VTOK: A UNIFIED VIDEO TOKENIZER WITH DECOUPLED SPATIAL-TEMPORAL LATENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

This work presents VTok, a unified video tokenization framework that can be used for both generation and understanding tasks. Unlike the leading vision–language systems that tokenize videos through a naïve frame-sampling strategy, we propose to decouple the spatial and temporal representations of videos by retaining the spatial features of a single key frame while encoding each subsequent frame into a single residual token, achieving compact yet expressive video tokenization. Our experiments suggest that VTok effectively reduces the complexity of video representation from the product of frame count and per-frame token count to their sum, while the residual tokens sufficiently capture viewpoint and motion changes relative to the key frame. Extensive evaluations demonstrate the efficacy and efficiency of VTok: it achieves notably higher performance on a range of video understanding and text-to-video generation benchmarks compared with baselines using naïve tokenization, all with shorter token sequences per video (e.g., 3.4% higher accuracy on our TV-Align benchmark and 1.9% higher VBench score). Remarkably, VTok produces more coherent motion and stronger guidance following in text-to-video generation, owing to its more consistent temporal encoding. We hope VTok can serve as a standardized video tokenization paradigm for future research in video understanding and generation.

1 INTRODUCTION

A central bottleneck in scaling multimodal models to video is the interface between pixels and language—how to compress long, high-resolution streams into tokens that an LLM can reason over. The dominant practice in recent video–language systems (*e.g.*, Qwen2.5-VL (Bai et al., 2025), LLaVANext-Video (Zhang et al., 2024), BLIP-3-Video (Ryoo et al., 2024)) is to first sample frames, encode each frame with a 2D image encoder, and then concatenate all visual tokens across frames into one long sequence that is fed to the language model. While straightforward, this pipeline entangles spatial and temporal information in a single monolithic token stream, making it difficult to allocate capacity to what actually changes over time. This design inherits strong spatial features but creates a hard trade-off: either sample very sparsely to keep the token budget manageable—thereby erasing micro-motions and transient events—or retain more frames and suffer a rapidly growing sequence length, which drives up quadratic attention cost and makes long-range temporal credit assignment brittle. As a consequence, for understanding, models often struggle with fine-grained temporal reasoning and event ordering (Bai et al., 2025; Ryoo et al., 2024; Zhang et al., 2025; 2024; Wang et al., 2022; 2024d; Bardes et al., 2024); for generation, textual guidance is often weakly grounded in motion, leading to timing inconsistencies and identity drift despite acceptable performance on coarse recognition (Yang et al., 2025b; Agarwal et al., 2025; Kong et al., 2024; Wan et al., 2025; Zheng et al., 2024; Peng et al., 2025; Team, 2024).

In fact, given the strong prior of inter-frame continuity and coherence, frames within a scene of moderate duration tend to maintain very similar spatial structure while changing smoothly in viewpoint, object pose, and location. This implies that frame-sampling–based video tokenization introduces substantial redundancy on the spatial side while excessively discarding subtle temporal variations. This imbalance motivates us to develop an efficient video tokenization method that streamlines the spatial structure of a video while preserving as much motion information as possible. Through a series of investigations across both video understanding and generation tasks, we surprisingly find that video tokenization can be explicitly decoupled into spatial and temporal components. Concretely,

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

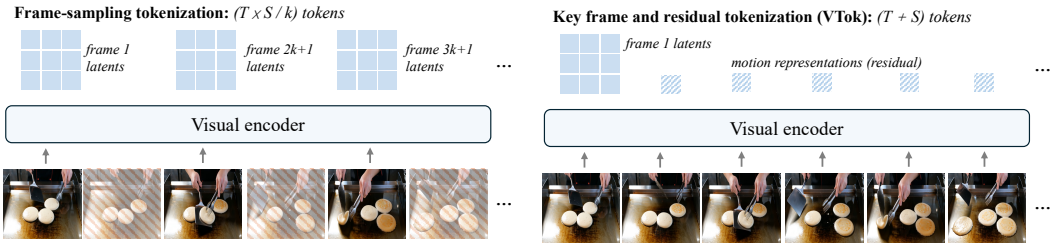

Figure 1: ***Video tokenization strategies.*** Leading VLMs such as Qwen2.5-VL (Bai et al., 2025) still rely on a naïve frame-sampling approach to process videos. We contend that this strategy introduces excessive redundancy in spatial information while omitting temporal details. We find that video tokenization can be decomposed into spatial and temporal components, which preserves spatiotemporal information while minimizing the overall token length.

as shown in Figure 1, a video clip suffices to retain the full spatial information of the first frame (often the key frame), while encoding each subsequent frame as a single token that captures its temporal change relative to the key frame. Under this paradigm, the token budget for a clip reduces from $\mathcal{O}(T \times S)$ to $\mathcal{O}(T + S)$—with S spatial tokens from the key frame and T frames—substantially improving efficiency while preserving action and motion details.

We validate this new tokenization paradigm by integrating it into a unified framework that jointly learns from video-to-text representation and text-to-video generation tasks. As shown in Figure 2, our model is built on a unified autoregressive Multimodal Large Language Model (MLLM) that shares the same tokenization and reasoning backbone across both understanding and generation. For the understanding branch, a video is encoded into a sequence of key-frame and residual tokens, which are projected into the same embedding space as language tokens. These visual tokens are then concatenated with an understanding-oriented textual prompt and processed by the MLLM to produce textual outputs such as captions, answers, or explanations. In the generation branch, the same MLLM takes a textual prompt as input and autoregressively samples visual tokens that conform to our key-frame-plus-residual format. The resulting token sequence is subsequently decoded by a pretrained diffusion transformer into the final video, ensuring both spatial fidelity and temporally consistent motion generation.

We evaluate our model by testifying its performance on video understanding benchmarks such as Video-MMMU (Hu et al., 2025), its generation quality under the VBench (Huang et al., 2024b) standard, as well as our newly proposed Text-to-Video Guidance Alignment Benchmark (TV-Align), which highlights whether generated videos faithfully follow high-level semantic instructions in prompts—such as object count, direction, and relative position. Results show that our model achieves notably stronger generation performance, outperforming the base model by 1.92 and 4.33 points on the total and semantic metrics of VBench (Huang et al., 2024b), respectively. This demonstrates that our tokenization paradigm provides more precise spatiotemporal guidance, leading to scenes and motions that are both semantically grounded and visually coherent. On the TV-Align benchmark, our model significantly surpasses leading open-source systems such as HunyuanVideo (Kong et al., 2024) and Wan2.2 (Wan et al., 2025), validating the superiority of our text-to-video generation framework—in which the MLLM itself generates visual tokens autoregressively, rather than merely conditioning a diffusion transformer on textual embeddings. This design enhances cross-modal semantic encoding, thereby achieving stronger text–video alignment. On video understanding benchmarks, our model likewise demonstrates consistent and significant gains. A fine-tuned LLaVA-Next model with a LLaMA-3 (Grattafiori et al., 2024) text encoder achieves an average improvement of 2.3% over the original base model across seven standard benchmarks, confirming that our unified training and tokenization strategy effectively benefits both understanding and generation tasks. We hope that our new video tokenizer and unified generative–understanding framework can provide valuable insights and foundations for future research on efficient and semantically aligned video–language modeling.

2 Related Work

Image and video tokenization A classical approach to visual tokenization is to compress visual information into a compact continuous (Rombach et al., 2022; Peebles & Xie, 2023; Ma et al., 2024;

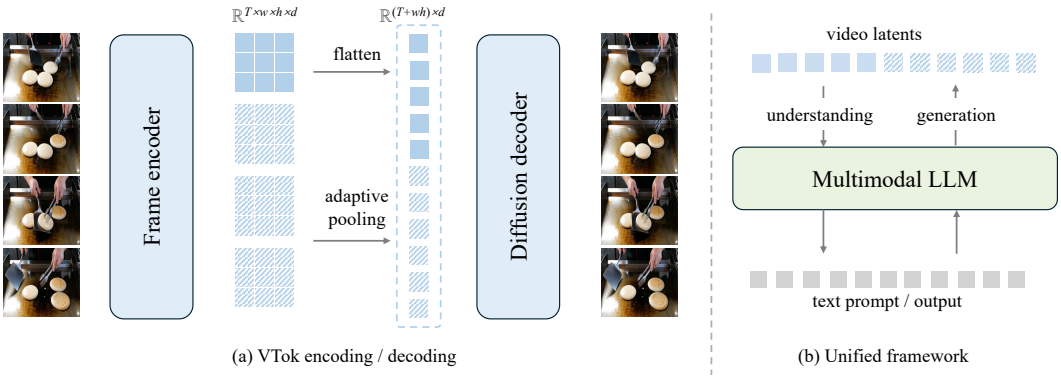

Figure 2: ***Our unified framework of video understanding and text-to-video generation.*** The model integrates video understanding and generation within a single autoregressive multimodal large language model (MLLM). In the understanding branch, videos are first tokenized into key-frame and residual tokens, aligned with the language space, and processed together with a textual prompt to produce semantic outputs such as captions or answers. In the generation branch, the MLLM samples visual tokens conditioned on text, following the same spatial–temporal format, and a diffusion transformer can decode these tokens back to a video.

Chen et al., 2025; Bao et al., 2023; Gao et al., 2023; Yu et al., 2025b) or discrete (Esser et al., 2021; Chang et al., 2022; Sun et al., 2024; Li et al., 2023; Yu et al., 2024; Jiao et al., 2025) latent space through a variational autoencoder (VAE) framework, where an encoder–decoder structure jointly performs information compression and reconstruction. This paradigm was initially developed for 2D images and was later extended to videos, showing that temporal data can be tokenized and reconstructed in a similar manner (Kong et al., 2024; Wan et al., 2025; Agarwal et al., 2025; Yang et al., 2025b; Zheng et al., 2024; Peng et al., 2025; Xi et al., 2025; Yan et al., 2024; Ryoo et al., 2021; Blattmann et al., 2023; Kondratyuk et al., 2024; Jang et al., 2025; Tang et al., 2024; Wang et al., 2024b;c;e; Luo et al., 2024; Lu et al., 2025; Xiong et al., 2025; Polyak et al., 2024; Yin et al., 2025; Huang et al., 2024a; Guo et al., 2025; Yuan et al., 2024; Ma et al., 2025; HaCohen et al., 2024; Guo et al., 2024; Chen et al., 2024; Fan et al., 2025). By pairing such video tokenizers with large-scale autoregressive language models or diffusion models, recent works have achieved impressive progress in both image and video generation and understanding (Liu et al., 2025; Wu et al., 2024b; Wang et al., 2024c; Zhang et al., 2024; Li et al., 2024a; Xue et al., 2025; Ryoo et al., 2024; Bai et al., 2025; Hansen-Estruch et al., 2025; Tschannen et al., 2025; Zhao et al., 2024; Yu et al., 2025a; Fang et al., 2024). An alternative line of work leverages pretrained visual encoders to convert visual inputs into latent representations, which are then used as conditioning signals for external models such as multimodal large language models (MLLMs) or diffusion transformers to de-tokenize the signal back to the visual domain (Yang et al., 2025a; Wang et al., 2025a; Ge et al., 2025; Zhou et al., 2025). Unlike VAE-style tokenization, this strategy does not emphasize precise pixel-level reconstruction; instead, it focuses on maintaining high-level semantic alignment between visual and textual modalities, enabling efficient multimodal reasoning and instruction following. This work mainly focuses on the later. Among these video tokenization studies, several recent works are most aligned with our perspective. For example, OmniTokenizer (Wang et al., 2024b) introduces a spatial–temporal dual-branch design, where a dedicated temporal branch provides additional motion information to complement key-frame encoding. Similarly, Video-LaVIT (Jin et al., 2024) incorporates a specialized motion encoder to obtain dual-stream video latents, motivated by the intuition that explicit motion cues can significantly enhance video understanding (Wang et al., 2023; Shen et al., 2024).

Foundational video models The rise of video foundation models marks a shift toward large-scale, general-purpose video understanding systems (Bai et al., 2025; Ryoo et al., 2024; Tschannen et al., 2025; Li et al., 2024a; Zhang et al., 2025; 2024; Wang et al., 2024c). These models aim to learn unified representations that capture both spatial semantics and temporal dynamics from massive video–text datasets. Early frameworks extended image–language pretraining (Radford et al., 2021; Wang et al., 2024a; Zhai et al., 2023; Alayrac et al., 2022; Yu et al., 2022; Jia et al., 2021; Pham et al., 2021; Yuan et al., 2021; Li et al., 2022) to videos by encoding sampled frames and aligning them with textual captions, enabling strong zero-shot reasoning across diverse tasks. Recent advances scale this idea further, training billions of parameters on diverse web videos to achieve robust cross-domain

transfer in captioning, retrieval, question answering, and generation (Kong et al., 2024; Luo et al., 2024; Wan et al., 2025; Zheng et al., 2024; Peng et al., 2025; Agarwal et al., 2025; Wang et al., 2024c). Despite impressive progress, most foundation models still rely on frame-based tokenization, which limits their efficiency and temporal fidelity, highlighting the need for more structured video representations like ours.

3 METHOD

3.1 UNIFIED FRAMEWORK

We build a unified model that jointly handles video understanding and video generation within a single multimodal framework. The architecture consists of two complementary branches: an understanding branch that maps videos to textual outputs, and a generation branch that synthesizes videos conditioned on textual prompts, both sharing the same autoregressive backbone and tokenization scheme. Formally, let a video be $X \in \mathbb{R}^{T\times H\times W\times 3}$ and a text prompt be p. Our framework couples a video encoder $\mathcal{T}_{\text{vid}}$ with a single autoregressive MLLM and a video decoder $\mathcal{D}_{\text{vid}}$. The same MLLM is used for both video-to-text understanding and text-to-video generation. We then denote the video tokenization pipeline as

$$\mathcal{T}_{\text{vid}} : \mathbb{R}^{T\times H\times W\times 3} \to \mathbb{R}^{L_v\times d_v}, \quad V = \mathcal{T}_{\text{vid}}(X), \tag{1}$$

which yields a sequence of L_v visual tokens $V = \{v_1, \ldots, v_{L_v}\}$. Text is tokenized by a standard tokenizer:

$$\mathcal{T}_{\text{text}} : \mathcal{V}_{\text{str}} \to \{0, \ldots, |\Sigma| - 1\}^{L_t}, \quad Y = \mathcal{T}_{\text{text}}(p), \tag{2}$$

with embeddings $\phi_{\text{text}} : \mathbb{N} \to \mathbb{R}^d$ and a visual projection $\phi_{\text{vis}} : \mathbb{R}^{d_v} \to \mathbb{R}^d$. The MLLM input is the concatenation

$$Z = \big[\, \phi_{\text{text}}(Y_{\text{in}}) \,;\, \phi_{\text{vis}}(V) \,\big] + P, \tag{3}$$

where P denotes positional or modality embeddings and the concatenation order depends on the branch. The MLLM defines a conditional distribution over token sequences:

$$f_\theta(Z) = \text{Transformer}_\theta(Z), \tag{4}$$

$$p_\theta(s_i \mid s_{<i}) = \text{softmax}\big(W\, f_\theta(Z)_i\big), \tag{5}$$

where s_i ranges over a unified vocabulary that includes both text and visual tokens after projection into the shared space.

In the understanding branch, given video X and instruction prompt p, the input is arranged as

$$Z_{\text{under}} = \big[\, \phi_{\text{text}}(\mathcal{T}_{\text{text}}(p)) \,;\, \phi_{\text{vis}}(\mathcal{T}_{\text{vid}}(X)) \,\big] + P. \tag{6}$$

The model autoregressively predicts the target text sequence $Y^\star$ (caption, answer, rationale) using the standard language modeling loss:

$$\mathcal{L}_{\text{under}} = -\sum_{i=1}^{|Y^\star|} \log p_\theta(Y_i^\star \mid Y_{<i}^\star,\, Z_{\text{under}})\,. \tag{7}$$

In the generation branch, given prompt p, the model samples visual tokens $\hat{V}$ conditioned on text:

$$Z_{\text{gen}} = \big[\, \phi_{\text{text}}(\mathcal{T}_{\text{text}}(p)) \,\big] + P, \qquad \hat{V} \sim p_\theta(\cdot \mid Z_{\text{gen}}). \tag{8}$$

A pretrained video decoder $\mathcal{D}_{\text{vid}}$, a diffusion transformer learned in the latent space reconstructs the video:

$$\hat{X} = \mathcal{D}_{\text{vid}}(\hat{V}). \tag{9}$$

Training uses a visual-token language modeling loss plus the decoder's diffusion objective:

$$\mathcal{L}_{\text{visLM}} = -\sum_{j=1}^{L_v} \log p_\theta(V_j \mid V_{<j},\, Z_{\text{gen}})\,, \tag{10}$$

$$\mathcal{L}_{\text{dec}} = \mathbb{E}_{t,\epsilon}\big[\|\epsilon - \epsilon_\psi(V_t, t)\|\big] \tag{11}$$

The overall training objective combines both branches:

$$\mathcal{L} = \mathcal{L}_{\text{under}} + \lambda_{\text{visLM}}\mathcal{L}_{\text{visLM}} + \lambda_{\text{dec}}\mathcal{L}_{\text{dec}}. \tag{12}$$

At inference, the same MLLM is used for: (i) *understanding*, by conditioning on (p, X) and decoding text; and (ii) *generation*, by conditioning on p, sampling visual tokens, and decoding them to video via $\mathcal{D}_{\text{vid}}$.

Model	#Param	TV-Alignment Benchmark							
		Count	Direction	Rel. pos.	Rel. size	Color	State	Motion	Avg.
OpenSora-1.3	1B	27.8	41.0	45.1	33.5	42.3	25.6	38.9	36.3
OpenSora-2.0	11B	28.9	42.7	47.0	34.9	43.1	26.8	40.1	37.6
Mochi	10B	29.4	43.2	48.0	35.2	43.6	27.1	40.5	38.1
CogVideoX1.5	5B	29.8	43.9	49.3	36.0	44.0	27.5	41.0	38.8
WAN2.1	14B	32.4	45.9	51.4	38.1	45.9	29.4	42.9	40.9
WAN2.2	14B	32.6	46.1	51.8	38.3	46.0	29.5	43.2	41.1
Our direct baseline:									
HunyuanVideo	13B	31.0	46.0	51.7	37.3	45.9	28.5	43.4	40.5
OmniTokenizer*	13B	31.8	46.8	52.5	38.6	46.4	29.4	44.0	41.4
Video-LaViT*	13B	32.4	47.3	53.1	39.8	46.9	30.2	44.6	42.1
VTok (ours)	13B	**33.7**	**48.3**	**55.0**	**42.9**	**47.7**	**32.5**	**47.1**	**43.9**

Table 1: ***TV-Align accuracy (%) of text-to-video generation.*** We compare with state-of-the-art open-source video generation baselines, where VTok demonstrates better guidance following of high-level semantics. * denotes using the same base model and training framework but replacing VTok by other tokenization strategy. Our model is highlighted in orange.

3.2 Spatial-Temporal Decoupled Tokenization

Given the unified framework defined above, our goal is to design a video tokenizer $\mathcal{T}_{\text{vid}}$ that produces compact yet expressive representations by explicitly separating spatial and temporal information. We denote a single video frame $x_t \in \mathbb{R}^{H\times W\times 3}$, where there are T frames in total from $X = \{x_t\}_{t=1}^{T}$. For best simplicity and generalizability, we select the first frame[1] x_1 as the key frame and encode it into a set of spatial tokens:

$$V^{(s)} = \mathcal{E}_{\text{key}}(x_1) \in \mathbb{R}^{S\times d_v}, \tag{13}$$

where $\mathcal{E}_{\text{key}}$ is a vision encoder and S denotes the number of spatial tokens. For each subsequent frame x_t $(t > 1)$, we compute a residual motion token that captures its temporal difference relative to the key frame:

$$v_t^{(m)} = \mathcal{E}_{\text{motion}}(x_t, x_1) = g_\phi(\mathcal{F}(x_t) - \mathcal{F}(x_1)), \tag{14}$$

where $\mathcal{F}$ is a shared feature extractor and g_ϕ projects the residual feature difference into the motion-token space $\mathbb{R}^{d_v}$. We use the same model and parameters for $\mathcal{E}_{\text{key}}$ and $\mathcal{F}$. The full video token sequence is then constructed as

$$V = \{\, V^{(s)};\ v_2^{(m)}, \ldots, v_T^{(m)} \,\}, \tag{15}$$

resulting in a total token length of $L_v = S + (T-1)$.

During generation, the MLLM autoregressively predicts a corresponding token sequence $\hat{V}$ with the same spatial–temporal structure:

$$\hat{V} = \{\, \hat{V}^{(s)};\ \hat{v}_2^{(m)}, \ldots, \hat{v}_T^{(m)} \,\}. \tag{16}$$

A pretrained video decoder $\mathcal{D}_{\text{vid}}$ reconstructs the video as

$$\hat{X} = \mathcal{D}_{\text{vid}}(\hat{V}) = \mathcal{D}_{\text{vid}}\big(\hat{V}^{(s)}, \hat{v}_{2:T}^{(m)}\big), \tag{17}$$

where $\hat{V}^{(s)}$ provides the spatial structure and $\hat{v}_{2:T}^{(m)}$ modulate motion dynamics. This spatial–temporal decoupled tokenization effectively captures key frame appearance and inter-frame motion while maintaining a compact token budget, enabling efficient end-to-end optimization within our unified video generation–understanding framework.

4 Experiments

4.1 Experimental Setup

We employ a pretrained LLaVA-Next (Zhang et al., 2024) equipped with a LLaMA3-8B (Grattafiori et al., 2024) language encoder as the core MLLM of our framework. Since the latent space of

[1]Employing a dedicated key frame detector can slightly improve performance. However, it compromises the overall compactness of the model, so we simply use the first frame as the key frame in all experiments.

this model has already been aligned with the video decoder in HunyuanVideo (Kong et al., 2024), no additional adapter is required to bridge the latent spaces between the MLLM and the video decoder. During training, we optimize only the MLLM parameters, while keeping both the visual encoder (CLIP-L/336px (Radford et al., 2021)) and the video decoder (HunyuanVideo-13B diffusion transformer (Kong et al., 2024)) frozen. We use the AdamW optimizer with a learning rate of 1e-5, a total batch size of 16, and an exponential moving average (EMA) decay of 0.999. The model is trained on approximately 5 million densely annotated video–caption pairs for both understanding and generation. The loss weights λ_{visLM} and λ_{dec} are simply set to 1. At the current stage, we have not performed extensive hyperparameter tuning, as the present configuration already shows a clear performance improvement over baseline models.

To evaluate the generalizability of our tokenizer, we further train a lightweight adapter (a two-layer MLP) to connect it with the state-of-the-art open-source video generation model WAN2.2 (Wan et al., 2025). Specifically, in this experiment, we project the video representations produced by VTok into the same latent space as WAN's text encoder, and then concatenate them with the original textual prompt features as the conditioning input to the DiT decoder. As shown in Figure 3, the quantitative results demonstrate that VTok transfers effectively to the WAN model, indicating its strong adaptability across different generative architectures.

4.2 Main Experiments

Text-to-video alignment. To explicitly evaluate whether the generated videos can well follow the instruction in their prompts, we introduce a new text-to-video generation benchmark called TV-Align. Unlike existing benchmarks such as VBench (Huang et al., 2024b), which primarily focus on assessing the visual quality, realism, or coherence of generated videos, TV-Align emphasizes a model's ability to understand and follow high-level semantic instructions embedded in textual prompts. The benchmark consists of approximately 1,000 (prompt, question, answer) triplets, covering seven key semantic categories in video generation: counting, direction, relative position, relative size, color alignment, state, and motion. Each triplet is carefully designed to isolate one semantic concept. For example, a counting case may use the prompt "A person places three oranges on the table", where the number of objects ("three") is the ground-truth semantic clue. The corresponding question would be "How many oranges are placed on the table?" with multiple-choice answers, enabling straightforward accuracy measurement. We show more specific examples of the TV-Align benchmark in Figure 3.

Given a trained text-to-video (T2V) model, we feed the prompt into it and pass the generated video to a video understanding model (we use Qwen2.5-VL in our experiments) to answer the question. The resulting accuracy directly reflects whether the T2V model correctly captured and executed the intended instruction during generation. To minimize ambiguity, both prompts and questions were manually curated, and we observed that automated understanding results are highly consistent with human judgments (within 1% accuracy difference). Notably, even the latest state-of-the-art open-source models—such as WAN2.2—achieve less than 50% accuracy on TV-Align, revealing significant limitations in their comprehension of high-level semantics. Throughout our experiments, we use TV-Align as a primary evaluation metric to objectively assess how well our unified MLLM and diffusion-based decoders understand and follow textual instructions in video generation.

Table 1 presents a quantitative comparison on TV-Align. As shown, our model VTok achieves the best performance across all seven categories, surpassing recent large-scale video generation systems including WAN2.2 (Wan et al., 2025), HunyuanVideo (Kong et al., 2024), and CogVideoX1.5 (Yang et al., 2025b). In particular, VTok shows notable gains in relative size, motion, and counting, indicating stronger grounding of quantitative and dynamic semantics. Compared to the base model HunyuanVideo with the same video decoder, VTok improves the overall alignment accuracy from 40.5% to 43.9%, demonstrating that the proposed tokenization and unified training framework effectively enhance instruction following and temporal coherence. We also observe that when replacing VTok by the existing video tokenizers such as OmniTokenizer and Video-LaViT, which decompose spatial-temporal information but still rely on frame-sampling, the alignment performance falls significant short to ours. These results highlight VTok's superior ability to translate textual intent into visually and semantically consistent video content.

Video quality evaluation. Table 2 provides a detailed comparison of model performance on VBench (baseline results are taken from (Wan et al., 2025) and (Yin et al., 2025)), where we assess overall

Model	Total	Quality	Semantic	Subject	Backg.	Smooth	Dynamic	Object	Multi.
Closed-source models:									
Sora	84.28	85.51	79.35	96.23	96.35	98.74	79.91	93.93	70.85
ARLON	82.31	83.58	77.27	95.59	98.65	97.56	72.22	90.60	74.49
Open-source models:									
WAN2.1	86.22	86.67	84.44	–	–	–	–	–	–
Show-1	78.93	80.42	72.98	95.53	96.02	98.43	44.44	93.07	45.47
Mochi-1	80.13	82.64	70.08	96.99	97.28	99.02	61.85	86.51	50.47
Emu3	80.96	–	–	95.32	97.29	97.64	98.93	86.17	44.64
CogVideoX-5B	81.61	82.75	77.04	96.23	96.52	96.92	70.97	85.23	62.11
RepVideo	81.94	82.70	78.16	96.25	96.56	98.13	57.78	87.83	71.18
LanDiff	85.43	86.13	82.61	96.11	98.73	97.08	92.71	94.94	86.69
Our direct baselines:									
HunyuanVideo	83.24	85.09	75.82	97.37	97.76	98.99	70.83	86.10	68.55
Frame sampling	83.95	85.14	77.20	**97.40**	97.81	98.99	72.29	88.26	70.45
VTok	**85.16**	**85.48**	**80.15**	**97.40**	**98.08**	**99.02**	**76.11**	**91.50**	**74.60**

Table 2: ***VBench evaluation results.*** We compare the models in terms of their total, quality, and semantic scores, and present eight representative sub-dimensions highlighting temporal quality and guidance alignment. VTok leads to clearly improved performance over its baseline, especially in the dynamic degree and semantic dimensions. Best results over the direct baseline are bolded.

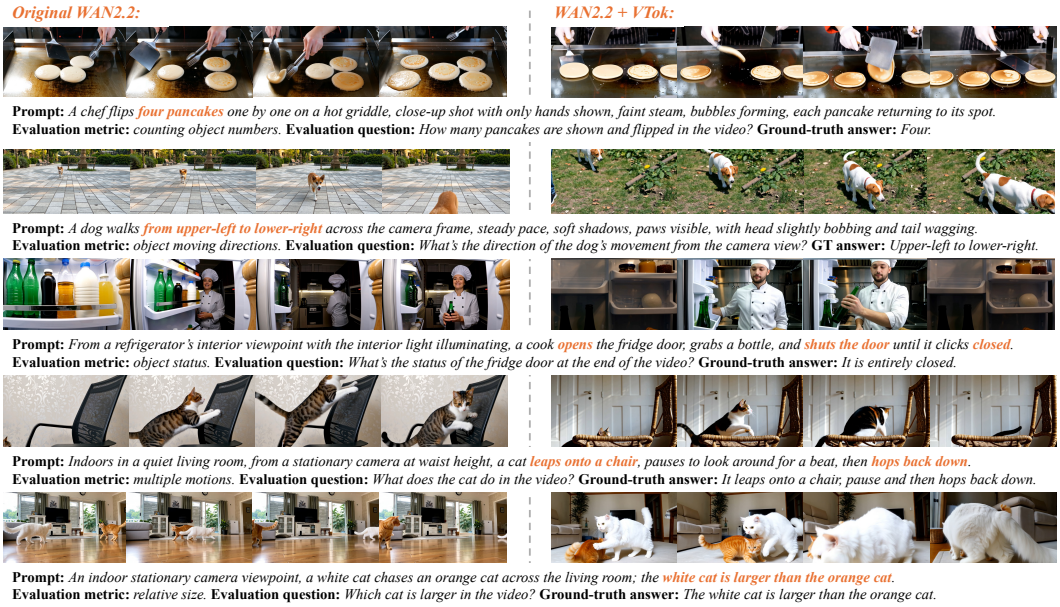

Figure 3: ***Video generation examples.*** We showcase the video generation results of the state-of-the-art open-source model WAN2.2 on our text-to-video alignment benchmark. By integrating VTok, the model demonstrates a more precise understanding of textual guidance, showing noticeably improved accuracy in following attributes such as object count, motion direction, position, and size.

quality, semantic alignment, and multiple temporal and compositional dimensions. Our VTok model achieves clear and consistent improvements over the HunyuanVideo baseline that shares the same video decoder. In particular, VTok records a dynamic degree score of 76.11%, which is substantially higher than Hunyuan's 70.83%, suggesting that our approach—letting the MLLM autoregressively sample visual tokens instead of relying solely on the diffusion process—enables the model to generate videos with faster and more complex motion dynamics. Moreover, compared with the naïve frame-sampling variant, our key-frame-plus-motion tokenization yields visibly better results across all motion-related metrics, confirming that the proposed temporal representation effectively enhances consistency and realism. Notably, VTok achieves a semantic score of 80.15%, representing a 4.33% absolute improvement over HunyuanVideo. Since this metric primarily captures text-to-video

Model	Video	Video sub.	V.-MMMU	MMVU	MVBench	LVB	LVBench
Gemini 1.5-Pro	75.0	81.3	53.9	65.4	60.5	64.0	33.1
GPT-4o	71.9	77.2	61.2	67.4	64.6	66.7	30.8
Qwen2.5-VL-72B	73.3	79.1	60.2	62.9	70.4	60.7	47.3
Qwen2.5-VL-7B	65.1	71.6	47.4	50.1	69.6	56.0	45.3
Qwen2.5-VL-3B	61.5	67.6	-	-	67.0	54.2	43.3
Our reproduced results:							
Qwen2.5-VL-7B	62.3	69.2	44.8	48.4	66.5	55.1	44.0
Qwen2.5-VL-7B VTok	62.8	69.4	45.2	48.9	67.4	55.5	44.3
Qwen2.5-VL-7B VTok finetuned	**64.9**	**70.8**	**47.3**	**50.2**	**68.5**	**56.3**	**45.0**
Llava-Next	57.8	64.9	41.3	43.2	64.0	54.1	40.6
Llava-Next VTok	58.0	65.2	41.2	43.5	64.1	54.5	40.7
Llava-Next VTok finetuned	**60.8**	**66.3**	**43.5**	**47.0**	**66.5**	**55.2**	**42.9**

Table 3: ***Video understanding results.*** We evaluate our model on seven commonly used video understanding benchmarks such as Video-MMMU (Hu et al., 2025), where the baseline results taken from (Bai et al., 2025) are de-emphasized. As shown, VTok consistently produces improved performance with or without finetuning. Column "LVB" denotes the LongVideoBench benchmark. Our results are highlighted in orange. Best results relative to direct baselines are bolded.

Model	#tokens	TV-Alignment Benchmark							
		Counting	Direction	Pos.	Size	Color	State	Motion	Avg.
No visual token (baseline)	0	31.0	46.0	51.7	37.3	45.9	28.5	43.4	40.5
Evenly sample 4 frames	64	32.3	47.2	53.5	39.1	46.9	30.5	44.5	42.0
Evenly Sample 8 frames	128	32.5	47.5	53.3	39.9	47.6	31.3	44.5	42.4
Keyframe and motion decoupled	46	**33.7**	**48.3**	**55.0**	**42.9**	**47.7**	**32.5**	**47.1**	**43.9**

Table 4: ***Ablation study of tokenization strategies.*** Column "# tokens" denotes the number of visual tokens sampled by MLLM. The HunyuanVideo (Kong et al., 2024) baseline directly conditions its decoder with text latent. We additionally ablate our framework with naive frame sampling strategies, which demonstrate degraded performance. Our default setup is highlighted in orange.

alignment, the gain highlights the advantage of our unified MLLM in learning richer and more coherent semantic representations.

Video understanding. We evaluate on seven standard benchmarks—Video-MMMU (Hu et al., 2025), MMVU (Zhao et al., 2025), MVBench (Li et al., 2024b), LongVideoBench (Wu et al., 2024a), LVBench (Wang et al., 2025b), and the video understanding benchmark (with and without sub-captions) used in Qwen2.5-VL (Bai et al., 2025)—covering short/long video comprehension, multimodal reasoning, and domain generalization. We fine-tune Qwen2.5-VL only for understanding with our tokenization, while LLaVA-Next is fine-tuned under the unified framework with joint generation and understanding objectives. As shown in Table 3, VTok consistently improves over the corresponding baselines across all benchmarks, with average gains of +1.8 for Qwen2.5-VL and +2.4 for LLaVA-Next, suggesting stronger temporal modeling and semantic reasoning. We also run a non-training control by simply replacing frame sampling with our keyframe-and-motion formulation; for this setting, we compress motion by encoding intermediate frames and pooling them into a single token. Even this lightweight swap yields noticeable gains, indicating that standard frame sampling discards important temporal cues and misses key frames, and confirming that our tokenization provides a more faithful and information-preserving representation for video understanding.

4.3 Qualitative Results

We present qualitative comparisons in Figure 3 using the same prompts, showing that WAN2.2 often misses fine-grained instructions while VTok follows them more faithfully, such as maintaining "four pancakes" consistently over time and correctly capturing the motion direction "from the top left to the bottom right." These results suggest a clearer semantic understanding of textual guidance with VTok, which we attribute to (i) an added sampling stage where the MLLM produces partial visual latents before diffusion decoding, providing the DiT with more concrete visual context for spatial-temporal

Model	#tokens	TV-Alignment Benchmark							
		Counting	Direction	Pos.	Size	Color	State	Motion	Avg.
No visual token (baseline)	0	31.0	46.0	51.7	37.3	45.9	28.5	43.4	40.5
Abalation of spatial granularity:									
VTok, 1 token for key frame	30	32.2	47.5	53.3	38.5	47.1	30.0	44.5	41.9
VTok, 4 tokens for key frame	34	33.1	47.8	54.0	39.9	47.6	31.4	45.8	42.8
VTok, 9 tokens for key frame	39	33.4	48.0	54.7	41.3	**47.7**	31.9	46.5	43.3
VTok, 16 tokens for key frame	46	33.7	48.3	55.0	42.9	**47.7**	32.5	47.1	43.9
VTok, 25 tokens for key frame	55	**33.8**	**48.5**	**55.2**	**43.0**	**47.7**	**32.7**	**47.4**	**44.0**
Ablation of temporal granularity:									
VTok, 3 frames per motion token	31	32.8	47.6	53.7	39.9	47.6	31.3	45.3	42.6
VTok, 6 frames per motion token	46	33.7	48.3	55.0	**42.9**	47.7	32.5	47.1	43.9
VTok, 12 frames per motion token	76	**33.8**	**48.4**	**55.3**	42.8	**47.9**	**32.8**	**47.4**	**44.1**

Table 5: ***Ablation study of tokenization granularity.*** To find a best efficiency-accuracy trade-off, we change the length of our visual token sequence by ablating the number of tokens used to represent each key frame (spatial granularity) and the number of frames represented by each motion token (temporal granularity). Our default setup is highlighted in orange.

conditioning, and (ii) joint training on generation and understanding tasks that improves the MLLM's high-level semantics and execution of complex instructions.

4.4 Ablation Studies

Tokenization strategy. We ablate our key insight that key-frame content and motion can be explicitly decoupled. In our default setup, each key frame is encoded as a 4×4 grid (16 spatial tokens) and motion is represented at 6 FPS; for a 5-second video at 24 FPS, this yields 46 visual tokens. To isolate tokenization effects, we also test a naïve frame-sampling variant where the MLLM predicts 4 or 8 key frames (64 or 128 tokens), while the HunyuanVideo baseline uses the MLLM only as a text encoder and conditions the DiT directly on text features. As shown in Table 4, the decoupled representation performs better (*e.g.*, 43.9% *vs.* 42.0% and 42.4%) despite using fewer tokens, suggesting that additional key frames are often redundant in short clips, whereas finer-grained temporal modeling better preserves motion continuity and improves semantic alignment.

Tokenization granularity. Table 5 studies how sequence length trades off with performance by varying spatial and temporal granularities. Increasing key-frame tokens from 1 to 16 steadily improves alignment, with 16 tokens (our default) offering the best balance; moving to 25 tokens brings only marginal gains at higher cost. "VTok, 1 token for key frame" performs the worst, indicating that video modeling benefits from explicitly preserving spatial structure rather than collapsing it into a purely 1D code (in contrast to some image-generation findings (Yu et al., 2024)). For temporal granularity, 3 frames/token slightly hurts performance, while 12 frames/token yields only minor gains over 6 frames/token but lowers temporal resolution. Overall, our default (16 spatial tokens per key frame, 6 frames per motion token) achieves the best efficiency–accuracy trade-off.

5 Conclusion

This work presents VTok, a unified video tokenizer and multimodal framework that efficiently bridges video understanding and generation. By explicitly decoupling spatial and temporal representations, VTok encodes key frames into spatial tokens and summarizes motion through compact temporal tokens, enabling a single MLLM to reason and generate within a shared token space. Experiments on VBench, our proposed TV-Align, and multiple video understanding benchmarks show that VTok achieves superior temporal fidelity, semantic alignment, and efficiency compared to frame-sampling baselines. Our ablations confirm that this spatial–temporal design provides the best trade-off between accuracy and token cost. Beyond benchmark performance, VTok also demonstrates strong generalizability across architectures such as WAN2.2 via a lightweight adapter, with qualitative results further showcasing that VTok enables better semantic grounding and more coherent motion dynamics. Overall, VTok offers a simple yet powerful foundation for building efficient, semantically aligned, and temporally consistent multimodal video generation and understanding systems.

AI USE STATEMENT

In this work, we used generative AI tools only for polishing the writing and assisting with LaTeX formatting. All AI-assisted edits were reviewed by the authors.

ETHICS STATEMENT

This paper aims to advance the field of Machine Learning. While our work may potentially have societal implications, we do not identify any that require specific discussion in this statement.

REFERENCES

Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025. 1, 3, 4

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *NeurIPS*, 2022. 3

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 1, 2, 3, 8

Fan Bao, Shen Nie, Kaiwen Xue, Yue Cao, Chongxuan Li, Hang Su, and Jun Zhu. All are worth words: A vit backbone for diffusion models. In *CVPR*, 2023. 3

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *arXiv preprint arXiv:2404.08471*, 2024. 1

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *CVPR*, 2023. 3

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *CVPR*, 2022. 3

Hao Chen, Ze Wang, Xiang Li, Ximeng Sun, Fangyi Chen, Jiang Liu, Jindong Wang, Bhiksha Raj, Zicheng Liu, and Emad Barsoum. Softvq-vae: Efficient 1-dimensional continuous tokenizer. In *CVPR*, 2025. 3

Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *CVPR*, 2024. 3

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, 2021. 3

Weichen Fan, Chenyang Si, Junhao Song, Zhenyu Yang, Yinan He, Long Zhuo, Ziqi Huang, Ziyue Dong, Jingwen He, Dongwei Pan, et al. Vchitect-2.0: Parallel transformer for scaling up video diffusion models. *arXiv preprint arXiv:2501.08453*, 2025. 3

Yuwei Fang, Willi Menapace, Aliaksandr Siarohin, Tsai-Shien Chen, Kuan-Chien Wang, Ivan Skorokhodov, Graham Neubig, and Sergey Tulyakov. Vimi: Grounding video generation through multi-modal instruction. *arXiv preprint arXiv:2407.06304*, 2024. 3

Shanghua Gao, Pan Zhou, Ming-Ming Cheng, and Shuicheng Yan. Mdtv2: Masked diffusion transformer is a strong image synthesizer. *arXiv preprint arXiv:2303.14389*, 2023. 3

Yuying Ge, Yizhuo Li, Yixiao Ge, and Ying Shan. Divot: Diffusion powers video tokenizer for comprehension and generation. In *CVPR*, 2025. 3

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 2, 5

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In *ICLR*, 2024. 3

Yuwei Guo, Ceyuan Yang, Ziyan Yang, Zhibei Ma, Zhijie Lin, Zhenheng Yang, Dahua Lin, and Lu Jiang. Long context tuning for video generation. *arXiv preprint arXiv:2503.10589*, 2025. 3

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, et al. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024. 3

Philippe Hansen-Estruch, David Yan, Ching-Yao Chung, Orr Zohar, Jialiang Wang, Tingbo Hou, Tao Xu, Sriram Vishwanath, Peter Vajda, and Xinlei Chen. Learnings from scaling visual tokenizers for reconstruction and generation. *ICML*, 2025. 3

Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos. *arXiv preprint arXiv:2501.13826*, 2025. 2, 8

Yuanhui Huang, Wenzhao Zheng, Yuan Gao, Xin Tao, Pengfei Wan, Di Zhang, Jie Zhou, and Jiwen Lu. Owl-1: Omni world model for consistent long video generation. *arXiv preprint arXiv:2412.09600*, 2024a. 3

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *CVPR*, 2024b. 2, 6

Huiwon Jang, Sihyun Yu, Jinwoo Shin, Pieter Abbeel, and Younggyo Seo. Efficient long video tokenization via coordinate-based patch reconstruction. In *CVPR*, 2025. 3

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021. 3

Yang Jiao, Haibo Qiu, Zequn Jie, Shaoxiang Chen, Jingjing Chen, Lin Ma, and Yu-Gang Jiang. Unitoken: Harmonizing multimodal understanding and generation through unified visual encoding. In *CVPR*, 2025. 3

Yang Jin, Zhicheng Sun, Kun Xu, Liwei Chen, Hao Jiang, Quzhe Huang, Chengru Song, Yuliang Liu, Di Zhang, Yang Song, et al. Video-lavit: Unified video-language pre-training with decoupled visual-motional tokenization. *ICML*, 2024. 3

Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Grant Schindler, Rachel Hornung, Vighnesh Birodkar, Jimmy Yan, Ming-Chang Chiu, et al. Videopoet: a large language model for zero-shot video generation. In *ICML*, 2024. 3

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024. 1, 2, 3, 4, 6, 8

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *Transactions on Machine Learning Research*, 2024a. 3

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 3

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *CVPR*, 2024b. 8

Tianhong Li, Huiwen Chang, Shlok Mishra, Han Zhang, Dina Katabi, and Dilip Krishnan. Mage: Masked generative encoder to unify representation learning and image synthesis. In *CVPR*, 2023. 3

Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. World model on million-length video and language with blockwise ringattention. *ICLR*, 2025. 3

Jiasen Lu, Liangchen Song, Mingze Xu, Byeongjoo Ahn, Yanjun Wang, Chen Chen, Afshin Dehghan, and Yinfei Yang. Atoken: A unified tokenizer for vision. *arXiv preprint arXiv:2509.14476*, 2025. 3

Zhuoyan Luo, Fengyuan Shi, Yixiao Ge, Yujiu Yang, Limin Wang, and Ying Shan. Open-magvit2: An open-source project toward democratizing auto-regressive visual generation. *arXiv preprint arXiv:2409.04410*, 2024. 3, 4

Nanye Ma, Mark Goldstein, Michael S Albergo, Nicholas M Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *ECCV*, 2024. 2

Xin Ma, Yaohui Wang, Xinyuan Chen, Gengyun Jia, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation. *Transactions on Machine Learning Research*, 2025. 3

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023. 2

Xiangyu Peng, Zangwei Zheng, Chenhui Shen, Tom Young, Xinying Guo, Binluo Wang, Hang Xu, Hongxin Liu, Mingyan Jiang, Wenjun Li, et al. Open-sora 2.0: Training a commercial-level video generation model in $200 k. *arXiv preprint arXiv:2503.09642*, 2025. 1, 3, 4

Hieu Pham, Zihang Dai, Golnaz Ghiasi, Kenji Kawaguchi, Hanxiao Liu, Adams Wei Yu, Jiahui Yu, Yi-Ting Chen, Minh-Thang Luong, Yonghui Wu, et al. Combined scaling for open-vocabulary image classification. *arXiv preprint arXiv:2111.10050*, 2021. 3

Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024. 3

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 3, 6

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2

Michael Ryoo, AJ Piergiovanni, Anurag Arnab, Mostafa Dehghani, and Anelia Angelova. Tokenlearner: Adaptive space-time tokenization for videos. *NeurIPS*, 2021. 3

Michael S Ryoo, Honglu Zhou, Shrikant Kendre, Can Qin, Le Xue, Manli Shu, Jongwoo Park, Kanchana Ranasinghe, Silvio Savarese, Ran Xu, et al. xgen-mm-vid (blip-3-video): You only need 32 tokens to represent a video even in vlms. *arXiv preprint arXiv:2410.16267*, 2024. 1, 3

Cuifeng Shen, Yulu Gan, Chen Chen, Xiongwei Zhu, Lele Cheng, Tingting Gao, and Jinzhi Wang. Decouple content and motion for conditional image-to-video generation. In *AAAI*, 2024. 3

Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024. 3

Anni Tang, Tianyu He, Junliang Guo, Xinle Cheng, Li Song, and Jiang Bian. Vidtok: A versatile and open-source video tokenizer. *arXiv preprint arXiv:2412.13061*, 2024. 3

Genmo Team. Mochi 1. https://github.com/genmoai/models, 2024. 1

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. 3

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 1, 2, 3, 4, 6

Feng Wang, Jieru Mei, and Alan Yuille. Sclip: Rethinking self-attention for dense vision-language inference. In *ECCV*, 2024a. 3

Hanyu Wang, Saksham Suri, Yixuan Ren, Hao Chen, and Abhinav Shrivastava. Larp: Tokenizing videos with a learned autoregressive generative prior. *ICLR*, 2025a. 3

Junke Wang, Yi Jiang, Zehuan Yuan, Bingyue Peng, Zuxuan Wu, and Yu-Gang Jiang. Omnitokenizer: A joint image-video tokenizer for visual generation. *NeurIPS*, 2024b. 3

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, et al. Lvbench: An extreme long video understanding benchmark. In *ICCV*, 2025b. 8

Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. *NeurIPS*, 2023. 3

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024c. 3, 4

Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, et al. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022. 1

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. Internvideo2: Scaling foundation models for multimodal video understanding. In *ECCV*, 2024d. 1

Yuqing Wang, Tianwei Xiong, Daquan Zhou, Zhijie Lin, Yang Zhao, Bingyi Kang, Jiashi Feng, and Xihui Liu. Loong: Generating minute-level long videos with autoregressive language models. *arXiv preprint arXiv:2410.02757*, 2024e. 3

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *NeurIPS*, 2024a. 8

Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, et al. Vila-u: a unified foundation model integrating visual understanding and generation. *arXiv preprint arXiv:2409.04429*, 2024b. 3

Haocheng Xi, Shuo Yang, Yilong Zhao, Chenfeng Xu, Muyang Li, Xiuyu Li, Yujun Lin, Han Cai, Jintao Zhang, Dacheng Li, et al. Sparse video-gen: Accelerating video diffusion transformers with spatial-temporal sparsity. In *ICML*, 2025. 3

Tianwei Xiong, Jun Hao Liew, Zilong Huang, Jiashi Feng, and Xihui Liu. Gigatok: Scaling visual tokenizers to 3 billion parameters for autoregressive image generation. *arXiv preprint arXiv:2504.08736*, 2025. 3

Le Xue, Manli Shu, Anas Awadalla, Jun Wang, An Yan, Senthil Purushwalkam, Honglu Zhou, Viraj Prabhu, Yutong Dai, Michael S Ryoo, et al. Blip-3: A family of open large multimodal models. In *ICCV*, 2025. 3

Wilson Yan, Volodymyr Mnih, Aleksandra Faust, Matei Zaharia, Pieter Abbeel, and Hao Liu. Elastictok: Adaptive tokenization for image and video. *arXiv preprint arXiv:2410.08368*, 2024. 3

Nianzu Yang, Pandeng Li, Liming Zhao, Yang Li, Chen-Wei Xie, Yehui Tang, Xudong Lu, Zhihang Liu, Yun Zheng, Yu Liu, et al. Rethinking video tokenization: A conditioned diffusion-based approach. *arXiv preprint arXiv:2503.03708*, 2025a. 3

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. In *ICLR*, 2025b. 1, 3, 6

Aoxiong Yin, Xu Tan, Kai Shen, Yichong Leng, Xinyu Zhou, Juncheng Li, and Siliang Tang. The best of both worlds: Integrating language models and diffusion models for video generation. In *ICCV*, 2025. 3, 6

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 3

Qihang Yu, Mark Weber, Xueqing Deng, Xiaohui Shen, Daniel Cremers, and Liang-Chieh Chen. An image is worth 32 tokens for reconstruction and generation. *NeurIPS*, 2024. 3, 9

Shoubin Yu, Difan Liu, Ziqiao Ma, Yicong Hong, Yang Zhou, Hao Tan, Joyce Chai, and Mohit Bansal. Veggie: Instructional editing and reasoning video concepts with grounded generation. In *ICCV*, 2025a. 3

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. *ICLR*, 2025b. 3

Hangjie Yuan, Shiwei Zhang, Xiang Wang, Yujie Wei, Tao Feng, Yining Pan, Yingya Zhang, Ziwei Liu, Samuel Albanie, and Dong Ni. Instructvideo: Instructing video diffusion models with human feedback. In *CVPR*, 2024. 3

Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021. 3

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, 2023. 3

Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model, 2024. URL https://llava-vl.github.io/blog/2024-04-30-llava-next-video/. 1, 3, 5

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Llava-video: Video instruction tuning with synthetic data. *Transactions on Machine Learning Research*, 2025. 1, 3

Long Zhao, Nitesh Bharadwaj Gundavarapu, Liangzhe Yuan, Hao Zhou, Shen Yan, Jennifer J Sun, Luke Friedman, Rui Qian, Tobias Weyand, Yue Zhao, et al. Videoprism: A foundational visual encoder for video understanding. In *ICML*, 2024. 3

Yilun Zhao, Haowei Zhang, Lujing Xie, Tongyan Hu, Guo Gan, Yitao Long, Zhiyuan Hu, Weiyuan Chen, Chuhan Li, Zhijian Xu, et al. Mmvu: Measuring expert-level multi-discipline video understanding. In *CVPR*, 2025. 8

Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*, 2024. 1, 3, 4

Ziqin Zhou, Yifan Yang, Yuqing Yang, Tianyu He, Houwen Peng, Kai Qiu, Qi Dai, Lili Qiu, Chong Luo, and Lingqiao Liu. Hitvideo: Hierarchical tokenizers for enhancing text-to-video generation with autoregressive large language models. *arXiv preprint arXiv:2503.11513*, 2025. 3