performance.

information, while the other terms are forced to learn from the motion residuals between the anchor and its neighbors, as shown in Fig. 2(b). We further apply TDC to model inter-frame interactions. Our temporal interaction module operates on two distinct streams: the anchor X_a itself and the motion residuals $(X_p - X_a)$. The final output is a learned combination of these two streams:

$$Y_{output} = W_a * X_a + W_p * (X_p - X_a) \tag{6}$$

In this way, the model is explicitly guided to distinguish between static scene structure and temporal dynamics, as shown in Fig. 2(c). For Causal Convolutions, the anchor is always the current frame X_t. The operation thus learns from the differences between past frames and the current frame:

$$Y_t = W_t * X_t + W_{t-1} * (X_{t-1} - X_t) + W_{t-2} * (X_{t-2} - X_t) \tag{7}$$

This forces the model to encode the current frame's content while explicitly learning from the historical residuals within its receptive field. This design ensures that at the start of training, its output is mathematically equivalent to that of the model initialized with the Image VAE, and it preserves the static reconstruction capability inherited from the image VAE throughout training, leading to enhanced stability and superior results.

3.4 Other Implementation Refinement

We identified a common issue in current causal Video VAEs [29, 34]: the first frame of a video sequence often has poorer reconstruction quality than later frames. This problem likely stems from the widely used Separate First-Frame Processing approach, which creates an architectural asymmetry since the first frame is processed differently due to the lack of prior temporal context. As illustrated in Fig. 6, this can cause a noticeable performance gap between first frame and the following sequence. To prevent our model from suffering from this flaw, we implement an In-Sequence First-Frame Processing pipeline that removes the special treatment of the first frame. This approach also unifies the processing of single images and video sequences by treating all inputs equally: a single image is turned into a minimal pseudo-video by repeating the frame. Consequently, all frames are processed with the same encoding-decoding process, eliminating the reconstruction inconsistencies caused by the Separate First-Frame Processing.

Table 2: Results on TokBench Video. T-ACC, T-NED measure the accuracy of text reconstruction, while F-Sim measures the similarity of face reconstruction. ↑ indicates higher is better.

Method	FCR	T-ACC ↑				T-NED ↑				F-Sim ↑			
		small	medium	large	mean	small	medium	large	mean	small	medium	large	mean
Resolution: 256x													
CogvideoX	4*8*8	24.80	72.47	86.34	61.21	43.06	82.29	92.41	72.59	0.58	0.78	0.91	0.76
Wanx2.1	4*8*8	17.88	69.52	87.56	58.32	37.27	81.04	93.18	70.50	0.59	0.79	0.92	0.77
IVVAE	4*8*8	18.43	72.08	89.52	60.01	38.29	82.49	94.36	71.72	0.57	0.79	0.93	0.76
Hunyuan	4*8*8	26.85	69.12	87.47	61.15	45.55	80.54	93.12	73.07	0.60	0.80	0.92	0.77
VC-VAE	4*8*8	**31.81**	**77.21**	**89.70**	**66.24**	**51.21**	**85.24**	**94.39**	**76.95**	**0.64**	**0.82**	**0.93**	**0.80**
Resolution: 480x													
CogvideoX	4*8*8	28.02	65.41	91.71	61.71	43.47	78.24	95.60	72.43	0.67	0.80	0.91	0.79
Wanx2.1	4*8*8	18.88	63.42	92.36	58.22	35.95	77.43	96.02	69.80	0.68	0.82	0.92	0.81
IVVAE	4*8*8	20.58	63.89	**92.98**	59.15	37.85	78.07	**96.37**	70.76	0.68	0.82	0.92	0.80
Hunyuan	4*8*8	28.65	64.49	91.83	61.66	44.43	77.83	95.83	72.70	0.69	0.82	0.92	0.81
VC-VAE	4*8*8	**31.24**	**72.27**	92.72	**65.41**	**48.11**	**82.90**	96.24	**75.75**	**0.72**	**0.83**	**0.92**	**0.83**

Table 3: Effectiveness of the TDC Operator.

Setting	PSNR↑	SSIM↑	LPIPS↓
Causal Conv Baseline	33.29	0.9298	0.04946
+ TDC	33.64	0.9336	0.04599
GCConv Baseline	33.49	0.9323	0.04720
+ TDC (Full model)	**33.79**	**0.9357**	**0.04455**

Table 4: Impact of the Image VAE's Performance.

Setting	PSNR↑	SSIM↑	LPIPS↓
SD3.5-VAE Init w/o TDC	33.13	0.9264	0.05519
Flux-VAE Init w/o TDC	33.49	0.9323	0.04720
SD3.5-VAE Init w TDC	33.74	0.9337	0.04691
Flux-VAE Init w TDC	**33.79**	**0.9357**	**0.04455**

4 Experiments

4.1 Experimental Setup

Training details. We train our Video VAE on the Kinetics-600 dataset [4] using a two-stage approach. We first pretrain the model for 500k steps, focusing on learning fundamental spatio-temporal dynamics from a large volume of short-duration, low-resolution videos, optimized with a combination of reconstruction losses (L1, LPIPS [40]) and KL regularization [12]. This is followed by a 50k steps fine-tuning stage on a diverse mix of videos with varying resolutions, frame rates, and durations. In this stage, we incorporate an adversarial loss from a 3D discriminator to improve high-frequency details and strengthen temporal extrapolation capabilities. As shown in Table 5, we compare with the total training steps required by state-of-the-art models. Notably, our total training compute is con-

Table 5: Comparison of the total training computation and reconstruction performance.

Method	Training steps	PSNR
WFVAE	1200k	35.60
IVVAE	1000k	35.21
VC-VAE	550k	36.04

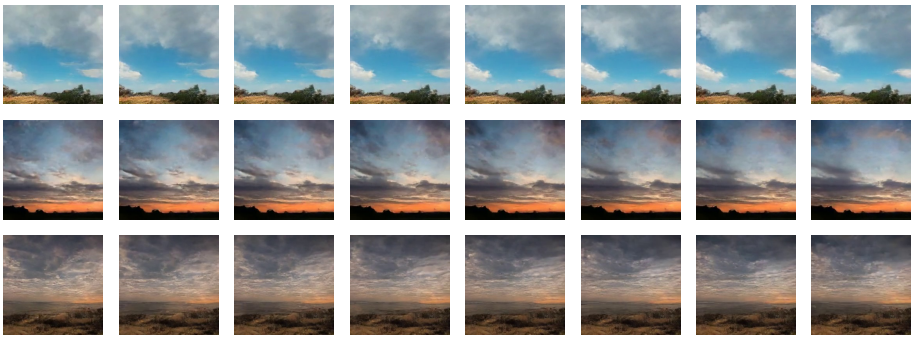

Fig. 5: Generative results from a Latent Video Diffusion Model built upon our proposed VC-VAE.

siderably lower than that of state-of-the-art models, highlighting the training efficiency of our proposed approach.

For the 4×8×8 compression version, which downsamples the video by a factor of 4 in the temporal dimension and 8 in each spatial dimension, we use 16 channels identical to most of the SOTA Video VAEs, and utilize Flux VAE [15] for initialization. For the 8×8×8 compression version, we employ a 32-channel implementation, where the first half is initialized from Flux VAE and the remaining half is zero-initialized. This strategy ensures that, after increasing the number of channels, our model begins with an initial image reconstruction capability identical to Flux VAE. Detailed hyperparameters, including learning rates and loss weights for each stage, are provided in the supplementary material.

Evaluation details. We employ WebVid-10M [2] and UCF-101 [26] to assess overall video reconstruction performance. To ensure an fair comparison, all metrics are calculated without the first frame for baseline models. We utilize FVD [28], PSNR [11], SSIM [31], and LPIPS [40] to measure reconstruction quality. Furthermore, we assess the VAE's ability to reconstruct text and faces within videos using the TokBench-Video [33] benchmark, with performance evaluated by reconstruction text accuracy and face similarity [8].

4.2 Performance

Our baseline models include several 4×8×8 SOTA methods: CogVideoX-VAE [37], Wan2.1-VAE [29], WFVAE [16], IVVAE [34], and Hunyuan-VAE [13]. Additionally, WFVAE provides an 8×8×8 version, which serves as the baseline for our experiments on high-ratio temporal compression.

Quantitative evaluation of reconstruction results. We present a comprehensive quantitative evaluation of our VC-VAE against several SOTA methods in Tab. 1, with results reported on both WebVid-10M and UCF-101. Under the standard 4×8×8 compression setting, with significantly lower training compute, our VC-VAE achieves SOTA performance across all datasets. We also evaluate a more challenging configuration with a higher temporal compression rate,

where our model continues to outperform the WFVAE baseline. These results validate the effectiveness and robustness of our proposed architecture.

Furthermore, to provide a more comprehensive analysis, we test our Video VAE on the challenging TokBench-Video benchmark. This benchmark is particularly demanding as it comprises long-duration videos with complex motion and intricate, high-frequency text. Because our method is specifically designed to enhance the modeling of both temporal dynamics and high-frequency details, our model achieves significant performance gains on this benchmark, especially in the reconstruction of small to medium-sized text and faces, as shown in Tab. 2.

Quantitative Evaluation of Generation Results. To evaluate the quality of the latent space and its suitability for downstream generative tasks, we conducted video generation experiments following IVVAE [34]. Specifically, we trained an unconditional diffusion model using Latte-XL [19] within our VC-VAE latent space and generated 2,000 video samples for evaluation.

The generation quality is measured using the generalized Fréchet Video Distance (gFVD). As shown in Tab.6, our model achieves a gFVD score that is superior to the baseline. These quantitative results demonstrate that VC-VAE does not merely excel at reconstruction, but also structures a more favorable latent distribution for diffusion-based generation. The competitive gFVD scores serve as direct evidence of our model's reliability as a backbone for high-fidelity video generation system.

Table 6: Comparison of gFVD on SkyTimelapse.

Method	gFVD↓
ODVAE	140.0
IVVAE	118.6
VC-VAE	112.8

Qualitative evaluation of reconstruction results. Our method's superior reconstruction capabilities are demonstrated across diverse and challenging scenes. Fig. 3 shows a sports scene with rapid motion where our method effectively captures movement while maintaining the stability of fine text details, unlike other methods that suffer from artifacts. Furthermore, in scenes demanding high-frequency texture reconstruction as shown in Fig. 4, our model continues to outperform baselines, producing visibly sharper results. This offers qualitative evidence for the effectiveness of our proposed TDC operator in decoupling static information from dynamic motion. Additional reconstruction results are included in the supplementary material.

4.3 Ablation Study

In this section, we conduct a series of ablation experiments. For the ablation study, we first initialize Video VAE with Flux VAE, then train it for 100K steps on 16-frame videos with 256×256 resolution using the Kinetics-600 dataset. The test dataset comprises 2000 videos from the Kinetics-600 validation set, each with 16-frames at 256×256 resolution.

Ablation of main components. We conduct ablation studies to verify the effectiveness of our TDC operator. First, we replace the core temporal operators in the standard Causal Convolution and Group Causal Convolution baselines with our TDC operator. As shown in Tab. 3, this substitution delivers significant

Table 7: Ablation on First-Frame Processing Strategy.

Method	PSNR↑	SSIM↑	LPIPS↓
Separate	32.15	0.9242	0.04059
In-Sequence	33.34	0.9346	0.03601

Table 8: Ablation on TDC operator under challenging compression rate.

Method	FCR	PSNR↑	SSIM↑	LPIPS↓
GCConv w/o TDC	8*8*8	31.11	0.9018	0.07641
GCConv w TDC	8*8*8	32.01	0.9122	0.07099

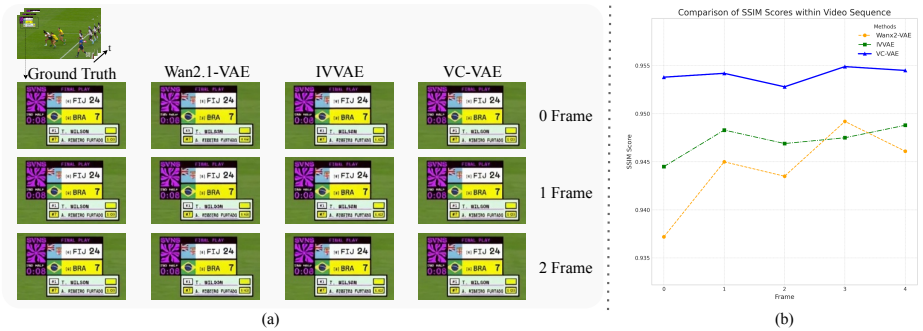

Fig. 6: (a). Visualization of the video sequence's reconstruction performance. (b). Average SSIM scores over the first five frames, calculated across 1000 videos from the validation set, where baseline models often exhibit lower reconstruction quality on the initial frame, while our VC-VAE maintains high and stable performance.

improvements in both PSNR and LPIPS metrics. This demonstrates that our design, inspired by video codec compression, substantially enhances the model's ability to preserve static details inherited from the ImageVAE while effectively modeling temporal dynamics. The consistent performance gains observed across different architectures confirm the generality of our approach. Furthermore, to evaluate the operator under more challenging conditions, we extend our analysis to models with higher temporal compression rates. As reported in Tab. 8, our method yields even more pronounced performance advantages as the compression rate increases, fully reinforcing the superiority and robustness of the TDC operator.

Ablation on First Frame Processing Strategy. To validate the effectiveness of our In-Sequence First-Frame Processing strategy, we compare our implementation with the widely used Separate First-Frame Processing variant. As shown in Tab. 7, by adopting the In-Sequence First-Frame Processing, our model achieves a 1.2 dB improvement in PSNR for the first frame, effectively resolving the inconsistency. More evaluation of our model's performance on single-image reconstruction is provided in the supplementary material.

Impact of the Initializing Image VAE. To evaluate our model's sensitivity to the quality of pre-trained Image VAE weights, we perform an ablation study using two image VAEs with different performance levels: a stronger Flux-VAE and a comparatively weaker SD3.5-VAE [9]. As shown in Tab. 4, when training a baseline model without TDC, performance strongly depends on the

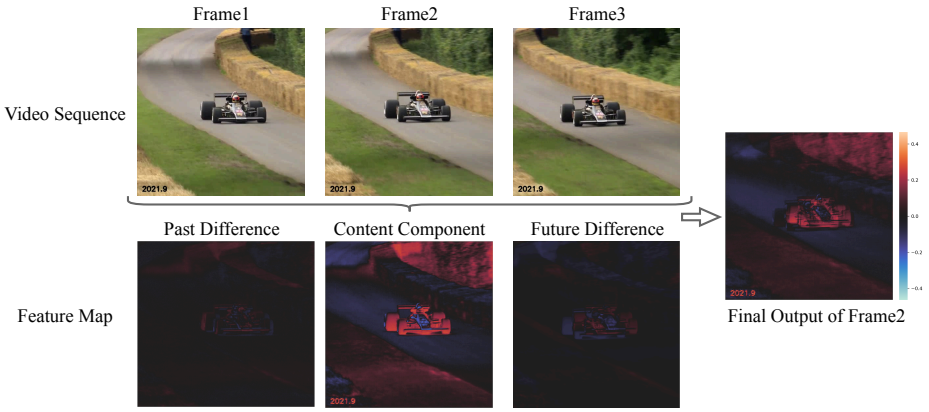

Fig. 7: Visulization of the proposed TDC operator.

quality of initialization; the model initialized with the stronger Flux-VAE notably outperforms the one starting from SD3.5-VAE. However, when our TDC operator is added, the performance gap between the two setups decreases significantly. Remarkably, adding TDC results in a more notable improvement for the model initialized with the weaker SD3.5-VAE, with a PSNR increase of 0.61 dB. This indicates that TDC not only enhances overall performance but also makes training more robust and less sensitive to the quality of initial spatial priors.

4.4 Visualization

Visualization of TDC. To intuitively understand how our Temporal Dynamic Difference Convolution operator explicitly separates the sparse motion residual from the static content, we perform visualization of its internal components. As shown in Fig. 7, we use three frames from a video sequence with significant motion to clearly illustrate the effect. The anchor component focuses solely on modelling the intra-frame spatial information of the current frame, successfully encoding the representation of static structures, including the car's appearance, the background, and the watermark, while the other components, derived from the temporal differences, isolate inter-frame temporal changes. Static regions such as the watermark and stationary background elements exhibit near-zero activation. Conversely, activations are concentrated around the moving object, effectively encoding its displacement between frames. These difference maps can be interpreted as learned, high-level motion representations, highlighting where and how the scene changes over time. The final output synthesizes these components by combining the rich spatial details from the content component with the precise motion information from the difference components, demonstrating that our TDC operator does not merely mix temporal information but explicitly disentangles the representation of content from its dynamics, enabling a more efficient and interpretable modelling of video.

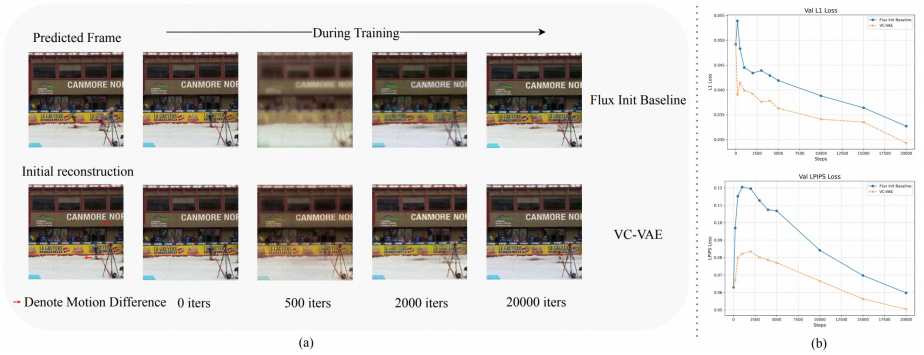

Fig. 8: (a). Visual comparison of reconstructed frames at different training iterations. While both models start with a perfect reconstruction (0 iters), the baseline quickly degrades and becomes blurry, whereas our model preserves high-frequency details throughout training. (b). Corresponding validation L1 and LPIPS loss curves of VC-VAE and baseline.

Visulization of First Frame Performance. As discussed in the method section, causal Video VAEs often suffer from inconsistent reconstruction quality, particularly for the first frame. Fig. 6 provides a clear qualitative demonstration of this phenomenon. In the 0 Frame column, it is evident that baseline models like Wan2.1-VAE and IVVAE exhibit significant degradation. This quality drop appears confined to the initial frame, as their performance on subsequent frames is visibly better, highlighting the performance discrepancy stemming from architectural asymmetry. In contrast, our VC-VAE maintains a consistently high level of reconstruction fidelity across all frames. This robustness is a direct result of our In-Sequence First-Frame Processing strategy, which eliminates the performance degradation of the initial frame. This visualization confirms that our approach effectively resolves this common phenomenon in causal Video VAEs and ensures stable, high-quality performance throughout the entire video sequence.

Analysis of the training process. We visualize the training process of the Video VAE initialized from a pre-trained image VAE to analyse its training process. As shown in Fig. 8(a), the baseline model, in its effort to learn inter-frame dynamics, experiences a severe performance degradation in the early training stages. The static content becomes heavily blurred, indicating that the model suffers from catastrophic forgetting of the powerful spatial priors provided by the image VAE. Our proposed VC-VAE maximally preserves the static reconstruction capability of the initial VAE while isolating the learning of the temporal dynamics. This is quantitatively confirmed by the training curves in Fig. 8(b). The baseline model's validation loss noticeably spikes at the beginning of training, whereas our method exhibits a stable and monotonic decrease, effectively mitigating the initial reconstruction degradation.

Generative Performance with VC-VAE. Following previous work [7,34], we demonstrate the video generation capability of our Video VAE's latent space.

We train an unconditional latent video diffusion model, using Latte-XL [19] as the generative backbone, on the SkyTimelapse [36] dataset. The qualitative results are visualized in Fig. 5. Leveraging our VC-VAE, the video diffusion model is capable of generating high-fidelity videos that exhibit both complex temporal dynamics and intricate textures.

5 Conclusion

Inspired by video codecs, which use keyframes to encode static content and motion residuals for inter-frame dynamics, we propose VC-VAE, a novel architecture that integrates this principle of decomposition into the design of Video VAEs. First, by initializing with a high-quality ImageVAE, we endow the model with strong keyframe reconstruction capabilities, ensuring the quality of its static spatial priors. Second, we design the Temporal Dynamic Difference Convolution operator, which explicitly computes temporal differences within convolution, successfully decoupling static spatial content from dynamic temporal motion information. The TDC operator significantly boosts the model's performance without introducing any additional parameters or computational costs, particularly enhancing its ability to model high-frequency motion details. Extensive experiments show that the proposed VC-VAE achieves SOTA results in video reconstruction and generation.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 62576076) and the Guangdong Basic and Applied Basic Research Foundation (Grant No. 2023A1515140037).

References

1. Agarwal, N., Ali, A., Bala, M., Balaji, Y., Barker, E., Cai, T., Chattopadhyay, P., Chen, Y., Cui, Y., Ding, Y., et al.: Cosmos world foundation model platform for physical ai. arXiv preprint arXiv:2501.03575 (2025) 2
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) 10
3. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023) 4
4. Carreira, J., Noland, E., Banki-Horvath, A., Hillier, C., Zisserman, A.: A short note about kinetics-600. arXiv preprint arXiv:1808.01340 (2018) 9
5. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017) 6

6. 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) 3
7. Chen, L., Li, Z., Lin, B., Zhu, B., Wang, Q., Yuan, S., Zhou, X., Cheng, X., Yuan, L.: Od-vae: An omni-dimensional video compressor for improving latent video diffusion model. arXiv preprint arXiv:2409.01199 (2024) 4, 6, 14
8. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4690–4699 (2019) 10
9. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: Forty-first International Conference on Machine Learning (2024) 12
10. 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) 3
11. Hore, A., Ziou, D.: Image quality metrics: Psnr vs. ssim. In: 2010 20th international conference on pattern recognition. pp. 2366–2369. IEEE (2010) 10
12. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013) 3, 9
13. 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) 1, 10
14. 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) 3
15. Labs, B.F.: Flux. Online (2024), https://github.com/black-forest-labs/flux 3, 10
16. Li, Z., Lin, B., Ye, Y., Chen, L., Cheng, X., Yuan, S., Yuan, L.: Wf-vae: Enhancing video vae by wavelet-driven energy flow for latent video diffusion model. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 17778–17788 (2025) 2, 4, 10
17. Lin, X., Tang, W., Wang, H., Liu, Y., Ju, Y., Wang, S., Yu, Z.: Exposing image splicing traces in scientific publications via uncertainty-guided refinement. Patterns **5**(9) (2024) 4
18. Ma, G., Huang, H., Yan, K., Chen, L., Duan, N., Yin, S., Wan, C., Ming, R., Song, X., Chen, X., et al.: Step-video-t2v technical report: The practice, challenges, and future of video foundation model. arXiv preprint arXiv:2502.10248 (2025) 1
19. Ma, X., Wang, Y., Jia, G., Chen, X., Liu, Z., Li, Y.F., Chen, C., Qiao, Y.: Latte: Latent diffusion transformer for video generation. arXiv preprint arXiv:2401.03048 (2024) 11, 15
20. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023) 4
21. Polyak, A., Zohar, A., Brown, A., Tjandra, A., Sinha, A., Lee, A., Vyas, A., Shi, B., Ma, C.Y., Chuang, C.Y., et al.: Movie gen: A cast of media foundation models. arXiv preprint arXiv:2410.13720 (2024) 1, 2
22. Procházka, A., Gráfová, L., Vyšata, O., Caregroup, N.: Three-dimensional wavelet transform in multi-dimensional biomedical volume processing. In: Proc. of the IASTED International Conference on Graphics and Virtual Reality, Cambridge. vol. 263, p. 268 (2011) 2

23. 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) 1, 3
24. Seawead, T., Yang, C., Lin, Z., Zhao, Y., Lin, S., Ma, Z., Guo, H., Chen, H., Qi, L., Wang, S., et al.: Seaweed-7b: Cost-effective training of video generation foundation model. arXiv preprint arXiv:2504.08685 (2025) 1
25. Skorokhodov, I., Girish, S., Hu, B., Menapace, W., Li, Y., Abdal, R., Tulyakov, S., Siarohin, A.: Improving the diffusability of autoencoders. arXiv preprint arXiv:2502.14831 (2025) 3
26. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012) 10
27. Tian, R., Dai, Q., Bao, J., Qiu, K., Yang, Y., Luo, C., Wu, Z., Jiang, Y.G.: Reducio! generating 1k video within 16 seconds using extremely compressed motion latents. arXiv preprint arXiv:2411.13552 (2024) 4
28. Unterthiner, T., Van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: Fvd: A new metric for video generation (2019) 10
29. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025) 1, 8, 10
30. Wang, Y., Guo, J., Xie, X., He, T., Sun, X., Bian, J.: Vidtwin: Video vae with decoupled structure and dynamics. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 22922–22932 (2025) 4
31. 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 **13**(4), 600–612 (2004) 10
32. Wiegand, T., Sullivan, G.J., Bjontegaard, G., Luthra, A.: Overview of the h. 264/avc video coding standard. IEEE Transactions on circuits and systems for video technology **13**(7), 560–576 (2003) 2
33. Wu, J., Luo, D., Zhao, W., Xie, Z., Wang, Y., Li, J., Xie, X., Liu, Y., Bai, X.: Tokbench: Evaluating your visual tokenizer before visual generation. arXiv preprint arXiv:2505.18142 (2025) 10
34. Wu, P., Zhu, K., Liu, Y., Zhao, L., Zhai, W., Cao, Y., Zha, Z.J.: Improved video vae for latent video diffusion model. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 18124–18133 (2025) 2, 4, 5, 6, 8, 10, 11, 14
35. Wu, Y., Li, Y., Skorokhodov, I., Kag, A., Menapace, W., Girish, S., Siarohin, A., Wang, Y., Tulyakov, S.: H3ae: High compression, high speed, and high quality autoencoder for video diffusion models. arXiv preprint arXiv:2504.10567 (2025) 4
36. Xiong, W., Luo, W., Ma, L., Liu, W., Luo, J.: Learning to generate time-lapse videos using multi-stage dynamic generative adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2364–2373 (2018) 15
37. 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. arXiv preprint arXiv:2408.06072 (2024) 2, 4, 5, 10
38. Yao, J., Yang, B., Wang, X.: Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 15703–15712 (2025) 3
39. Yu, L., Lezama, J., Gundavarapu, N.B., Versari, L., Sohn, K., Minnen, D., Cheng, Y., Birodkar, V., Gupta, A., Gu, X., et al.: Language model beats diffusion–

tokenizer is key to visual generation. arXiv preprint arXiv:2310.05737 (2023) 2, 4, 5
40. 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) 9, 10
41. 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) 4
42. Zhou, L., Ruan, C., Ling, N., Chen, Z., Wang, W., Jiang, W.: Tvc: tokenized video compression with ultra-low bit rate. Visual Intelligence **3**(1), 25 (2025) 3
43. Zou, Y., Yao, J., Yu, S., Zhang, S., Liu, W., Wang, X.: Turbo-vaed: Fast and stable transfer of video-vaes to mobile devices. arXiv preprint arXiv:2508.09136 (2025) 3