HiTVideo: Hierarchical Tokenizers for Enhancing Text-to-Video Generation with Autoregressive Large Language Models

Ziqin Zhou[1,2*], Yifan Yang[2†], Yuqing Yang[2], Tianyu He[2], Houwen Peng[2], Kai Qiu[2], Qi Dai[2], Lili Qiu[2], Chong Luo[2], Lingqiao Liu[1†]

[1]The University of Adelaide
[2]Microsoft Research Asia
{ziqin.zhou, lingqiao.liu}@adelaide.edu.au, yifanyang@microsoft.com

Abstract

Text-to-video generation poses significant challenges due to the inherent complexity of video data, which spans both temporal and spatial dimensions. It introduces additional redundancy, abrupt variations, and a domain gap between language and vision tokens while generation. Addressing these challenges requires an effective video tokenizer that can efficiently encode video data while preserving essential semantic and spatiotemporal information, serving as a critical bridge between text and vision. Inspired by the observation in VQ-VAE-2, we propose **HiTVideo**, a novel approach for text-to-video generation with hierarchical tokenizers. It utilizes a 3D causal VAE with a multi-layer discrete token framework, encoding video content into hierarchically structured codebooks. Higher layers capture semantic information with higher compression, while lower layers focus on fine-grained spatiotemporal details, striking a balance between compression efficiency and reconstruction quality. Our approach efficiently encodes longer video sequences (e.g., 8 seconds, 64 frames), reducing bits per pixel (bpp) by approximately 70% compared to previous tokenizers, while maintaining competitive reconstruction quality. We explore the trade-offs between compression and reconstruction, while emphasizing the advantages of high-compressed semantic tokens in text-to-video tasks. HiTVideo aims to address the potential limitations of existing video tokenizers in text-to-video generation tasks, striving for higher compression ratios, improved token quality, and simplify LLMs modeling under language guidance, offering a scalable and promising framework for advancing text to video generation.

Code — https://ziqinzhou66.github.io/project/HiTVideo
Datasets — https://www.pexels.com/search/videos/videos

Introduction

Visual generation is a challenging and rapidly evolving task that has seen significant advancements over the years. Initially, GANs (Generative Adversarial Networks) (Goodfellow et al. 2014) shows impressive capabilities in creating realistic images (Karras 2017) and videos (Skorokhodov, Tulyakov, and Elhoseiny 2022). In recent years, the development of diffusion models (Rombach et al. 2022; Ho et al. 2022a) and Large Language Models (LLMs) (Raffel et al. 2020; Sanh et al. 2021) has provided new solutions for tackling visual generation tasks, especially for generating content based on human instructions, such as text-to-video generation (He et al. 2022; Hong et al. 2022). While diffusion models achieve high visual fidelity through denoising, LLMs offer an alternative that enables unified multimodal generation with efficient sampling and broader task generalization.

Leveraging autoregressive large language models (LLMs) to generate videos natively, rather than relying on diffusion-based decoders, offers natural advantages: temporal causality can be modeled explicitly, and the model's intrinsic ability to follow complex language instructions is retained throughout generation. However, two obstacles limit practical adoption. First, video carries vast information—token counts quickly explode. For instance, encoding a standard 360p video at 33 fps with an $8\times8\times4$ VAE compression as used in WAN 2.1 (Wan et al. 2025) would demand roughly 22k tokens per second, far exceeding mainstream context windows and rendering large-scale pre-training prohibitively costly. Consequently, we target the design of a much higher-compression video tokenizer that brings autoregressive generation into a tractable cost regime. Second, textual prompts are inherently high-level abstractions, whereas low-level video tokens remain semantically distant; predicting them in a raster-scan, frame-by-frame fashion is inefficient and often ineffective. Human animators typically draft coarse storyboards before refining details. Inspired by this workflow, we advocate an autoregressive ordering that first generates high-level semantic tokens and then progressively fills in fine-grained details, enabling LLM-based video synthesis to proceed more naturally and with markedly improved efficiency.

Despite these developments, training generative models for text-to-video generation presents significant challenges, particularly in visual encoding and multimodal alignment. Unlike image data, video sequences have greater redundancy and variability across frames, complicating the encoding process and making video tokenization especially challenging. Additionally, the gap between text tokens and vi-

*This work was completed during an internship at Microsoft Research Asia.
†Corresponding authors.

sual tokens is a major barrier in aligning language and visual representations for generation tasks. A well-designed tokenizer can better represent semantic content, a critical requirement for generating coherent tokens cross language to visual. Related video autoencoder, such as MAGVIT-V2 (Yu et al. 2023b), have demonstrated the utility of dynamic masking strategies in enhancing generation efficiency and reconstruction quality (Chang et al. 2022). Similarly, VAR (Tian et al. 2024) has shown that a coarse-to-fine strategy improves text-to-image generation.

To enable human-animation–style workflows—where animators sketch a high-level storyboard before refining details—we introduce a **multi-layer codebook tokenizer** as shown in Fig. 1. Inspired by the hierarchical scheme of VQ-VAE-2 (Razavi, Van den Oord, and Vinyals 2019), our encoder produces a stack of discrete latent maps at progressively finer spatiotemporal scales, each endowed with its own vocabulary that captures information at a specific level of abstraction. During decoding we first reconstruct a highly compressed, semantic overview, and then iteratively inject residual latent maps from successive layers, adding high-frequency details in a codec-like coarse-to-fine fashion. This procedure mirrors the "draft-then-polish" pipeline of traditional animation, preserves global coherence, and keeps the total token budget tractable for autoregressive language models while still delivering visually rich results.

Our contributions are as follows: **Design Hierarchical Video Tokenizer**: We design a multi-layer codebook that balances reconstruction quality and compression efficiency. By encoding videos at different granularities, high-level tokens capture semantic content, while lower-level tokens reconstruct spatiotemporal details.

Increase Generation Performance and Compression Ratio: We demonstrate that multi-layer tokenizers significantly improve generation performance compared to single-layer approaches. The high-level semantic tokens facilitate a smoother transition between textual and visual representations, leading to better alignment in text-to-video tasks.

Dynamic Encoding and Masked Decoding: We proposes dynamic encoding mechanisms that reduce redundancy and proves masked decoding to improve adaptability.

Our experiments validate the effectiveness of this approach, showcasing strong performance in both reconstruction and generation. Compared to existing methods, our hierarchical tokenizer enables better semantic representation under higher compression, facilitating efficient and coherent text-to-video generation. These results highlight the potential of our design for advancing scalable and multimodal generative frameworks.

Related Work

Vision Tokenizer

The original VQ-VAE (Van Den Oord, Vinyals et al. 2017) model introduced the concept of vector quantization for images, encoding continuous visual data into discrete codes that could be used in autoregressive modeling. VQ-VAE-2 (Razavi, Van den Oord, and Vinyals 2019) extended this to hierarchical representations, improving the quality and resolution of generated images. This framework laid the groundwork for transforming images into discrete tokens that could be processed similarly to text. DALL-E (Ramesh et al. 2021) utilized dVAE to tokenize images into discrete codes, leveraging the transformer architecture for image synthesis.

Method	Input size	Tokens	Compression Ratio
MAGVIT-v2	17×128^2	1280	217.6
CogVideoX	-	-	256
EMU3	4×512^2	4096	256
Ours	64×256^2	2448	**1713.4**

Table 1: Video tokenizers compression ratio comparison.

Building on VQ-VAE, VideoGPT (Yan et al. 2021) applied tokenization to video data, encoding each frame into discrete tokens to support autoregressive generation across temporal sequences. MAGVIT (Yu et al. 2023a) further refined this approach by introducing a masked generative video transformer, which encodes videos into multi-layer tokens, enhancing the representation of complex video scenes. MAGVIT-v2 (Yu et al. 2023b) improved upon MAGVIT by introducing a lookup-free quantization approach, enhancing reconstruction fidelity even in large vocabularies.

Generation Based on Diffusion Models

Diffusion models (Ho, Jain, and Abbeel 2020; Nichol and Dhariwal 2021; Rombach et al. 2022) have become central to generative tasks for their ability to produce high-quality images (Saharia et al. 2022) and videos (Blattmann et al. 2023; Chen et al. 2024) via iterative denoising. Recent architectures like DiT (Peebles and Xie 2023) and U-ViT (Bao et al. 2023) evolve from U-Net (Ronneberger, Fischer, and Brox 2015), improving scalability for spatiotemporal modeling. In text-to-video generation, Make-a-Video (Singer et al. 2022) repurposes text-to-image diffusion models for unpaired video synthesis. VDM (Ho et al. 2022b) introduces spatiotemporal conditioning, while Latent-Shift (Blattmann et al. 2023) extends Stable Diffusion through latent-space temporal shifts. VideoCrafter (Chen et al. 2024) applies a two-stage strategy with temporal modules to enhance coherence. MicroCinema (Wang et al. 2024b) further improves efficiency via hierarchical latents. Recent systems such as SORA (Brooks et al. 2024), Vidu (Bao et al. 2024), and Lumiere (Bar-Tal et al. 2024) showcase HD-level long-form video generation. CogVideoX (Yang et al. 2024) employs a hierarchical coarse-to-fine framework with text-conditioned token modeling and Wan (Wan et al. 2025) proposed .

Generation Based on LLM Models

Large Language Models (LLMs) (Touvron et al. 2023; Achiam et al. 2023) have emerged as a powerful alternative to diffusion models for generative tasks, offering faster sampling (Brown 2020), better scalability to large datasets, and strong capabilities in multimodal alignment (**?**Wang et al. 2024a). In contrast to diffusion models' iterative denoising, LLMs generate content by predicting tokens either autoregressively (Chen et al. 2020) or through masked modeling (Devlin 2018), enabling flexible and efficient generation.

Masked prediction frameworks, such as MaskGit (Chang et al. 2022) and MagViT-v2 (Yu et al. 2023b), refine content by iteratively filling in missing tokens, offering efficient parallelism and enhanced spatial-temporal consistency. These models are especially effective in multi-task learning setups and benefit from lookup-free quantization and token reuse strategies. Autoregressive LLMs, on the other hand, predict tokens sequentially, capturing long-range dependencies and producing coherent outputs. LlamaGen (Sun et al. 2024) and VAR (Tian et al. 2024) demonstrate strong performance in image generation using coarse-to-fine prediction. In the video domain, works like VideoPoet (Kondratyuk et al. 2023), Show-O (Xie et al. 2024), and EMU3 (Wang et al. 2024a) extend autoregressive modeling to text-to-video generation. These models integrate multiple modalities—including text, images, and audio—and generate consistent and high-quality video sequences.

Method

Rethinking multi-modal generation

Recent advances in multi-modal generative models show promise for unified framework (Wu et al. 2024; Wang et al. 2024a), and the typical pipeline framework consists of two main training stages:

Stage 1: Train visual autoencoder;

Stage 2: Train a generative model (e.g., diffusion or LLM) to predict generation results.

However, video generation faces key challenges:

Complexity of Video Data: Videos are inherently more complex than images, as both spatial and temporal dimensions introduce redundancy and variation across frames.

Encoding Trade-Offs: The autoencoder must balance detail preservation and computational efficiency, with designs tailored to specific use cases, though challenges may arise in certain downstream applications.

Modality Gap: Bridging the structural and semantic gap between text and video tokens remains a significant hurdle.

These challenges highlight the need for autoencoders that balance high compression, reconstruction quality, and effective text-video alignment to support generative tasks.

Compared to diffusion models, large language models (LLMs) exhibit flexibility in handling various input-output configurations (Yan et al. 2021; Wu et al. 2024; Wang et al. 2024a). To validate the effectiveness of our hierarchical video tokenizer in bridging the gap between text and visual modalities, we adopt one of the most challenging downstream tasks: autoregressive next-token prediction for long video sequences using an LLM. Our experiments demonstrate that the hierarchical video tokenizer offers significant advantages in text-to-video generation, achieving high compression ratios while preserving superior reconstruction quality. Unlike single-layer codebooks, which often struggle under high compression, the multi-layer design produces coherent and competitive results, effectively capturing key semantic prompts. Additionally, hierarchical structure simplifies LLM modeling by providing compact visual tokens conditioned with text prompts, while residual dense tokens refine fine-grained details during generation. Rather than aiming to create the best autoregressive LLM with post-training modifications, this work emphasizes the potential of high-compression tokenizers as a robust intermediary for multi-modal tasks.

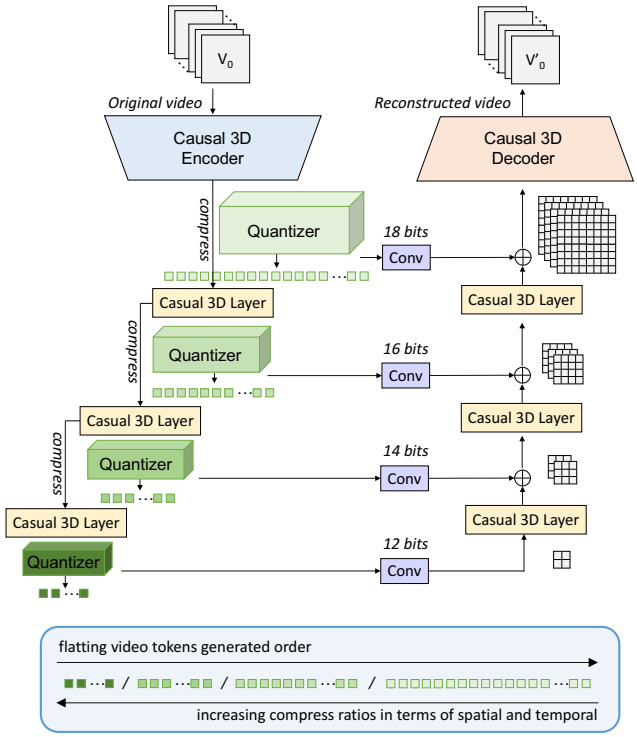


Figure 1: The overall architecture of **HiTVideo** tokenizer.

Multi-layer discrete tokenizers design

Tokenizer configuration: Previous vision tokenizers based on VQ-VAE and VQ-GAN have shown the feasibility of generating images and videos from text prompts (Kondratyuk et al. 2023; Tian et al. 2024; Sun et al. 2024) and demonstrated potential for unified cross-modality token understanding (Wang et al. 2024a). Among recent advancements, MAGVIT-v2 (Yu et al. 2023b) stands out as a highly advanced video tokenizer, achieving high-quality reconstruction by utilizing a large vocabulary size with lookup-free quantization (LFQ). MAGVIT-v2 compresses only 17 frames (approximately 2.125 seconds) of 128×128 resolution video into a $5 \times 16 \times 16$ latent representation, flattened to 1280 tokens. This relatively low compression ratio limits the efficiency of redundant information handling, making the generation of long video sequences computationally expensive when using large language models (LLMs) as the generative framework.

To address these limitations, we draw inspiration from the average scene shot duration in film production and propose a video tokenizer capable of handling extended video sequences. Specifically, our approach encodes 64 frames at 8 FPS which is 8s to provide a broader temporal context for capturing scene-level coherence and continuity. This longer-

duration tokenizer is designed to process input sequences that better represent complete scenes, effectively capturing temporal dependencies and narrative flow. Our design prioritizes maintaining high reconstruction quality while achieving a higher compression ratio, enabling more efficient long-sequence video generation. The detailed compression configurations for the video tokenizer are provided in Tab. 1.

Architecture: Our approach builds on the foundational 3D VAE architecture used in VideoPoet (Yan et al. 2021). This model employs a causal 3D encoder for down-sampling across spatial and temporal dimensions, paired with a symmetrical causal 3D decoder for up-sampling. These components, together with a quantizer (e.g., a look-up mechanism for regularization), enable efficient encoding. Inspired by advancements in 2D VAEs for image encoding (Esser, Rombach, and Ommer 2021; Rombach et al. 2022), 3D VAEs achieve considerably higher compression ratios (Yang et al. 2024), which improves both video quality and temporal consistency in reconstructions. The use of causal 3D CNNs further strengthens our model's ability to generate coherent, visually accurate video sequences (Yu et al. 2023b). The detailed architecture can be found in Fig. 1.

The primary 3D causal encoder and decoder in our model consist of four ResNet blocks, each equipped with causal 3D CNN layers along the temporal dimension. For both compression in the encoder and uncompress in the decoder, spatial processing is applied at each of the four block layers, while temporal processing occurs specifically within the first three block layers. More specifically, starting from an input video resolution of $64 \times 256 \times 256$, the encoder progressively reduces the video data to a latent size of $8 \times 16 \times 16$. Compared to other methods (Yu et al. 2023b; Yang et al. 2024; Wang et al. 2024a), our approach achieves an additional 8x compression when combining both spatial and temporal dimensions. This allows for a substantial reduction in data storage requirements while still retaining essential spatiotemporal information.

Although our tokenizer achieves high-quality reconstruction of long videos at a high compression ratio, it encounters sampling failures in generation tasks based on large language models (LLMs). We attribute this to a substantial semantic gap between vision tokens, which are optimized for visual reconstruction, and text tokens, which capture abstract semantic information. Inspired by VAR (Tian et al. 2024), which demonstrated the effectiveness of a coarse-to-fine approach in image generation using a GPT-style autoregressive model conditioned on class embeddings or text, we adopted a hierarchical strategy similar to VQ-VAE-2 (Razavi, Van den Oord, and Vinyals 2019). Following the main 3D causal encoder, our approach progressively compresses latent features with lightweight causal 3D CNN layers. Notably, as the latent space is further compressed, the quantization dim in LFQ is reduced accordingly. In the decoder phase, a symmetrical network structure is introduced to progressively reconstruct the video content.

Loss Function: Our tokenizer is trained jointly using a combination of L1 loss, LPIPS perceptual loss (Zhang et al. 2018), and adversarial loss (Goodfellow et al. 2014) with a discriminator to enhance reconstruction quality. Following MAGVIT-v2 (Yu et al. 2023b), we also incorporate an entropy penalty to encourage effective vocabulary utilization across each codebook layer. To facilitate more effective training of the multi-layer codebook, we adopt a progressive training strategy. This approach initially focuses on training with higher-compression codebooks to reconstruct a coarse version of the video, gradually adjusting and refining the model on denser codebooks, thereby enhancing the detail in reconstructed outputs.

Text-to-video Generation by Autoregressive LLM

We utilize the Llama-3B model (Touvron et al. 2023; Sun et al. 2024) as the large language model (LLM) to predict the next token, training it to interpret and generate content based on the text prompt. The text prompt is first processed by a frozen Flan-T5-XL model (Chung et al. 2024) to extract text embeddings, which serve as conditioning input for the autoregressive model. The video tokens are then appended after the initial text tokens with right padding applied. Inspired by the animation creation process, where animators start with high-level storyboards and progressively refine details, the sequence of video tokens begins with the most compressed (high-level) tokens and gradually transitions to denser tokens. This ordered sequence of tokens are shown in the bottom of Fig. 1 and can be represented as:

$$p(v|t) = \prod_{m=1}^{M} \prod_{n=1}^{N^m} p(v_n^m \mid t, v_{<n}^{\leq m})$$

where m represents the $m-$th tokenizer layer and N^m denotes the total number of tokens for the $m-$th vocabulary, the next token v_n^m is generated based on previous tokens $v_{<n}^{\leq m}$, as well as all text condition tokens t.

Rotary Position Embedding (RoPE) (Su et al. 2024) effectively captures relative positional information between tokens and excels in modeling long token sequences (Yang et al. 2024), outperforming learnable absolute position embeddings. To represent the positions of video tokens across multi-layer vocabularies in spatial and temporal dimensions, we extend 2D RoPE embeddings to 3D, allocating feature dimensions across three axes. Each layer's codebook applies position encoding independently, while layer-specific positional encodings are incorporated to inform the GPT model about the layer, temporal frame, and spatial location of the token being predicted. Inspired by diffusion models, we include learnable unconditional embeddings to support classifier-free guidance (CFG) generation (Ho and Salimans 2022; Dhariwal and Nichol 2021) to improve diversity.

During training, we use shifted supervision to predict the next token, applying cross-entropy loss independently for the classification logits of multi-layer vocabularies. To ensure balanced learning across hierarchical layers, distinct loss weights are assigned to each layer. During inference, condition text embeddings are used to prefill the first token, and key-value caching (KV-cache) mechanism (Dai et al. 2019; Pope et al. 2023) is employed with positional input to optimize memory usage and reduce computational overhead. The CFG factor is incorporated to balance diversity and fidelity during generation. This framework allows us to

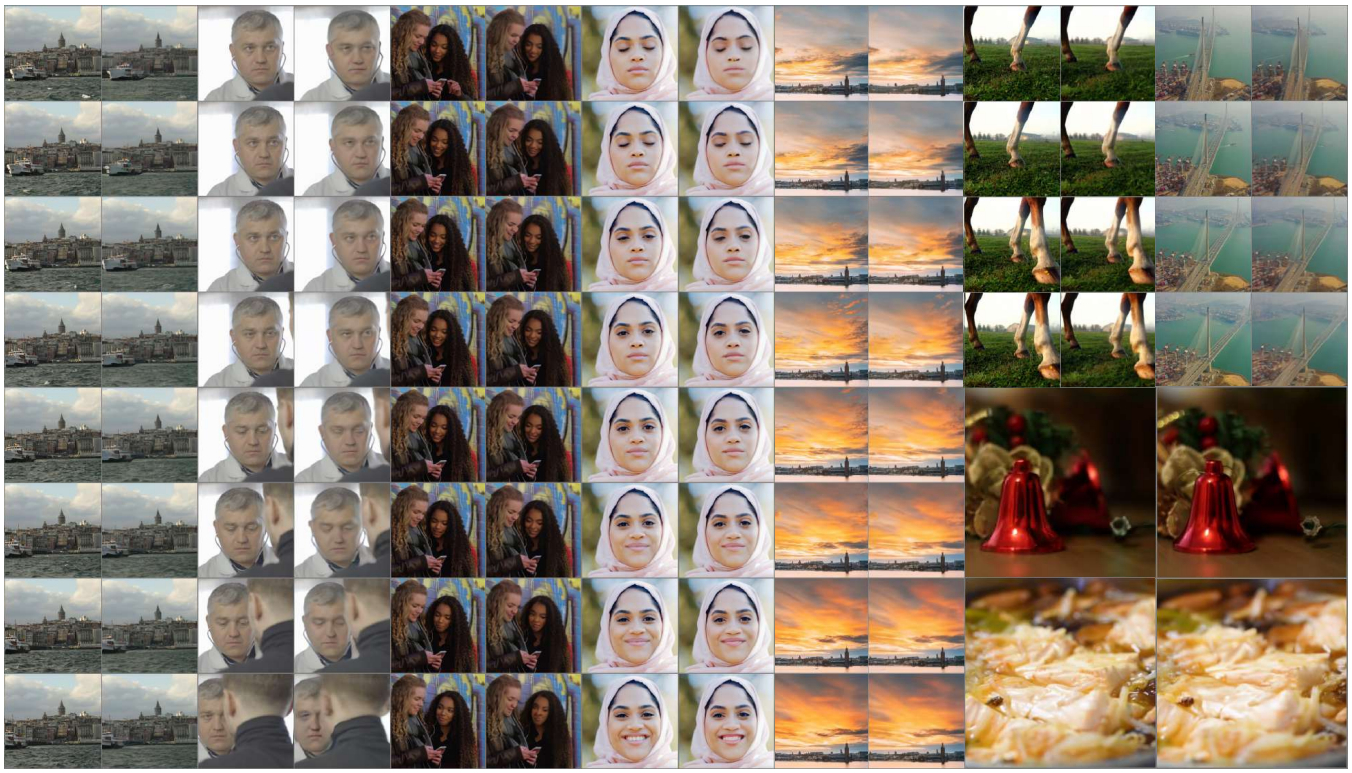


Figure 2: Qualitative evaluation of our proposed hierarchical tokenizers. In each case, the left column represent partial frames sampled from the 64 input video, while the right column presents the corresponding reconstruction results.

Setting	Value	Setting	Value
Parameters	3.1B	Heads	32
Layers	24	Text Length	120
Hidden Size	3200	Residual Dropout	0.1
FFN Dropout	0.1	Token Dropout	0.1

Table 2: Detailed configuration of the autoregressive model.

Method	bpp	LPIPS↓	PSNR↑
HEVC (Bross et al. 2021)	0.03–0.06	0.199	30.10
VVC (Bross et al. 2021)	0.02–0.06	0.153	32.65
MAGVIT (Yu et al. 2023a)	0.038	0.144	23.70
MAGVIT-v2 (Yu et al. 2023b)	0.038	**0.104**	26.18
EMU3 (Wang et al. 2024a)	0.038	0.109	21.59
Ours	**0.012**	0.108	**27.53**

Table 3: Quantitative comparison on Pexels Dataset.

sample video sequences of up to 64 frames efficiently, maintaining an acceptable speed for long sequences. Our primary objective is to demonstrate the benefits of hierarchical tokenizers for simplifying the modeling complexity of video generation with LLM. We do not apply post-training techniques, such as fine-tuning on aesthetic datasets or super-resolution, to further enhance inference results.

Experiments

Video Tokenizers

Implementation details: We train our hierarchical tokenizers as architecture details in Tab. 2 on the Pexels dataset at a resolution of 256×256 and 8 FPS. To assess the effectiveness of our method, we evaluate the bits per pixel (bpp) alongside reconstruction metrics, including LPIPS (Zhang et al. 2018), PSNR (Jähne 2005), SSIM (Wang et al. 2004), and MS-SSIM (Wang, Simoncelli, and Bovik 2003), on the validation set, as summarized in Tab.3. Qualitative comparisons are provided in Fig.2. Compared to prior methods, our proposed video tokenizer achieves state-of-the-art reconstruction quality with a 70% reduction in bits per pixel. Tab. 3 demonstrates its ability to balance compression and reconstruction for long video sequences. Subsequent experiments further show improved multimodal alignment during generation compared to single-layer dense tokenizers at low compression ratios. Furthermore, ablation study on the tokenizer design reveals several key insights and potentials:

Improvements with hierarchical tokenizers: For a fair comparison, we maintain the total token count consistent with the single-layer setup by interpolating the spatial dimensions before downsampling. The detailed compression configurations are provided in Tab. 4. Multi-layer codebooks achieve superior reconstruction quality compared to single-layer codebooks. The quantitative comparison of reconstruction quality on the Pexels dataset is presented inas shown in Fig. 4-(2). Beyond reconstruction, multi-layer codebooks

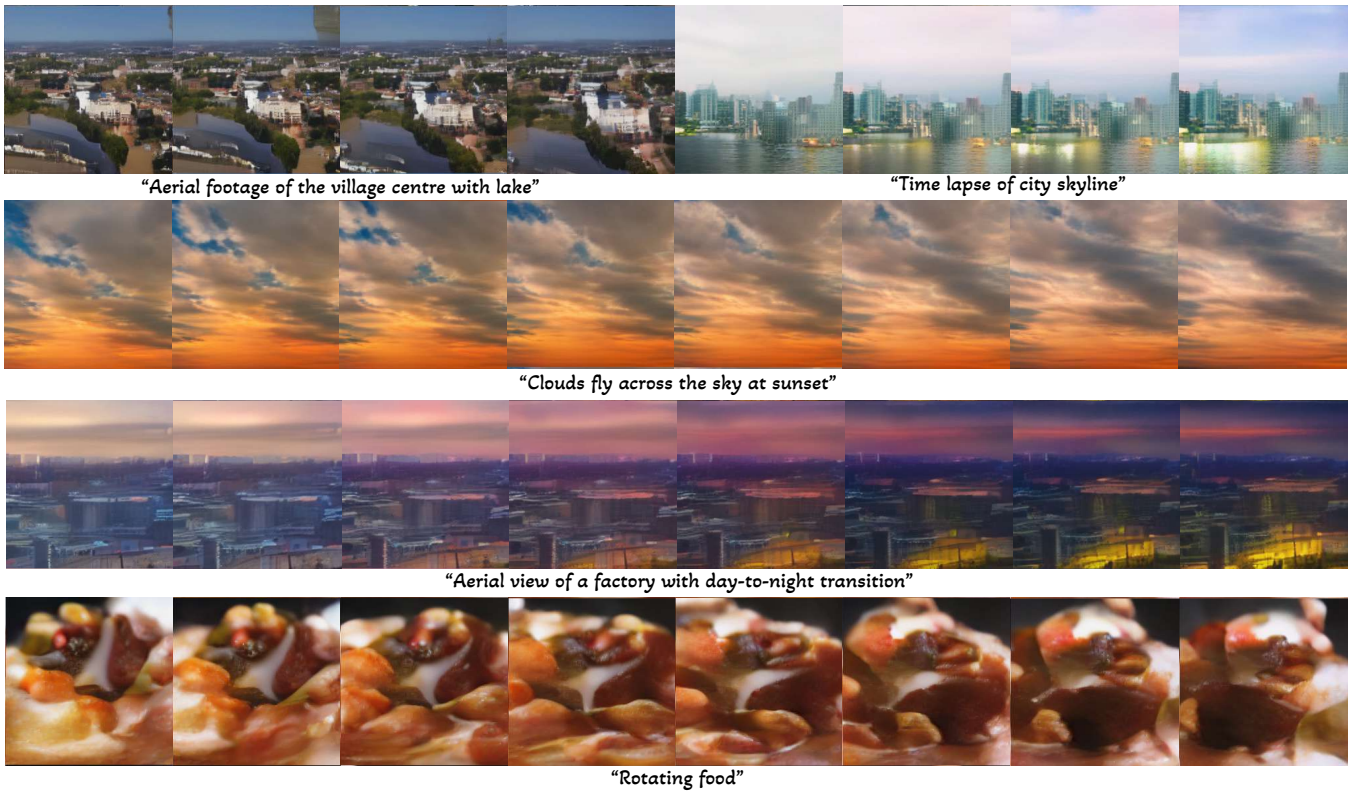


Figure 3: Qualitative generative results of our HiTVideo tokenizers.

Config	Layer Index			
	0	1	2	3
Codebook Size	$2^{18} = 262144$	$2^{16} = 65536$	$2^{14} = 16384$	$2^{12} = 4096$
Compression	$8 \times 8 \times 8$	$16 \times 32 \times 32$	$32 \times 32 \times 32$	$64 \times 64 \times 64$
Latent Size	$8 \times 16 \times 16$	$4 \times 8 \times 8$	$2 \times 8 \times 8$	$1 \times 4 \times 4$
Quant Dim	18	16	14	3
Num of Tokens	2048	256	128	16

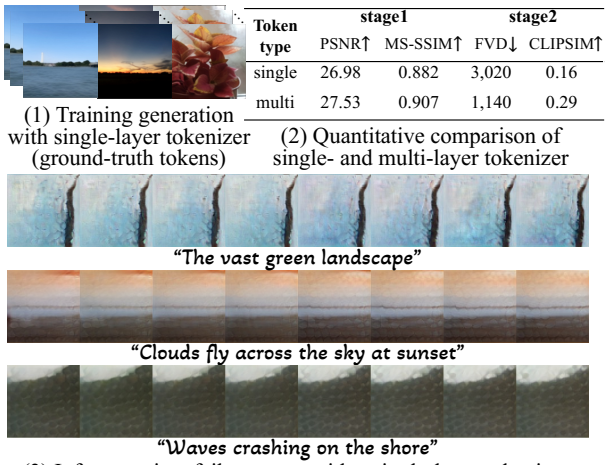
Config	Layer Index
	0
Codebook Size	$2^{18} = 262144$
Compression	$8 \times 8 \times 8$
Latent Size	$8 \times 17 \times 17$
Quant Dim	18
Num of Tokens	2312

Table 4: Configuration details for HiTVideo tokenizer. Left: multi-layer structure. Right: single-layer structure.

show significant advantages in autoregressive generation tasks, as highlighted in Fig. 4. Notably, using a dense single-layer codebook, the LLM fails to generate a coherent 64-frame video from a language prompt as shown in Fig. 4-(3). This underscores the potential of multi-layer structures not only for efficient compression but for bridging the semantic gap between language and video tokens.

Higher Resolution with Higher Compression Ratios Improves Performance: For a fixed total number of video tokens, we observe that utilizing higher input resolutions with larger compression ratios enables the model to capture finer details within the same number of training iterations. This approach ensures that, despite significant compression, the higher-resolution input preserves more spatial information, leading to enhanced reconstruction quality. To ensure a fair comparison, both the 128-resolution and 256-resolution models were initialized with identical parameters and quantization strategies. As illustrated in Fig. 5, the higher-resolution model demonstrates faster convergence in terms of PSNR, indicating superior reconstruction quality. At the same training iteration, videos reconstructed by the higher-resolution model exhibit greater visual fidelity and more detailed features, closely resembling the ground truth, underscoring the advantages of this strategy under high compression settings.

Dynamic video encoding with hierarchical tokenizers: As shown in Tab. 4, dense-layer tokenization generates 2048 tokens per 64-frame video, accounting for approximately 84% of the total token budget and incurring high computational costs. Through VAE pretraining with multi-layer codebooks, we observed that consecutive frames often exhibit high similarity after quantization, indicating redundancy in dense tokenizers due to the inherent properties of video data illustrated in Fig. 6-(1). To quantify this redundancy, we compute the differential matrix across video frames. The pink mask highlights video tokens with differ-



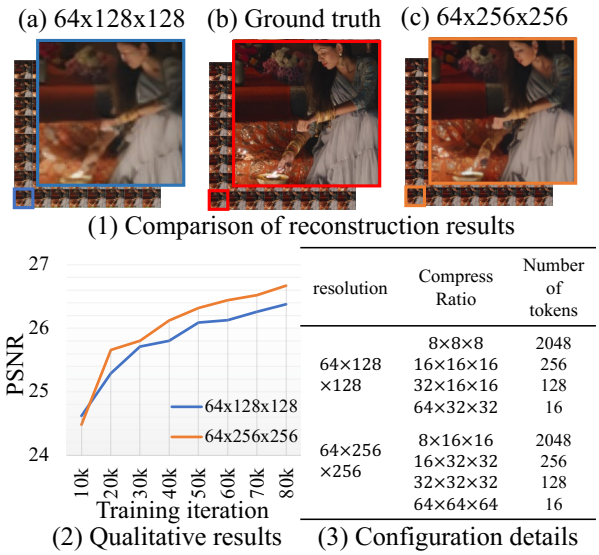
Token type	stage1 PSNR↑	stage1 MS-SSIM↑	stage2 FVD↓	stage2 CLIPSIM↑
single	26.98	0.882	3,020	0.16
multi	27.53	0.907	1,140	0.29

(1) Training generation with single-layer tokenizer (ground-truth tokens)

(2) Quantitative comparison of single- and multi-layer tokenizer

(3) Inference-time failure cases with a single-layer tokenizer

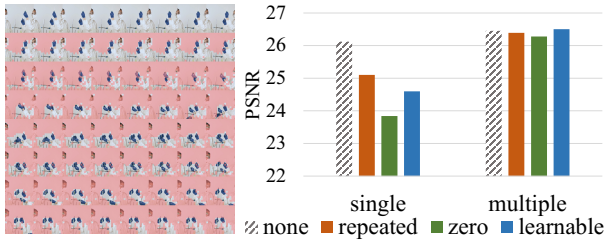
Figure 4: Comparison of using single and multi-layer video tokenizers as $64 \times 256 \times 256$ video resolution.



resolution	Compress Ratio	Number of tokens
64×128 ×128	8×8×8	2048
	16×16×16	256
	32×16×16	128
	64×32×32	16
64×256 ×256	8×16×16	2048
	16×32×32	256
	32×32×32	128
	64×64×64	16

(3) Configuration details

Figure 5: Comparison of qualitative and quantitative results across different input resolutions with same total number.



(1) Redundancy visualization (2) PSNR results w/wo masked decoding

Figure 6: Dynamic video encoding capability with hierarchical tokenizers.

ence scores below the matrix average, identifying redundant patches within the video.

We apply three masking strategies: repeated tokens from the previous frame, zero padding tokens, and learned mask tokens. To balance static and dynamic content, we limit the masking ratio to 85%, ensuring a minimum 15% unmasked patches by random sampling if necessary. As shown in Fig. 6-(2), all masking strategies allow video reconstruction, with scores reported for the same training iterations. Notably, single codebooks with mask significantly degrade reconstruction performance, highlighting the potential of hierarchical tokenizers for dynamic compression and efficient video generation with fewer tokens.

Text-to-video Generation

To demonstrate the effectiveness of our hierarchical tokenizers in generative models, we trained the Llama-3B model to predict the next token autoregressively on Pexels dataset. During inference, the first token is prefilled according to text-conditioned embeddings, and unconditioned embeddings are adjusted using a classifier-free guidance (CFG) scalar. Qualitative results are shown in Fig. 3, while quantitative performance on Pexels prompts is evaluated using FVD (Unterthiner et al. 2018) and CLIPSIM (Radford et al. 2021), as reported in Fig. 4-(2).

For a fair comparison, we trained the autoregressive LLM using a pre-trained single-layer tokenizer on 64 frames. While the model successfully sampled video tokens during training when provided with ground truth causal token indices, as shown in Fig. 4-(1), it failed to generate coherent and meaningful videos when conditioned solely on text prompts, as illustrated in Fig. 4-(2-3). These findings highlight the effectiveness of our highly compressed hierarchical tokenizers in bridging the semantic gap between text and dense visual tokens to force robust text-to-video generation.

Conclusion

Text-to-video generation remains a challenging yet promising frontier in generative modeling. Inspired by traditional animation workflows and prior research, we propose a multi-layer codebook video tokenizer that strikes an effective balance between compression efficiency and reconstruction quality. By capturing high-level semantic content while preserving essential spatiotemporal details, our hierarchical design enhances generation performance under high compression. Compared to single-layer approaches, it improves reconstruction fidelity and facilitates LLM-based modeling by providing abstract, semantically rich representations. Further gains are achieved by introducing dynamic encoding and a masked decoding strategy to reduce redundancy and enhance adaptability.

Limitations. Integrating hierarchical tokenizers with diffusion models remains an open direction for future research.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Bao, F.; Nie, S.; Xue, K.; Li, C.; Pu, S.; Wang, Y.; Yue, G.; Cao, Y.; Su, H.; and Zhu, J. 2023. One transformer fits all distributions in multi-modal diffusion at scale. In *International Conference on Machine Learning*, 1692–1717. PMLR.

Bao, F.; Xiang, C.; Yue, G.; He, G.; Zhu, H.; Zheng, K.; Zhao, M.; Liu, S.; Wang, Y.; and Zhu, J. 2024. Vidu: a highly consistent, dynamic and skilled text-to-video generator with diffusion models. *arXiv preprint arXiv:2405.04233*.

Bar-Tal, O.; Chefer, H.; Tov, O.; Herrmann, C.; Paiss, R.; Zada, S.; Ephrat, A.; Hur, J.; Liu, G.; Raj, A.; et al. 2024. Lumiere: A space-time diffusion model for video generation. *arXiv preprint arXiv:2401.12945*.

Blattmann, A.; Dockhorn, T.; Kulal, S.; Mendelevitch, D.; Kilian, M.; Lorenz, D.; Levi, Y.; English, Z.; Voleti, V.; Letts, A.; et al. 2023. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*.

Brooks, T.; Peebles, B.; Holmes, C.; DePue, W.; Guo, Y.; Jing, L.; Schnurr, D.; Taylor, J.; Luhman, T.; Luhman, E.; et al. 2024. Video generation models as world simulators.

Bross, B.; Wang, Y.-K.; Ye, Y.; Liu, S.; Chen, J.; Sullivan, G. J.; and Ohm, J.-R. 2021. Overview of the versatile video coding (VVC) standard and its applications. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(10): 3736–3764.

Brown, T. B. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Chang, H.; Zhang, H.; Jiang, L.; Liu, C.; and Freeman, W. T. 2022. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11315–11325.

Chen, H.; Zhang, Y.; Cun, X.; Xia, M.; Wang, X.; Weng, C.; and Shan, Y. 2024. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7310–7320.

Chen, M.; Radford, A.; Child, R.; Wu, J.; Jun, H.; Luan, D.; and Sutskever, I. 2020. Generative pretraining from pixels. In *International Conference on Machine Learning*, 1691–1703. PMLR.

Chung, H. W.; Hou, L.; Longpre, S.; Zoph, B.; Tay, Y.; Fedus, W.; Li, Y.; Wang, X.; Dehghani, M.; Brahma, S.; et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70): 1–53.

Dai, Z.; Yang, Z.; Yang, Y.; Carbonell, J.; Le, Q. V.; and Salakhutdinov, R. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*.

Devlin, J. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34: 8780–8794.

Esser, P.; Rombach, R.; and Ommer, B. 2021. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12873–12883.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27.

He, Y.; Yang, T.; Zhang, Y.; Shan, Y.; and Chen, Q. 2022. Latent video diffusion models for high-fidelity long video generation. *arXiv preprint arXiv:2211.13221*.

Ho, J.; Chan, W.; Saharia, C.; Whang, J.; Gao, R.; Gritsenko, A.; Kingma, D. P.; Poole, B.; Norouzi, M.; Fleet, D. J.; et al. 2022a. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33: 6840–6851.

Ho, J.; and Salimans, T. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Ho, J.; Salimans, T.; Gritsenko, A.; Chan, W.; Norouzi, M.; and Fleet, D. J. 2022b. Video diffusion models. *Advances in Neural Information Processing Systems*, 35: 8633–8646.

Hong, W.; Ding, M.; Zheng, W.; Liu, X.; and Tang, J. 2022. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*.

Jähne, B. 2005. *Digital image processing*. Springer Science & Business Media.

Karras, T. 2017. Progressive Growing of GANs for Improved Quality, Stability, and Variation. *arXiv preprint arXiv:1710.10196*.

Kondratyuk, D.; Yu, L.; Gu, X.; Lezama, J.; Huang, J.; Schindler, G.; Hornung, R.; Birodkar, V.; Yan, J.; Chiu, M.-C.; et al. 2023. Videopoet: A large language model for zero-shot video generation. *arXiv preprint arXiv:2312.14125*.

Nichol, A. Q.; and Dhariwal, P. 2021. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, 8162–8171. PMLR.

Peebles, W.; and Xie, S. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4195–4205.

Pope, R.; Douglas, S.; Chowdhery, A.; Devlin, J.; Bradbury, J.; Heek, J.; Xiao, K.; Agrawal, S.; and Dean, J. 2023. Efficiently scaling transformer inference. *Proceedings of Machine Learning and Systems*, 5: 606–624.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text

transformer. *Journal of machine learning research*, 21(140): 1–67.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, 8821–8831. Pmlr.

Razavi, A.; Van den Oord, A.; and Vinyals, O. 2019. Generating diverse high-fidelity images with vq-vae-2. *Advances in Neural Information Processing Systems*, 32.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, 234–241. Springer.

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. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35: 36479–36494.

Sanh, V.; Webson, A.; Raffel, C.; Bach, S. H.; Sutawika, L.; Alyafeai, Z.; Chaffin, A.; Stiegler, A.; Scao, T. L.; Raja, A.; et al. 2021. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*.

Singer, U.; Polyak, A.; Hayes, T.; Yin, X.; An, J.; Zhang, S.; Hu, Q.; Yang, H.; Ashual, O.; Gafni, O.; et al. 2022. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*.

Skorokhodov, I.; Tulyakov, S.; and Elhoseiny, M. 2022. Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3626–3636.

Su, J.; Ahmed, M.; Lu, Y.; Pan, S.; Bo, W.; and Liu, Y. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568: 127063.

Sun, P.; Jiang, Y.; Chen, S.; Zhang, S.; Peng, B.; Luo, P.; and Yuan, Z. 2024. Autoregressive Model Beats Diffusion: Llama for Scalable Image Generation. *arXiv preprint arXiv:2406.06525*.

Tian, K.; Jiang, Y.; Yuan, Z.; Peng, B.; and Wang, L. 2024. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *arXiv preprint arXiv:2404.02905*.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Unterthiner, T.; Van Steenkiste, S.; Kurach, K.; Marinier, R.; Michalski, M.; and Gelly, S. 2018. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*.

Van Den Oord, A.; Vinyals, O.; et al. 2017. Neural discrete representation learning. *Advances in Neural Information Processing Systems*, 30.

Wan, T.; Wang, A.; Ai, B.; Wen, B.; Mao, C.; Xie, C.-W.; Chen, D.; Yu, F.; Zhao, H.; Yang, J.; et al. 2025. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*.

Wang, X.; Zhang, X.; Luo, Z.; Sun, Q.; Cui, Y.; Wang, J.; Zhang, F.; Wang, Y.; Li, Z.; Yu, Q.; et al. 2024a. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*.

Wang, Y.; Bao, J.; Weng, W.; Feng, R.; Yin, D.; Yang, T.; Zhang, J.; Dai, Q.; Zhao, Z.; Wang, C.; et al. 2024b. Microcinema: A divide-and-conquer approach for text-to-video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8414–8424.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612.

Wang, Z.; Simoncelli, E. P.; and Bovik, A. C. 2003. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, 1398–1402. Ieee.

Wu, C.; Chen, X.; Wu, Z.; Ma, Y.; Liu, X.; Pan, Z.; Liu, W.; Xie, Z.; Yu, X.; Ruan, C.; et al. 2024. Janus: Decoupling visual encoding for unified multimodal understanding and generation. *arXiv preprint arXiv:2410.13848*.

Xie, J.; Mao, W.; Bai, Z.; Zhang, D. J.; Wang, W.; Lin, K. Q.; Gu, Y.; Chen, Z.; Yang, Z.; and Shou, M. Z. 2024. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*.

Yan, W.; Zhang, Y.; Abbeel, P.; and Srinivas, A. 2021. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*.

Yang, Z.; Teng, J.; Zheng, W.; Ding, M.; Huang, S.; Xu, J.; Yang, Y.; Hong, W.; Zhang, X.; Feng, G.; et al. 2024. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*.

Yu, L.; Cheng, Y.; Sohn, K.; Lezama, J.; Zhang, H.; Chang, H.; Hauptmann, A. G.; Yang, M.-H.; Hao, Y.; Essa, I.; et al. 2023a. Magvit: Masked generative video transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10459–10469.

Yu, L.; Lezama, J.; Gundavarapu, N. B.; Versari, L.; Sohn, K.; Minnen, D.; Cheng, Y.; Birodkar, V.; Gupta, A.; Gu, X.; et al. 2023b. Language Model Beats Diffusion–Tokenizer is Key to Visual Generation. *arXiv preprint arXiv:2310.05737*.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 586–595.