Video-LaVIT: Unified Video-Language Pre-training with Decoupled Visual-Motional Tokenization

Yang Jin [1] **Zhicheng Sun** [1] **Kun Xu** [2] **Kun Xu** [2] **Liwei Chen** [2] **Hao Jiang** [1] **Quzhe Huang** [1] **Chengru Song** [2]
Yuliang Liu [2] **Di Zhang** [2] **Yang Song** [2] **Kun Gai** [2] **Yadong Mu** [1]

Abstract

In light of recent advances in multimodal Large Language Models (LLMs), there is increasing attention to scaling them from image-text data to more informative real-world videos. Compared to static images, video poses unique challenges for effective large-scale pre-training due to the modeling of its spatiotemporal dynamics. In this paper, we address such limitations in video-language pre-training with an efficient video decomposition that represents each video as keyframes and temporal motions. These are then adapted to an LLM using well-designed tokenizers that discretize visual and temporal information as a few tokens, thus enabling unified generative pre-training of videos, images, and text. At inference, the generated tokens from the LLM are carefully recovered to the original continuous pixel space to create various video content. Our proposed framework is both capable of comprehending and generating image and video content, as demonstrated by its competitive performance across 13 multimodal benchmarks in image and video understanding and generation. Our code and models are available at https://video-lavit.github.io.

1. Introduction

Recently, the significant breakthrough of Large Language Models (LLMs) (Brown et al., 2020; Touvron et al., 2023a) has brought a surge in building general-purpose multimodal AI assistants (OpenAI, 2023b; Gemini Team, 2023) that can follow both textual and visual instructions. Drawing on the remarkable reasoning abilities of LLMs and knowledge in massive alignment corpus (e.g., image-text pairs), they showcase the great potential of accurately compre-

Efficient motion vectors (saving > 90% tokens)
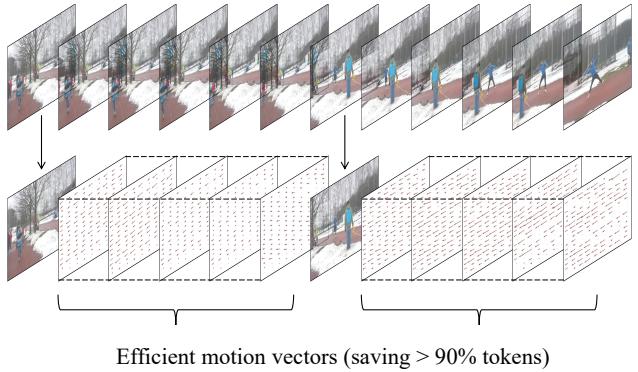
Figure 1. The key observation in this work is: most video parts have a high degree of temporal redundancy that may be described by motion vectors. By exploiting these motion vectors, the video can be efficiently tokenized for pre-training of multimodal LLMs.

hending and generating visual content (Sun et al., 2024; Jin et al., 2024; Dong et al., 2024). Despite their success, these multimodal LLMs (Alayrac et al., 2022; Liu et al., 2023c) predominantly concentrate on the image-text data, leaving the adaptation for video modality less explored. In contrast to static images, video serves as a dynamic media form that is more in line with human visual perception. Learning effectively from video is particularly essential for enhancing machine intelligence to comprehend the real world.

To this end, several approaches have made attempts at harnessing the generative capabilities of LLMs for handling video data. Inheriting the successful paradigm from the image domain, they represent video as a sequence of visual tokens that aligns with LLMs' semantic space by utilizing a pre-trained 2D image model (Li et al., 2023d; Zhang et al., 2023) or a 3D video backbone (Kondratyuk et al., 2023). Nevertheless, the existing designs are still not competent for effectively encoding videos. Compared to images, videos pose unique challenges associated with higher demands for learning complex spatiotemporal clues, such as time-varying actions and scene changes. In this regard, encoding individual video frames separately by the 2D visual encoder falls short of capturing the temporal motion information, which plays a vital role in identifying distinct behaviors and events within the video content. Although the recent

---

concurrent work VideoPoet (Kondratyuk et al., 2023) crafts a 3D video tokenizer for video generation with LLM, its applicability is constrained to short video clips due to the use of long token sequences (e.g., 1280 tokens for a 2.2s clip). When it comes to understanding or generating long videos, inputting excessive numbers of tokens into LLMs is deemed unacceptable in terms of computational resources.

This work addresses the limitation in video-language pre-training by exploring an efficient video representation that decomposes video into keyframes and temporal motions. Our motivation is built upon the natural characteristics of video data itself. As illustrated in Figure 1, a video is typically divided into several shots, where video frames within each shot often exhibit substantial information redundancy. It is superfluous to encode all of these frames as tokens and incorporate them into the generative pre-training of LLMs. This fact strongly spurs us to decompose each video into alternating keyframes and motion vectors, where the former encapsulate the primary visual semantics and the latter depict the dynamic evolution of its corresponding keyframe over time. There are several benefits to such decomposed representation: (1) Compared to processing consecutive video frames utilizing 3D encoders, the combination of a single keyframe and motion vectors requires fewer tokens to represent video temporal dynamics, which is more efficient for large-scale pre-training. (2) The model can inherit the acquired visual knowledge from an off-the-shelf image-only LLM and focus solely on modeling temporal information without learning from scratch.

Based on the above motivations, we present **Video-LaVIT** (**La**nguage-**VI**sion **T**ransformer), a new multimodal pre-training approach that effectively empowers LLMs to comprehend and generate video content in a unified framework. Specifically, Video-LaVIT incorporates two core components: a *tokenizer* and a *detokenizer* to handle video modality. The video tokenizer aims to transform the continuous video data into a sequence of compact discrete tokens akin to a foreign language, where the keyframes are processed by utilizing an established image tokenizer (Jin et al., 2024). For converting the temporal motions into the compatible discrete format, a spatiotemporal motion encoder is devised. It can capture the time-varying contextual information contained in extracted motion vectors, thereby significantly enhancing LLMs' ability to comprehend the intricate actions in video. The video detokenizer is responsible for mapping the discretized video token generated by LLMs back into its original continuous pixel space. During training, video is represented as an alternating discrete visual-motion token sequence, and thus can be optimized under the same next-token prediction objective together with different modalities. Since video is inherently a time series, this joint autoregressive pre-training contributes to learning the sequential relationships of different video clips. We found that Video-

LaVIT, is capable of serving as a multimodal generalist to achieve promising results in both understanding and generation tasks without further fine-tuning. The key contributions of this work are summarized as:

- We introduce Video-LaVIT, a multimodal pre-training method that pushes the limit of LLMs' unified understanding and generation capability towards video.

- To efficiently model visual and temporal information in video, Video-LaVIT incorporates a novel video tokenizer and detokenizer that operates on the decomposed representations of keyframes and motion vectors.

- Experiments on 13 multimodal benchmarks demonstrate that Video-LaVIT achieves very competitive performance, ranging from image and video comprehension to zero-shot text-to-image and text-to-video generation.

2. Related Work

Vision-language pre-training. Following the success of using large-scale image-text pairs for contrastive learning of vision-language models (Radford et al., 2021), a similar idea has been exploited in generative pre-training, where visual and language data are jointly modeled under an autoregressive process. In practice, this is typically achieved by adapting visual image inputs to pre-trained LLMs (Raffel et al., 2020; Brown et al., 2020; Touvron et al., 2023a) via an intermediate module like cross-attention (Alayrac et al., 2022), Q-Former (Li et al., 2023c), or linear projection (Liu et al., 2023c). More recent approaches such as CM3Leon (Yu et al., 2023a) and LaVIT (Jin et al., 2024) advocate the use of discrete visual tokenizers (van den Oord et al., 2017; Esser et al., 2021) to form a unified next token prediction objective. However, these methods are primarily focused on image-text data and cannot be directly extended to videos due to the significantly higher computational cost.

Video understanding and generation. By unifying videos in the above pre-training framework, remarkable progress has been made in video comprehension with masked (Yang et al., 2022) and autoregressive language models (Li et al., 2023d; Zhang et al., 2023; Maaz et al., 2023). However, for video generation, the mainstream approaches are still based on diffusion models (Sohl-Dickstein et al., 2015; Song & Ermon, 2019; Ho et al., 2020), which enhance existing image pre-trained models with better temporal consistency (Ho et al., 2022; Singer et al., 2023; Blattmann et al., 2023b; Esser et al., 2023; Blattmann et al., 2023a). Language model based counterparts (Yan et al., 2021; Hong et al., 2023; Kondratyuk et al., 2023), on the other hand, face the critical challenge of efficiently encoding video temporal dynamics with limited context windows and computational resources. In response, our work leverages motion vectors, a classic
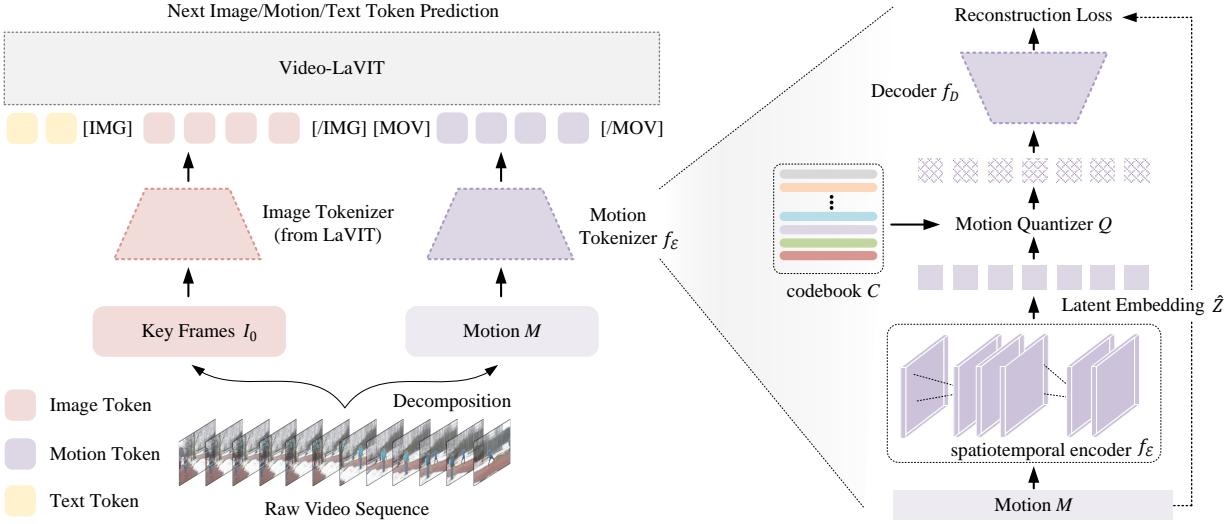
Figure 2. For each video-text pair, Video-LaVIT decomposes the video into keyframes and motion vectors for efficient tokenization. The tokenizers are learned by maximally reconstructing original inputs (e.g., the motion tokenizer is shown on the right). Finally, the encoded tokens are concatenated with text tokens to form a multimodal sequence, allowing for unified generative pre-training of the LLM (left).

and effective cue in video modeling (Zhang et al., 2016; Wang et al., 2023b; Shen et al., 2024), for improving the efficacy of LLM-based video comprehension and generation.

3. Method

This work aims to present an effective pre-training framework that harnesses the exceptional modeling capability of Large Language Models (LLMs) to facilitate the learning of video modality. In pursuit of this goal, we highlight two core designs: a video *tokenizer* (Section 3.1) which allows for the representation of all modalities in a unified discrete form, and a video *detokenizer* (Section 3.2) to map the generated discrete tokens back to the continuous pixel space. Coped with these two main components, Video-LaVIT can be optimized through a unified autoregressive training paradigm (Section 3.3), enabling it to simultaneously comprehend and generate various multimodal content.

3.1. Video Tokenization

To encode an untrimmed video as inputs to LLMs, the prevailing approaches (Lin et al., 2023; Li et al., 2023d) mainly uniformly downsample the original video into a series of frames. Then, a pre-trained ViT encoder (Radford et al., 2021; Fang et al., 2023) is employed to separately encode these frames and produce a sequence of frame-level embeddings as the video representation. This straightforward way disregards the modeling of temporal dynamics between frames, thus impeding the capacity to understand the actions and camera transitions occurring in the video. While the utilization of 3D video encoders in very recent (Kondratyuk

et al., 2023) enables the encoding of temporal information, it only applies to short video clips and inevitably yields a substantial proliferation of tokens (e.g., 1280 tokens for one 2.2s clip), resulting in a heavy computational overhead.

Motion-aware Video Decomposition. Given the above concerns, our proposed video tokenizer seeks to integrate temporal dynamics into the video representations efficiently. We observe that a video clip captured in the same shot can convey its primary semantics through a single keyframe, while the subsequent frames only illustrate the temporal evolvement based on that keyframe. This property empowers the decomposed video tokenization for keyframe and temporal motion. For the keyframe, we employ an off-the-shelf image tokenizer from LaVIT (Jin et al., 2024) to inherit the learned visual codebook and prior knowledge without training from scratch. For encoding temporal motion information, a common alternative is to calculate hand-crafted dense optical flow between adjacent frames (Beauchemin & Barron, 1995). Despite providing a fine-grained depiction of object motions in videos, the expensive computations render it unsuitable for scaling to large-scale video data during pre-training. Hence, we resort to motion vectors, which can be directly extracted at high speed on the CPU (Wu et al., 2018) during the compressed video decoding process.

As illustrated in Figure 2, we employ the MPEG-4 (Le Gall, 1991) compression technique to extract keyframe and motion information. For simplicity, the I-frames in MPEG-4 are considered as the keyframes requiring tokenization. More sophisticated (but expensive) keyframe selection schemes can also be considered, but are not the main focus of this work. Formally, each video frame is partitioned into

(a) Detokenizer Training (b) Long Video Decoding
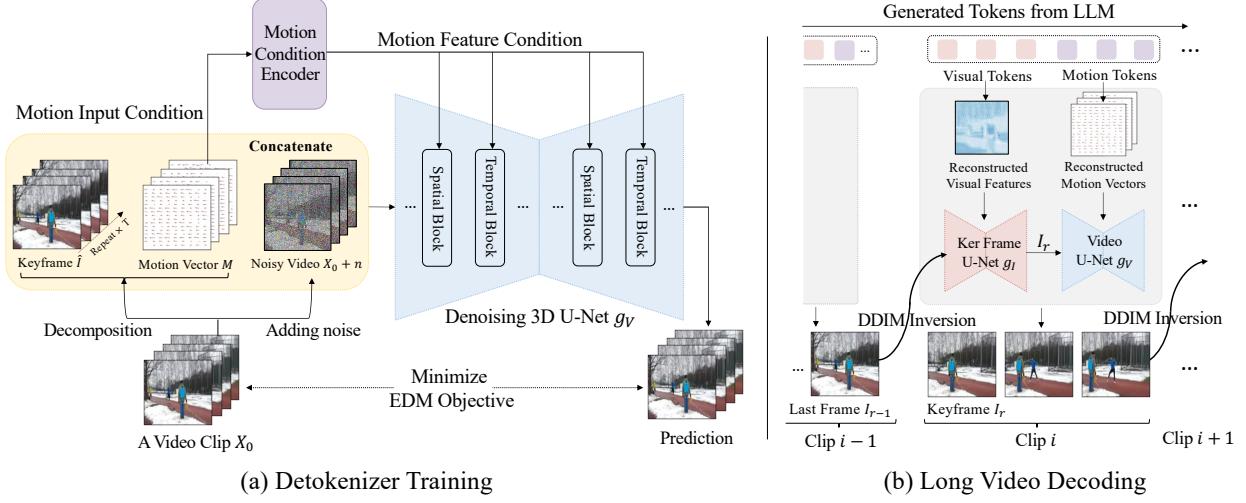
Figure 3. Illustrations for video detokenization in Video-LaVIT. (a) Training pipeline for the video detokenizer, which aims to reconstruct the original video clip using one keyframe and the subsequent motion vectors. (b) Autoregressive inference for long video decoding.

16×16 non-overlapping macroblocks. Motion vectors $\vec{m}$ of the t-th frame are estimated by finding the best macroblock correspondence between adjacent frames I_t and I_{t-1}:

$$\vec{m}(p, q) = \arg\min_{i,j} \|I_t(p, q) - I_{t-1}(p - i, q - j)\|, \quad (1)$$

where $I(p, q)$ indicates the pixel values of the macroblock at location (p, q), and (i, j) is the coordinate offset between the center of two macroblocks. Then, a video clip can be decomposed into a keyframe $I_0 \in \mathbb{R}^{H \times W \times 3}$ and the motion vectors $M \in \mathbb{R}^{T \times \frac{H}{16} \times \frac{W}{16} \times 2}$ of its subsequent T frames.

Motion Vector Tokenization. To transform the motion vectors into a sequence of discrete tokens like a foreign language, we develop a motion-specific tokenizer based on the VQ-VAE architecture (van den Oord et al., 2017). It includes a spatiotemporal encoder $f_\mathcal{E}$, a learnable codebook $\mathcal{C} = \{c_k\}_{k=1}^K$, and a decoder $f_\mathcal{D}$. The encoder $f_\mathcal{E}$ has L stacked transformer blocks consisting of spatial and temporal attention layers to fuse the contextual motion information among the T frames. It maps the motion vectors $M \in \mathbb{R}^{T \times \frac{H}{16} \times \frac{W}{16} \times 2}$ into a 1D latent embedding sequence $\hat{Z} \in \mathbb{R}^{N \times d}$. Each embedding vector $\hat{z} \in \mathbb{R}^d$ is then tokenized by a vector quantizer Q, which assigns it to the closest code in $\mathcal{C}$:

$$z_i = \arg\min_j \|l_2(\hat{z}_i) - l_2(c_j)\|_2, \quad (2)$$

where l_2 indicates the L_2 normalization. The decoder $f_\mathcal{D}$ has a similar structure to the encoder and is obliged to map the discrete motion codes $\{z_i\}_{i=1}^N$ back to the original motion vectors. The whole motion tokenizer can be updated by optimizing the reconstruction quality. To prevent codebook collapse during training, we follow Yu et al. (2022) to project the motion embeddings $\hat{Z}$ into a low-dimensional

space before quantization and use exponential moving average (EMA) updates. More details about the motion tokenizer can be found in Appendix A.1. Finally, a video is tokenized into alternating $\langle visual, motion, ... \rangle$ codes that serve as the supervision signals in LLMs during generative pre-training. Such a factorized tokenization significantly reduces the inter-frame redundancy in one video shot while efficiently capturing the temporal motion information.

3.2. Video Detokenization

The video detokenizer of Video-LaVIT is in charge of converting them back into the original continuous pixel space for video generation. Considering the challenge in learning a direct mapping from discrete tokens to the high-dimensional video space, we take a sequential decoding strategy, wherein the keyframe is initially recovered based on the visual token. The subsequent frames are then decoded by taking both the keyframe and motion tokens as the conditions. The efficacy of this strategy in enhancing video generation quality has also been validated by recent work (Girdhar et al., 2023).

Specifically, the keyframe and video detokenizers both use conditional denoising U-Net (Rombach et al., 2022). Similar to LaVIT (Jin et al., 2024), the keyframe U-Net g_I takes the reconstructed visual features that contain image semantics as conditions to infill visual details from a Gaussian noise. Here, we primarily focus on the newly proposed video detokenizer g_V. As illustrated in Figure 3(a), it is a 3D variant of the original 2D U-Net architecture by inserting temporal convolution and attention layers after the spatial modules, following Blattmann et al. (2023b; 2023a).

Enhanced Motion Conditioning. The objective of the video detokenizer g_V is to rigorously adhere to the guidance of the motion vectors, thereby facilitating the recovery of T

frames following the keyframe. To this end, we highlight two different forms of motion conditions in g_V. Given the motion vectors $M \in \mathbb{R}^{T \times \frac{H}{16} \times \frac{W}{16} \times 2}$ of a sampled video clip, we adopt the nearest neighbor interpolation to ensure that it matches the spatial shape of the U-Net input. Also, the latent state $\hat{I}$ of the keyframe from the VAE is repeated T times along the temporal axis to form visual conditioning. The motion vector M, the keyframe latent $\hat{I}$, and the noisy video frames are concatenated channel-wise as the input condition to g_V. Except for direct input conditioning, we also enhance conditioning with motion feature embedding via the spatial and temporal cross-attention layers in the 3D U-Net blocks. Here, the motion features are from a conditioning encoder that has a similar architecture to $f_{\mathcal{E}}$ excluding the downsample layer to reduce potential information loss. The parameters of the video detokenizer g_V are updated by minimizing the following EDM training objective (Karras et al., 2022) on a video training dataset $\mathcal{D}$:

$$\mathbb{E}_{(X_0, \hat{I}, \hat{M}) \sim \mathcal{D}, \sigma, n} \left[\lambda_\sigma ||g_V(X_0 + n, \sigma, \hat{I}, M) - X_0|| \right],$$
(3)

where $\sigma \sim p(\sigma)$ is the noise level during training, $n \sim \mathcal{N}(n; 0, \sigma^2)$ is a random noise added to video sample X_0, and λ_σ is loss weighting function. At inference, the $\langle visual, motion \rangle$ tokens yielded by LLM are first mapped into visual features and motion vectors by their corresponding tokenizers. The reconstructed visual features are fed into g_I to generate a keyframe, which is subsequently combined with reconstructed motion vectors to serve as conditions for g_V to decode the video clip (See Figure 3(b)).

Long Video Decoding. Since a video is expressed as multiple alternating $\langle visual, motion \rangle$ sequences, the interdependencies among different video fragments can be effectively learned by autoregressive pre-training of LLMs. Hence, Video-LaVIT naturally supports the generation of longer videos by progressive decoding multiple clips. However, separate decoding will bring inconsistencies in some fine-grained visual details among different clips (See Figure 5). To mitigate this, we incorporate an explicit noise constraint when decoding the keyframe I_r of a video clip. As illustrated in Figure 3(b), we reverse its last frame I_{r-1} from the previously generated clip into an intermediate noisy state $x_{\Delta T}$ by reversing the DDIM sampling (Song et al., 2020) process ΔT times. Each inversion step is formulated by:

$$x_{t+1} = \sqrt{\frac{\alpha_{t+1}}{\alpha_t}} x_t + \left(\sqrt{\frac{1}{\alpha_{t+1}} - 1} - \sqrt{\frac{1}{\alpha_t} - 1} \right) g_I(x_t, t, \hat{I}),$$
(4)

where α_t is the noise level, $\hat{I}$ is the visual feature condition. The reversed noisy state $x_{\Delta T}$ is then considered as the initial noise in the denoising loop for keyframe I_r. As illustrated in Figure 5, adding this noise constraint can improve the temporal consistency between video clips.

3.3. Unified Generative Modeling

Based on the developed decomposed video tokenization strategy, it is feasible to indiscriminately treat all the modalities (video, image, and text) as 1D discrete tokens fed into LLMs. Following LaVIT (Jin et al., 2024), special tokens (e.g., [MOV] and [/MOV] for motion modality) are inserted at the beginning and end of the visual and motion token sequence for differentiating modalities in the input data. During pre-training, we also exchange the order of multimodal data pairs to form both [video(image), text] and [text, video(image)] as input sequences. Formally, given a multimodal sequence $y = (y_1, y_2, .., y_S)$, Video-LaVIT inherits the successful generative language modeling paradigm from LLM to directly maximize the likelihood of each token y_i in an autoregressive manner:

$$p(y) = \sum_{y \in \mathcal{D}} \sum_{i=1}^{S} \log P_\theta(y_i | y_{<i}).$$
(5)

After pre-training, Video-LaVIT is capable of serving as a multimodal generalist to achieve both multimodal comprehension and generation of data in any modality.

Model Training. Video-LaVIT undergoes a three-stage training procedure on the large-scale multimodal corpora. The purpose of each stage can be summarized as follows: i) Tokenizer and Detokenizer Training. This stage requires only pure video data without corresponding textual captions. It aims to produce compact video tokens that serve as supervision signals to guide the subsequent generative pre-training, as well as to facilitate an accurate reconstruction of the original videos. ii) Generative Pre-training. Stage-2 empowers the model to learn the inter-correlation among the data of different modalities via unified generative modeling within the LLM. iii) Instruction Tuning. To fully unleash the acquired knowledge, the last stage further improves the instruction-following ability to accomplish various multimodal tasks. More details about the model architectures and training data for each stage are provided in Appendix A.1.

4. Experiments

4.1. Multimodal Understanding

With the decomposed video representation, Video-LaVIT is naturally capable of understanding both videos and images. Here, we demonstrate its multimodal understanding capability on 11 commonly used image and video benchmarks.

Image Understanding. Table 1 presents an extensive comparison across eight widely used image question answering and multimodal benchmarks: VQA v2 (Goyal et al., 2017), GQA (Hudson & Manning, 2019), VizWiz (Gurari et al., 2018), ScienceQA-IMG (Lu et al., 2022), MME (Fu et al., 2023), MMBench (Liu et al., 2023e), SEED (Li et al.,

Table 1. Image understanding performance (↑) on 8 benchmarks. Video-LaVIT achieves state-of-the-art results on most of the benchmarks. For convenience, SQA[I] denotes ScienceQA-IMG (Lu et al., 2022), and MMB denotes MMBench (Liu et al., 2023e). * indicates that there is some overlap with the training data. Note that only LLaVA-1.5 (Liu et al., 2023a) is reported with a higher image resolution of 336. The Video-LLaVA, LLaMA-VID and LLaVA-1.5 use Vicuna-1.5 (Chiang et al., 2023) as the language model.

Method	LLM size	Image Question Answering				Multimodal			
		VQA^{v2}	GQA	VizWiz	SQA^{I}	MME	MMB	SEED	MM-Vet
Flamingo (Alayrac et al., 2022)	9B	51.8	-	28.8	-	-	-	-	-
BLIP-2 (Li et al., 2023b)	13B	41.0	41.0	19.6	61.0	1293.8	-	46.4	22.4
InstructBLIP (Dai et al., 2023)	13B	-	49.5	34.3	63.1	1212.8	44.0	-	25.6
CM3Leon (Yu et al., 2023a)	7B	47.6	-	37.6	-	-	-	-	-
Emu (Sun et al., 2024)	13B	52.0	-	34.2	-	-	-	-	36.3
DreamLLM (Dong et al., 2024)	7B	72.9*	-	49.3	-	-	58.2	-	**36.6**
Video-LLaVA (Lin et al., 2023)	7B	74.7*	60.3*	48.1	66.4	-	60.9	-	32.0
LLaMA-VID (Li et al., 2023f)	7B	78.3*	63.0*	52.5	67.7	1405.6	65.3	59.7	-
LLaVA-1.5 (Liu et al., 2023a)	7B	78.5*	62.0*	50.0	66.8	1510.7	64.3	58.6	30.5
Video-LaVIT	7B	**80.3***	**64.4***	**56.0**	**70.0**	**1551.8**	**67.3**	**64.0**	33.2

Table 2. Zero-shot video question answering accuracy (↑). Video-LaVIT demonstrates state-of-the-art accuracy on all three benchmarks. The evaluation uses a GPT assistant (Maaz et al., 2023), with "Score" denoting a relative score from 0 to 5 assigned by the GPT model. The Video-LLaVA and LLaMA-VID both use Vicuna-1.5 (Chiang et al., 2023) as the language model.

Method	LLM size	MSVD-QA		MSRVTT-QA		ActivityNet-QA	
		Accuracy	Score	Accuracy	Score	Accuracy	Score
FrozenBiLM (Yang et al., 2022)	1B	32.2	-	16.8	-	24.7	-
Video-LLaMA (Zhang et al., 2023)	7B	51.6	2.5	29.6	1.8	12.4	1.1
VideoChat (Li et al., 2023d)	7B	56.3	2.8	45.0	2.5	26.5	2.2
Video-ChatGPT (Maaz et al., 2023)	7B	64.9	3.3	49.3	2.8	35.2	2.7
LLaMA-VID (Li et al., 2023f)	7B	69.7	3.7	57.7	3.2	47.4	3.3
Video-LLaVA (Lin et al., 2023)	7B	70.7	3.9	59.2	**3.5**	45.3	3.3
Video-LaVIT	7B	**73.2**	**3.9**	**59.3**	3.3	**50.1**	**3.3**

2023a), MM-Vet (Yu et al., 2023b). Our model successfully generalizes the pre-training knowledge to image comprehension tasks and provides the best overall performance. Specifically, with the same instruction dataset and the base model as LLaVA-1.5 (Liu et al., 2023a), our method consistently yields the best results on all image question answering datasets. For example on SQA^{I}, it surpasses LLaVA-1.5 which has a higher input resolution by 3.2%, while consistently outperforming the other video-language models. The same advantages are further validated on more comprehensive multimodal benchmarks, where our model leads on three out of four benchmarks.

Zero-Shot Video Question Answering. Table 2 compares our proposed Video-LaVIT with multiple recent video-language models on three common video benchmarks: MSVD-QA (Chen & Dolan, 2011), MSRVTT-QA (Xu et al., 2016) and ActivityNet-QA (Yu et al., 2019), in terms of accuracy and relative score measured by a GPT-3.5 assistant (Maaz et al., 2023). We achieve state-of-the-art accuracies and very competitive relative scores on the three benchmarks, such as surpassing the previous leading model Video-LLaVA (Lin et al., 2023) by 2.5% on MSVD-QA.

Using the same 100k video-text instruction dataset from Video-ChatGPT (Maaz et al., 2023) which is also adopted by Video-LLaVA, our method outperforms these alternatives by explicitly modeling temporal dynamics with motion tokens. Especially for the ActivityNet-QA benchmark, which contains various human behaviors, incorporating motion information contributes to the recognition of different actions. For the only metric where our performance is not the best, namely the relative score on MSRVTT-QA, we deliver a high score only second to Video-LLaVA (by a margin of 0.2), again confirming the effectiveness of our method.

Zero-Shot Video Understanding. Besides the widely-used video question answering datasets, we also evaluated Video-LaVIT on Perception Test (Patraucean et al., 2024) or EgoSchema (Mangalam et al., 2024). These two benchmarks aim to evaluate the understanding and reasoning capability of long-term videos, rather than exploiting the hallucination capabilities of LLMs. The detailed evaluation results are shown in Table 3. On the Perception Test, Video-LaVIT delivers the highest zero-shot performance. Notably, it outperforms VideoChat2 (Li et al., 2023e), which uses 1.9M additional instruction tuning data (including videos)

Table 3. Zero-shot understanding (↑) on the test set of Perception Test (Patraucean et al., 2024) and EgoSchema (Mangalam et al., 2024).

Method	Flamingo (Alayrac et al., 2022)	BLIP-2 (Li et al., 2023c)	VideoChat2 (Li et al., 2023e)	Video-LaVIT
Accuracy	33.5	39.2	47.3	**47.9**

Method	FrozenBiLM (Yang et al., 2022)	mPLUG-Owl (Ye et al., 2023)	InternVideo (Wang et al., 2022)	Video-LaVIT
Accuracy	26.9	28.7	32.1	**37.3**

Table 4. Zero-shot text-to-video generation performance. Video-LaVIT delivers competitive results against state-of-the-art models trained on more proprietary data, with data size reported in terms of the number of training video clips. The next best results are underlined.

Method	Data size	Public data	MSR-VTT			UCF-101	
			CLIPSIM (↑)	FVD (↓)	FID (↓)	IS (↑)	FVD (↓)
CogVideo (Hong et al., 2023)	5.4M	✓	0.2631	1294	23.59	25.27	701.59
Video LDM (Blattmann et al., 2023b)	10M	✓	0.2929	-	-	33.45	550.61
VideoComposer (Wang et al., 2023b)	10M	✓	0.2932	580	-	-	-
InternVid (Wang et al., 2024)	28M	✓	0.2951	-	-	21.04	616.51
Make-A-Video (Singer et al., 2023)	20M	✓	**0.3049**	-	13.17	33.00	367.23
VideoPoet (Kondratyuk et al., 2023)	270M	✗	**0.3049**	<u>213</u>	-	38.44	355.00
PYoCo (Ge et al., 2023)	22.5M	✗	-	-	**9.73**	**47.76**	355.19
SVD (Blattmann et al., 2023a)	152M	✗	-	-	-	-	**242.02**
Video-LaVIT	10M	✓	<u>0.3012</u>	**188.36**	<u>11.27</u>	<u>44.26</u>	<u>280.57</u>

to improve video understanding. In comparison, our advantageous performance is achieved with the standard instructions from LLaVA-1.5 (Liu et al., 2023a) and Video-ChatGPT (Maaz et al., 2023) (amounting to 765K), demonstrating the effectiveness of our proposed method. As for EgoSchema, which focuses on long video understanding, Video-LaVIT is able to analyze 16 keyframes (with the motion vectors in between) spanning 64 seconds, thereby deliver better results. For example, it outperforms Intern-Video (Wang et al., 2022), which uses up to 90 frames, by a significant 5.2% in zero-shot QA accuracy. This validates the efficacy of the visual-motional decomposition for modeling long-term temporal information.

4.2. Multimodal Generation

By unified generative pre-training, Video-LaVIT can flexibly generate both video and images. Due to page limitations, we present here its text-to-video generation results, while the text-to-image evaluation is discussed in Appendix B.1.

Zero-Shot Text-to-Video Generation. Table 4 summarizes the model performance on MSR-VTT (Xu et al., 2016) and UCF-101 (Soomro et al., 2012), in terms of CLIP similarity (CLIPSIM) (Wu et al., 2021), Fréchet video distance (FVD) (Unterthiner et al., 2018), Fréchet Inception distance (FID) (Heusel et al., 2017), and Inception score (IS) (Saito et al., 2020). Overall, our model significantly outperforms most baselines using similar public datasets, and is highly competitive against models trained on much larger proprietary data, for example leading the FVD on MSR-VTT. In particular, when compared to language model-based

text-to-video generators, our method consistently outscores CogVideo (Hong et al., 2023), while surpassing the recent concurrent work VideoPoet (Kondratyuk et al., 2023), which uses a 3D video tokenizer trained on the much larger data. This clearly validates the superiority of our tokenizer design.

Zero-Shot Long Video Generation. We also conducted quantitative evaluation experiments for long video generation, following the setting from FreeNoise (Qiu et al., 2023). Specifically, it is evaluated on 2048 long videos (64 frames) generated using the prompts from EvalCrafter (Liu et al., 2023d). As shown in the table Table 5, our approach yields highly competitive performance among the specialists curated for long video generation. In particular, it surpasses all baselines on the KVD metric, which measures the discrepancy between short videos (first 16 frames) and subsets of long videos (last 16 frames). These results confirm the effectiveness of our proposed long video decoding strategy with explicit noise constraint.

4.3. Qualitative Results

This section compares videos created by Video-LaVIT with state-of-the-art results under both text and image conditions. It also presents our special ability to generate long videos. More visualization examples are provided in Appendix B.1.

The text-to-video and image-to-video generation results are visualized in Figure 4. For text-to-video generation, our method can produce visual quality not much far from the closed-source model Gen-2 (Runaway, 2023), thanks to the unified pre-training framework with images. Meanwhile, Video-LaVIT is advantageous in reasoning abilities, such as

A majestic eagle soars gracefully over a breathtaking mountain range.

A steam train moving on a mountainside by Vincent van Gogh.
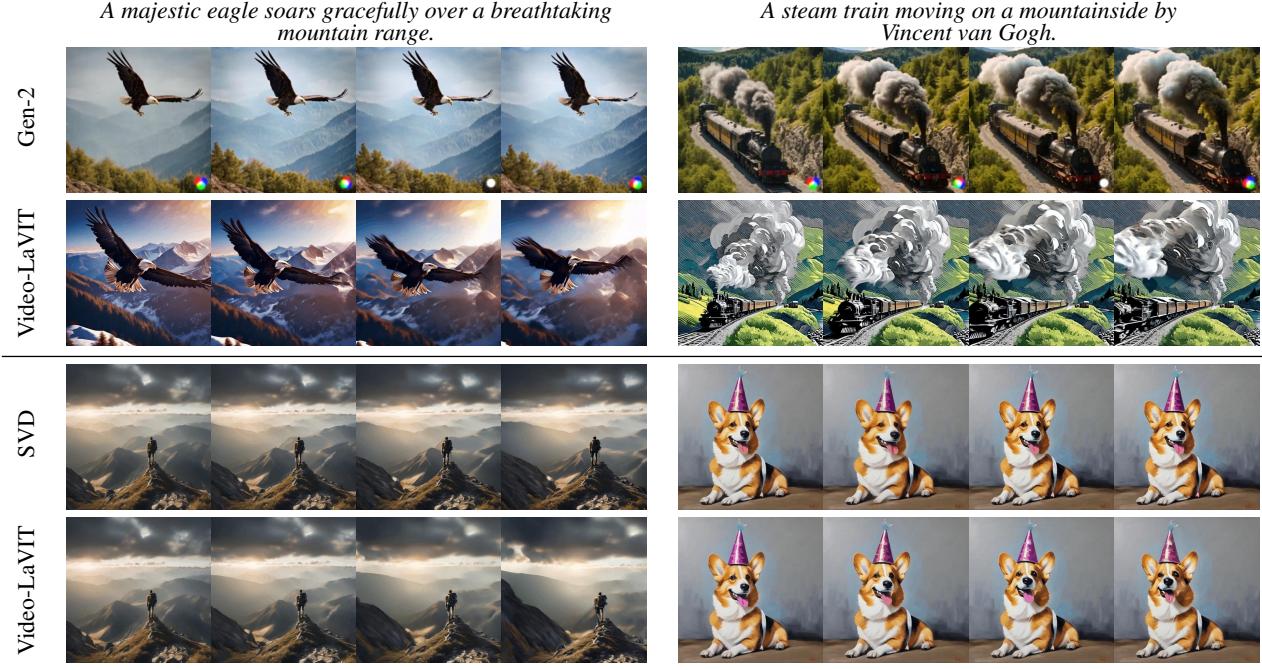

Figure 4. Text-to-video (top) and image-to-video (bottom) generation comparison with Gen-2 (Runaway, 2023) and SVD-XT (Blattmann et al., 2023a). Text prompts are from Emu Video (Girdhar et al., 2023) and SVD. The I2V generation is conditioned on the leftmost frame.

inferring better motion (in the top-left example) and adding artistic touches based on the text prompt (as in both cases). For image-to-video generation, our method is comparable to the state-of-the-art model SVD (Blattmann et al., 2023a) in generating both coherent and highly aesthetic video clips (the bottom-left example). In addition, the decomposed video representation enables the video decoder to produce more salient and vivid movements given a relatively difficult synthetic image prompt (the bottom-right example).

Furthermore, our autoregressive model can be naturally extended to long video generation, as shown in Figure 5. Thanks to the proposed explicit noise constraint when decoding consecutive video clips, the temporal consistency between decoded clips is greatly improved. In contrast, decoding each video clip separately will result in the incoherence of fine-grained visual details among the video frames of different clips (See the bottom of Figure 5).

4.4. Ablation Study

This section investigates the impact of motion tokenization and different motion token lengths. Due to limited space, the ablation for proposed enhanced motion conditioning strategy is provided in Appendix B.1.

Effect of Motion Tokenization. We design two baselines to validate the effectiveness of motion tokenization in video pre-training. For video understanding, the w/o motion in Table 6 indicates the independent tokenization of 16 uniformly sampled video frames by the 2D visual encoder without any

Table 5. Zero-shot text-to-long video generation performance. It is evaluated on 2048 long videos (64 frames) generated using the prompts from EvalCrafter (Liu et al., 2023d).

Method	FVD ($\downarrow$)	KVD ($\downarrow$)	CLIPSIM ($\uparrow$)
Direct	737.61	359.11	0.9104
Sliding	224.55	44.09	0.9438
Gen-L-Video (Wang et al.)	177.63	21.06	0.9370
FreeNoise (Qiu et al.)	**85.83**	6.07	**0.9732**
Video-LaVIT	113.37	**4.94**	0.9621

motion tokens. Most existing methods use a similar strategy to encode video content fed into the LLM. As observed in Table 6, the question-answering accuracy decreases without explicitly modeling temporal information. For the video synthesis, the w/o motion baseline divides the text-to-video process into two separate stages: text-to-image and image-to-video. Specifically, given a textual prompt, we generate only a keyframe (without producing motion tokens) and then feed this keyframe into the image-to-video generation model svd-img2vid-xt (Blattmann et al., 2023a) to synthesize the final video. Since svd-img2vid-xt takes only an image as condition, the video generation process of this baseline lacks temporal motion as guidance. In comparison, our model can generate text-related motion tokens and thus generate more accurate video content following the prompt.

Effect of Token Length. We also explore the influence of different motion token lengths when encoding temporal motion information. The detailed results are reported in Table 7. It can be observed that a very small number suffice

Figure 5. Long video generation example with *"a 360 shot of a sleek yacht sailing gracefully through the crystal-clear waters of the Caribbean"*. The top two rows use the noise constraint in Equation (4) to improve temporal consistency, while the bottom row does not.

Table 6. Ablation of proposed motion tokenization strategy in zero-shot video understanding (left) and generation (right).

Method	MSVD	ActivityNet	UCF-101	
	Accuracy	Accuracy	IS ($\uparrow$)	FVD ($\downarrow$)
w/o motion	67.3	47.4	29.56	442.80
w/ motion	**73.2**	**50.1**	**44.26**	**280.57**

Table 7. Ablation of the number of motion tokens (denoted by N) in zero-shot video understanding (left) and generation (right).

Method	MSVD	ActivityNet	UCF-101	
	Accuracy	Accuracy	IS ($\uparrow$)	FVD ($\downarrow$)
$N = 256$	69.2	48.8	37.57	281.24
$N = 135$	**73.2**	**50.1**	**44.26**	**280.57**

to yield high understanding and generation performance. More token numbers may lead to representation redundancy and bring more duplicate motion token IDs when encoding videos without obvious motions, rendering the next-token prediction learning paradigm of LLM less effective. Using fewer motion tokens also allows for more video clips as input conditions under the same context length of LLM, which is useful for long video understanding.

5. Conclusion

This paper introduces Video-LaVIT, a multimodal generative pre-training method that empowers LLMs with unified comprehension and generation of videos, images, and language. At the core of our method is a video decomposition scheme that allows for more effective modeling of temporal information while reusing visual knowledge from image-only multimodal LLMs. The decomposed keyframes and motion vectors can be efficiently tokenized to be adapted to LLMs for unified generative pre-training. Finally, the understanding and generative capabilities of Video-LaVIT are verified by extensive quantitative and qualitative results.

Impact Statement

While this work advances the pre-training of large multimodal models in both performance and efficiency, its reasoning and generative capabilities should be treated care-fully. Some well-known problems include hallucination in multimodal understanding and exploitation to create misinformation through personalized generation. The model could also produce harmful responses due to inherent data bias and lack of alignment procedures.

One positive aspect we'd like to highlight is that, unlike many generation methods compared, this model is fully trained with *public* datasets. This allows the research community to correct for bias and harmful content in the training data. We hope that such a practice will help address important safety and alignment issues in multimodal models.

Acknowledgements

References

Aghajanyan, A., Huang, B., Ross, C., Karpukhin, V., Xu, H., Goyal, N., Okhonko, D., Joshi, M., Ghosh, G., Lewis, M., et al. CM3: A causal masked multimodal model of the internet. *arXiv preprint arXiv:2201.07520*, 2022.

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I.,

Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al. Flamingo: A visual language model for few-shot learning. In *NeurIPS*, pp. 23716–23736, 2022.

Awadalla, A., Gao, I., Gardner, J., Hessel, J., Hanafy, Y., Zhu, W., Marathe, K., Bitton, Y., Gadre, S., Sagawa, S., et al. OpenFlamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023.

Baek, Y., Lee, B., Han, D., Yun, S., and Lee, H. Character region awareness for text detection. In *CVPR*, pp. 9365–9374, 2019.

Bain, M., Nagrani, A., Varol, G., and Zisserman, A. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *ICCV*, pp. 1728–1738, 2021.

Beauchemin, S. S. and Barron, J. L. The computation of optical flow. *ACM Computing Surveys*, 27(3):433–466, 1995.

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

Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S. W., Fidler, S., and Kreis, K. Align your latents: High-resolution video synthesis with latent diffusion models. In *CVPR*, pp. 22563–22575, 2023b.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. In *NeurIPS*, pp. 1877–1901, 2020.

Carreira, J. and Zisserman, A. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pp. 6299–6308, 2017.

Changpinyo, S., Sharma, P., Ding, N., and Soricut, R. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, pp. 3558–3568, 2021.

Chen, D. and Dolan, W. B. Collecting highly parallel data for paraphrase evaluation. In *ACL*, pp. 190–200, 2011.

Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J. E., Stoica, I., and Xing, E. P. Vicuna: An open-source chatbot impressing GPT-4 with 90%* ChatGPT quality. https://lmsys.org/blog/2023-03-30-vicuna, 2023.

Dai, W., Li, J., Li, D., Tiong, A. M. H., Zhao, J., Wang, W., Li, B., Fung, P., and Hoi, S. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *NeurIPS*, 2023.

Dong, R., Han, C., Peng, Y., Qi, Z., Ge, Z., Yang, J., Zhao, L., Sun, J., Zhou, H., Wei, H., et al. DreamLLM: Synergistic multimodal comprehension and creation. In *ICLR*, 2024.

Esser, P., Rombach, R., and Ommer, B. Taming transformers for high-resolution image synthesis. In *CVPR*, pp. 12873–12883, 2021.

Esser, P., Chiu, J., Atighehchian, P., Granskog, J., and Germanidis, A. Structure and content-guided video synthesis with diffusion models. In *ICCV*, pp. 7346–7356, 2023.

Fang, Y., Wang, W., Xie, B., Sun, Q., Wu, L., Wang, X., Huang, T., Wang, X., and Cao, Y. EVA: Exploring the limits of masked visual representation learning at scale. In *CVPR*, pp. 19358–19369, 2023.

Feng, W., Zhu, W., Fu, T.-j., Jampani, V., Akula, A., He, X., Basu, S., Wang, X. E., and Wang, W. Y. LayoutGPT: Compositional visual planning and generation with large language models. In *ICLR*, 2024.

Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Qiu, Z., Lin, W., Yang, J., Zheng, X., et al. MME: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

Ge, S., Nah, S., Liu, G., Poon, T., Tao, A., Catanzaro, B., Jacobs, D., Huang, J.-B., Liu, M.-Y., and Balaji, Y. Preserve your own correlation: A noise prior for video diffusion models. In *ICCV*, pp. 22930–22941, 2023.

Gemini Team, G. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Girdhar, R., Singh, M., Brown, A., Duval, Q., Azadi, S., Rambhatla, S. S., Shah, A., Yin, X., Parikh, D., and Misra, I. Emu video: Factorizing text-to-video generation by explicit image conditioning. *arXiv preprint arXiv:2311.10709*, 2023.

Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., and Parikh, D. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *CVPR*, pp. 6904–6913, 2017.

Gurari, D., Li, Q., Stangl, A. J., Guo, A., Lin, C., Grauman, K., Luo, J., and Bigham, J. P. VizWiz grand challenge: Answering visual questions from blind people. In *CVPR*, pp. 3608–3617, 2018.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring massive multitask language understanding. In *ICLR*, 2021.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *NeurIPS*, pp. 6629–6640, 2017.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. In *NeurIPS*, pp. 6840–6851, 2020.

Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., and Fleet, D. J. Video diffusion models. In *NeurIPS*, pp. 8633–8646, 2022.

Hong, W., Ding, M., Zheng, W., Liu, X., and Tang, J. CogVideo: Large-scale pretraining for text-to-video generation via transformers. In *ICLR*, 2023.

Hudson, D. A. and Manning, C. D. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, pp. 6700–6709, 2019.

Jin, Y., Xu, K., Xu, K., Chen, L., Liao, C., Tan, J., Huang, Q., Chen, B., Lei, C., Liu, A., et al. Unified language-vision pretraining in LLM with dynamic discrete visual tokenization. In *ICLR*, 2024.

Karras, T., Aittala, M., Aila, T., and Laine, S. Elucidating the design space of diffusion-based generative models. In *NeurIPS*, pp. 26565–26577, 2022.

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al. The Kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

Kondratyuk, D., Yu, L., Gu, X., Lezama, J., Huang, J., Hornung, R., Adam, H., Akbari, H., Alon, Y., Birodkar, V., et al. VideoPoet: A large language model for zero-shot video generation. *arXiv preprint arXiv:2312.14125*, 2023.

Kynkäänniemi, T., Karras, T., Aittala, M., Aila, T., and Lehtinen, J. The role of ImageNet classes in Fréchet Inception distance. In *ICLR*, 2023.

Le Gall, D. MPEG: A video compression standard for multimedia applications. *Communications of the ACM*, 34(4):46–58, 1991.

Li, B., Wang, R., Wang, G., Ge, Y., Ge, Y., and Shan, Y. SEED-Bench: Benchmarking multimodal LLMs with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023a.

Li, D., Li, J., and Hoi, S. BLIP-Diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. In *NeurIPS*, 2023b.

Li, J., Li, D., Xiong, C., and Hoi, S. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, pp. 12888–12900, 2022.

Li, J., Li, D., Savarese, S., and Hoi, S. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, pp. 19730–19742, 2023c.

Li, K., He, Y., Wang, Y., Li, Y., Wang, W., Luo, P., Wang, Y., Wang, L., and Qiao, Y. VideoChat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023d.

Li, K., Wang, Y., He, Y., Li, Y., Wang, Y., Liu, Y., Wang, Z., Xu, J., Chen, G., Luo, P., et al. MVBench: A comprehensive multi-modal video understanding benchmark. *arXiv preprint arXiv:2311.17005*, 2023e.

Li, Y., Wang, C., and Jia, J. LLaMA-VID: An image is worth 2 tokens in large language models. *arXiv preprint arXiv:2311.17043*, 2023f.

Lian, L., Li, B., Yala, A., and Darrell, T. LLM-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models. *arXiv preprint arXiv:2305.13655*, 2023.

Lin, B., Zhu, B., Ye, Y., Ning, M., Jin, P., and Yuan, L. Video-LLaVA: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft COCO: Common objects in context. In *ECCV*, pp. 740–755, 2014.

Liu, H., Li, C., Li, Y., and Lee, Y. J. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023a.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. In *NeurIPS*, 2023b.

Liu, S., Cheng, H., Liu, H., Zhang, H., Li, F., Ren, T., Zou, X., Yang, J., Su, H., Zhu, J., et al. LLaVA-Plus: Learning to use tools for creating multimodal agents. *arXiv preprint arXiv:2311.05437*, 2023c.

Liu, Y., Cun, X., Liu, X., Wang, X., Zhang, Y., Chen, H., Liu, Y., Zeng, T., Chan, R., and Shan, Y. Evalcrafter: Benchmarking and evaluating large video generation models. *arXiv preprint arXiv:2310.11440*, 2023d.

Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., et al. MMBench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023e.

Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.-W., Zhu, S.-C., Tafjord, O., Clark, P., and Kalyan, A. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *NeurIPS*, pp. 2507–2521, 2022.

Maaz, M., Rasheed, H., Khan, S., and Khan, F. S. Video-ChatGPT: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023.

Mangalam, K., Akshulakov, R., and Malik, J. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36, 2024.

OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023a.

OpenAI. GPT-4V(ision) system card. https://openai.com/research/gpt-4v-system-card, 2023b.

Ordonez, V., Kulkarni, G., and Berg, T. L. Im2Text: Describing images using 1 million captioned photographs. In *NeurIPS*, pp. 1143–1151, 2011.

Patraucean, V., Smaira, L., Gupta, A., Recasens, A., Markeeva, L., Banarse, D., Koppula, S., Malinowski, M., Yang, Y., Doersch, C., et al. Perception test: A diagnostic benchmark for multimodal video models. *Advances in Neural Information Processing Systems*, 36, 2024.

Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., and Rombach, R. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *ICLR*, 2024.

Qiu, H., Xia, M., Zhang, Y., He, Y., Wang, X., Shan, Y., and Liu, Z. Freenoise: Tuning-free longer video diffusion via noise rescheduling. *arXiv preprint arXiv:2310.15169*, 2023.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *ICML*, pp. 8748–8763, 2021.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 21(1):5485–5551, 2020.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *CVPR*, pp. 10684–10695, 2022.

Runaway. Gen-2. https://research.runwayml.com/gen2, 2023.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, pp. 36479–36494, 2022.

Saito, M., Saito, S., Koyama, M., and Kobayashi, S. Train sparsely, generate densely: Memory-efficient unsupervised training of high-resolution temporal GAN. *IJCV*, 128(10-11):2586–2606, 2020.

Sharma, P., Ding, N., Goodman, S., and Soricut, R. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, pp. 2556–2565, 2018.

Shen, C., Gan, Y., Chen, C., Zhu, X., Cheng, L., and Wang, J. Decouple content and motion for conditional image-to-video generation. In *AAAI*, 2024.

Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., et al. Make-A-Video: Text-to-video generation without text-video data. In *ICLR*, 2023.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, pp. 2256–2265, 2015.

Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models. In *ICLR*, 2020.

Song, Y. and Ermon, S. Generative modeling by estimating gradients of the data distribution. In *NeurIPS*, pp. 11918–11930, 2019.

Soomro, K., Zamir, A. R., and Shah, M. A dataset of 101 human action classes from videos in the wild. *Center for Research in Computer Vision*, 2(11), 2012.

Sun, Q., Yu, Q., Cui, Y., Zhang, F., Zhang, X., Wang, Y., Gao, H., Liu, J., Huang, T., and Wang, X. Emu: Generative pretraining in multimodality. In *ICLR*, 2024.

Together Computer. RedPajama: an open dataset for training large language models, 2023. URL https://github.com/togethercomputer/RedPajama-Data.

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

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b.

Tran, D., Bourdev, L., Fergus, R., Torresani, L., and Paluri, M. Learning spatiotemporal features with 3D convolutional networks. In *ICCV*, pp. 4489–4497, 2015.

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

van den Oord, A., Vinyals, O., and Kavukcuoglu, K. Neural discrete representation learning. In *NeurIPS*, pp. 6309–6318, 2017.

Villegas, R., Babaeizadeh, M., Kindermans, P.-J., Moraldo, H., Zhang, H., Saffar, M. T., Castro, S., Kunze, J., and Erhan, D. Phenaki: Variable length video generation from open domain textual descriptions. In *ICLR*, 2023.

Wang, F.-Y., Chen, W., Song, G., Ye, H.-J., Liu, Y., and Li, H. Gen-L-Video: Multi-text to long video generation via temporal co-denoising. *arXiv preprint arXiv:2305.18264*, 2023a.

Wang, X., Yuan, H., Zhang, S., Chen, D., Wang, J., Zhang, Y., Shen, Y., Zhao, D., and Zhou, J. VideoComposer: Compositional video synthesis with motion controllability. In *NeurIPS*, 2023b.

Wang, Y., Li, K., Li, Y., He, Y., Huang, B., Zhao, Z., Zhang, H., Xu, J., Liu, Y., Wang, Z., et al. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022.

Wang, Y., He, Y., Li, Y., Li, K., Yu, J., Ma, X., Chen, X., Wang, Y., Luo, P., Liu, Z., et al. InternVid: A large-scale video-text dataset for multimodal understanding and generation. In *ICLR*, 2024.

Wu, C., Huang, L., Zhang, Q., Li, B., Ji, L., Yang, F., Sapiro, G., and Duan, N. GODIVA: Generating open-domain videos from natural descriptions. *arXiv preprint arXiv:2104.14806*, 2021.

Wu, C.-Y., Zaheer, M., Hu, H., Manmatha, R., Smola, A. J., and Krähenbühl, P. Compressed video action recognition. In *CVPR*, pp. 6026–6035, 2018.

Xu, J., Mei, T., Yao, T., and Rui, Y. MSR-VTT: A large video description dataset for bridging video and language. In *CVPR*, pp. 5288–5296, 2016.

Yan, W., Zhang, Y., Abbeel, P., and Srinivas, A. VideoGPT: Video generation using VQ-VAE and transformers. *arXiv preprint arXiv:2104.10157*, 2021.

Yang, A., Miech, A., Sivic, J., Laptev, I., and Schmid, C. Zero-shot video question answering via frozen bidirectional language models. In *NeurIPS*, pp. 124–141, 2022.

Ye, Q., Xu, H., Xu, G., Ye, J., Yan, M., Zhou, Y., Wang, J., Hu, A., Shi, P., Shi, Y., et al. mPLUG-Owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023.

Yu, J., Li, X., Koh, J. Y., Zhang, H., Pang, R., Qin, J., Ku, A., Xu, Y., Baldridge, J., and Wu, Y. Vector-quantized image modeling with improved VQGAN. In *ICLR*, 2022.

Yu, L., Shi, B., Pasunuru, R., Muller, B., Golovneva, O., Wang, T., Babu, A., Tang, B., Karrer, B., Sheynin, S., et al. Scaling autoregressive multi-modal models: Pretraining and instruction tuning. *arXiv preprint arXiv:2309.02591*, 2023a.

Yu, W., Yang, Z., Li, L., Wang, J., Lin, K., Liu, Z., Wang, X., and Wang, L. MM-Vet: Evaluating large multi-modal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023b.

Yu, Z., Xu, D., Yu, J., Yu, T., Zhao, Z., Zhuang, Y., and Tao, D. ActivityNet-QA: A dataset for understanding complex web videos via question answering. In *AAAI*, pp. 9127–9134, 2019.

Zeng, Y., Wei, G., Zheng, J., Zou, J., Wei, Y., Zhang, Y., and Li, H. Make pixels dance: High-dynamic video generation. *arXiv preprint arXiv:2311.10982*, 2023.

Zhang, B., Wang, L., Wang, Z., Qiao, Y., and Wang, H. Real-time action recognition with enhanced motion vector CNNs. In *CVPR*, pp. 2718–2726, 2016.

Zhang, H., Li, X., and Bing, L. Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023.

A. Experimental Settings

A.1. Model Implementation Details

Video Tokenizer We employ the off-the-shelf visual tokenizer from LaVIT (Jin et al., 2024) to transform the video keyframe into 90 visual tokens on average, which follows most existing MLLMs to utilize the ViT-G/14 of EVA-CLIP (Fang et al., 2023) as the visual encoder. The visual codebook size is set to 16384. Please refer to the original paper for more details. During training and inference, images and keyframes are resized to 224×224 resolution as input.

As for motion tokenization, we downsample the original videos at 6 fps and then take 24 consecutive frames as a video clip to compute the motion vector M. It is further divided by the width and height of the corresponding video to normalize the value within the range of $[-1, 1]$. Before feeding into the motion tokenizer, the motion vector M is resized to a resolution of 20×36, resulting in the final input tensor shape being $B \times 24 \times 20 \times 36 \times 2$. The encoder $f_{\mathcal{E}}$ and decoder $f_{\mathcal{D}}$ in our motion tokenizer both have $L = 12$ transformer blocks with 512 hidden states and 8 attention heads. Each block consists of spatial, temporal attention, and feed-forward layers. Before the attention computation, the motion input is reshaped into $[(BT) \times (HW) \times D]$ and $[(BHW) \times T \times D]$ for the spatial and temporal layers, respectively. We insert the spatial or temporal downsampling layers after the $[3, 6, 9, 12]$ encoder blocks to reduce the dimension of motion embeddings, which will then be quantized into 135 $(3 \times 9 \times 5)$ discrete motion tokens. The decoder $f_{\mathcal{D}}$ includes symmetrical upsampling layers to recover the original input motion vector during training. The size of learned motion codebook is set to 1024. To improve the training stability of the motion codebook, we leverage exponential moving average (EMA) updates with a weight of 0.995. Before quantization, the motion embeddings are projected into a low-dimensional space (dim=32) to improve the codebook usage, following the experience of Yu et al. (2022).

Video Detokenizer During training of the video detokenizer, we randomly sampled 24 consecutive frames from videos downsampled at 6 fps. The motion conditioning encoder has the same transformer architecture (12 blocks) as $f_{\mathcal{E}}$, except that the downsample layers are removed to keep the same temporal dimension with the input video frames. This strategy reduces the compression of motion information during encoding and provides explicit guidance for each frame to be denoised in the 3D U-Net. The detailed architecture of the 3D U-Net employed follows the same implementations as Blattmann et al. (2023b; 2023a). During the EDM-preconditioning optimization for the detokenizer, the distribution of $\log \sigma$ is set to $\mathcal{N}(1.0, 1.2^2)$ to encourage a higher noise level, which is found effective for the high-resolution generation (Girdhar et al., 2023). We train the motion conditioning encoder, the input encoding layer, and all the cross-attention layers in the 3D U-Net from scratch and initialize the other weights from the SVD img2vid-xt (Blattmann et al., 2023a). To reduce the computational complexity, the detokenizer is first trained with a resolution 384×384 for 50k steps, and then further fine-tuned at two types of resolutions: 768×768 or 1024×576 for another 10k steps.

Language Model We utilize Llama 2 7B (Touvron et al., 2023b) as the default large language model for the generative pre-training. The weight of the language model is initialized from LaVIT (Jin et al., 2024) to preserve the learned visual prior knowledge to support the comprehension and generation for the image domain. During pre-training, we mix the image-text, video-text pairs, and textual data in one batch to form the final multimodal input sequence.

A.2. Pre-training Data

The training dataset used by Video-LaVIT only consists of publicly available image and video datasets. In the following, we present a detailed elaboration of the dataset usage at each training stage.

Stage 1: The video tokenizer and detokenizer are trained on the WebVid-10M (Bain et al., 2021), which is an open-source video-text dataset containing 10 million video-text pairs scraped from the stock footage sites. Since both our tokenizer and detokenizer do not rely on textual data, we only employ pure video data at this stage. Due to the common watermarks in WebVid-10M, during the training of the video detokenizer, we incorporate a subset of InterVid-14M-aesthetics (Wang et al., 2024) to remove watermarks in the generated videos. It has also been shown useful in PixelDance (Zeng et al., 2023). Specifically, we first select a subset of 4s–10s video clips with the highest aesthetic scores and then follow SVD (Blattmann et al., 2023a) in applying CRAFT (Baek et al., 2019) to filter out those videos with unwanted written text. The result contains about 300k publicly available video clips. **Noting that the 300k video subset is only used during the training of the video detokenizer to improve the aesthetics of the generated videos, the results reported in all the experiments are tested on the checkpoint that uses only the WebVid-10M dataset.**

Stage 2: The language model is pre-trained on a mixture of video, image and text data, including WebVid-10M (Bain et al., 2021); 93M samples from Conceptual Caption (Sharma et al., 2018; Changpinyo et al., 2021), SBU (Ordonez et al.,

2011), and BLIP-Capfilt (Li et al., 2022). Moreover, we also employ the English text corpus from RedPajama (Together Computer, 2023), which is open-source data like the original one to train LLaMA from scratch. The purpose of including the English text corpus during pre-training is to preserve the already learned language understanding ability of LLM (e.g., the performance on linguistic benchmarks like MMLU (Hendrycks et al., 2021)) while acquiring good multimodal capabilities.

Stage 3: For a fair comparison, we employ the same instruction tuning dataset as the existing works (Lin et al., 2023; Li et al., 2023f) during this stage. It includes a 665k image-text instruction dataset from LLaVA v1.5 (Liu et al., 2023a) and a 100k video-text instruction dataset from Video-ChatGPT (Maaz et al., 2023). All the understanding results are tested by the model trained after Stage 3.

A.3. Training Settings

The detailed training hyper-parameter settings for the video tokenizer, detokenizer, and language model in Video-LaVIT are reported in Table 8. We adopt the same instruction tuning setting as LLaVA v1.5 (Liu et al., 2023a).

Configuration	Language Model	Tokenizer	Detokenizer
LLM init	LaVIT-7B	-	-
Optimizer	AdamW	AdamW	AdamW
Optimizer Hyperparameters	$\beta_1 = 0.9, \beta_2 = 0.95, \epsilon = 1e^{-6}$	$\beta_1 = 0.9, \beta_2 = 0.99, \epsilon = 1e^{-6}$	
Global batch size	2048	512	128
Peak learning rate of LLM	2e-5	-	-
Peak learning rate of other Part	5e-5	1e-4	5e-5
Learning rate schedule	Cosine	Cosine	Cosine
Training Steps	30K	100K	60K
Warm-up steps	2k	5K	3K
Weight decay	0.1	0.001	0.001
Gradient clipping	1.0	1.0	1.0
Input sequence to LLM	2048	-	-
Numerical precision	bfloat16	bfloat16	bfloat16
GPU Usage	128 NVIDIA A100	64 NVIDIA A100	64 NVIDIA A100
Framework	Megatron	DeepSpeed	DeepSpeed
Training Time	60h	10h	48h

Table 8. The detailed training hyperparameters of Video-LaVIT

A.4. Evaluation

Image Understanding is evaluated using eight popular image question answering and multimodal benchmarks: VQA v2 (Goyal et al., 2017), GQA (Hudson & Manning, 2019), VizWiz (Gurari et al., 2018), ScienceQA-IMG (Lu et al., 2022), MME (Fu et al., 2023), MMBench (Liu et al., 2023e), SEED (Li et al., 2023a), MM-Vet (Yu et al., 2023b). For question-answering datasets, we use the same prompts as in LLaVA-1.5 (Liu et al., 2023a), and adopt the widely used VQA accuracy as the evaluation metric.

Video Question Answering. Three common datasets are considered: MSVD-QA (Chen & Dolan, 2011), MSRVTT-QA (Xu et al., 2016) and ActivityNet-QA (Yu et al., 2019). To assess model accuracy, a GPT-3.5 assistant (Maaz et al., 2023) is employed, which also produces outputs a relative score ranging from 0 to 5.

Text-to-Image Generation. We adopt the validation set of MS-COCO (Lin et al., 2014) and randomly select 30K samples. The quality of the generated images is evaluated by Fréchet Inception distance (FID) (Heusel et al., 2017), which computes its Fréchet distance to the ground truth in the feature space of a pre-trained Inception V3 model.

Text-to-Video Generation is measured on MSR-VTT (Xu et al., 2016) and UCF-101 (Soomro et al., 2012). For MSR-VTT, we use all 2990 videos and sample one caption for each video, resulting in 2990 video-text pairs; for UCF-101, we sample 20 videos per class and follow PYoCo (Ge et al., 2023) to curate prompts for each class, producing 2020 video-text pairs. Their evaluation metrics are detailed below.

- CLIP similarity (CLIPSIM) (Wu et al., 2021) measures the semantic similarity between video-text pairs. We follow Phenaki (Villegas et al., 2023) and VideoPoet (Kondratyuk et al., 2023) in using a ViT-B/16 (Radford et al., 2021) to compute CLIP scores between 224×224 sized video frames and their corresponding captions. The final score is

averaged over all generated video frames.

- Fréchet video distance (FVD) (Unterthiner et al., 2018) evaluates the Fréchet distance between generated and real videos in the feature space of an I3D action classification model (Carreira & Zisserman, 2017) pre-trained on Kinetics-400 (Kay et al., 2017).

- Fréchet Inception distance (FID) (Heusel et al., 2017) measures the Fréchet distance between generated and real video frames. Following PYoCo (Ge et al., 2023), we use a ViT-B/32 model (Kynkäänniemi et al., 2023) to extract the frame features. The final result is averaged over all video frames.

- Inception score (IS) (Saito et al., 2020) evaluates the distribution of our generated video frames. We employ a C3D model (Tran et al., 2015) fine-tuned on UCF-101 to calculate a video version of the inception score. The model takes the central 16 frames of each video as the input.

Note that there are slight variations in the evaluation protocols of different papers. We have sought to keep our protocol the same as or similar to most of the top-ranked methods.

B. Additional Results

B.1. Multimodal Generation

This section provides additional qualitative results and an ablation study to demonstrate the effectiveness of our design for multimodal generation, complementing the existing comparisons in the main paper.

Text-to-Image Generation. Figure 6 illustrates the comparison of text-to-image generation between Video-LaVIT and SDXL (Podell et al., 2024). Overall, our method achieves competitive visual quality while having better language understanding and reasoning capabilities. For example, in the top-left case of a young woman in front of a UFO, our method produces highly aesthetic headshots of the woman, while capturing the detail of "sharp focus" in the text prompt. And in the bottom-left example of apple painting, our model successfully infers from the prompt "neither is red and both are green" to draw two green apples, thanks to the better logical reasoning ability of the LLM-based generation approach we adopted.

Text-to-Video Generation. Figure 7 compares Video-LaVIT to a closed-source model Gen-2 (Runaway, 2023). As can be seen, our model produces high-quality videos that are generally comparable to Gen-2, which is especially evident in the last two examples where it successfully captures details such as "moss and many flowers" and "autumn" in the text prompt and yields very similar results to Gen-2. Moreover, the first two comparisons demonstrate a favorable prompt following ability of our model. In the first case with the keyword "running", our model produces significant camera motion toward the cabin, while the movement in Gen-2 is relatively nuanced. In the second case, our model correctly displays multiple "pirate ships" as the prompt specified, with artistic details such as all the ships being on fire, according to the implication of "intense battle". These results support the benefits of unified video-language pre-training in prompt following capabilities.

Image-to-video generation. Figure 8 presents a comparison of Video-LaVIT with the open-source model SVD (Blattmann et al., 2023a), both conditioned on synthetic image prompts. Moving to some unseen test cases, our method produces video clips featuring both natural and refined motions, thanks to the decomposed video representation that can better transfer motion-related knowledge to new visual inputs. For example, in the middle case, our generated goat smoothly lowers its head and blinks as if it were a human to think, while the goat in the video produced by SVD hardly moved. In the bottom case, where the image prompt shows a teddy is riding a motorcycle, our generated full video looks very natural and similar to a human riding a motorcycle, while SVD constantly produces a scenario where the motorcycle is moving a different direction from where its tire is pointing (which is physically wrong). Overall, our model demonstrates superior image-to-video generation performance with the inclusion of decoupled visual-motion tokenization and LLM pre-training.

Long Video Generation is showcased in Figure 9. By explicitly constraining the noise when decoding successive video clips, our model can provide a high temporal consistency during long video generation. For example, in the first two cases, the dog and the jeep car maintain the same identity across different clips with highly coherent visual details. In the last example which features large camera movement, the moving trajectory remains consistent as it approaches the cabin according to the text prompt. These examples all illustrate our reasonably good quality of long video generation. Note that all the generated videos are provided at `https://video-lavit.github.io`.

The Effect of Enhanced Motion Conditioning. To rigorously reconstruct original video content, we employ the enhanced

Table 9. The impact of incorporating motion tokens on image comprehension.

Method	VQAv2	GQA	VizWiz	SQA	MME	MMB	SEED	MM-Vet
w/o motion	80.0	63.7	54.4	**71.5**	1533.2	**67.5**	**64.7**	**34.5**
w motion	**80.3**	**64.4**	**56.0**	70.0	**1551.8**	67.3	64.0	33.2

Table 10. The impact of svd-img2vid-xt weight initialization on text-to-video generation.

Method	MSR-VTT			UCF-101	
	CLIPSIM ($\uparrow$)	FVD ($\downarrow$)	FID ($\downarrow$)	IS ($\uparrow$)	FVD ($\downarrow$)
w/ svd-img2vid-xt	0.3010	**169.51**	11.80	37.96	**274.96**
w/o svd-img2vid-xt	**0.3012**	188.36	**11.27**	**44.26**	280.57

conditioning: motion input condition and motion feature condition for training the 3D video U-Net g_V. We illustrate the effect of proposed enhanced motion conditioning (EMC) strategy on video decoding in Figure 10. The variant "w/o EMC" only leverages motion vectors as the input condition. Compared with using EMC, it is incapable of recovering the motion of original input videos. For example, the "train" and "fish" barely moved in the shown video samples, which demonstrates the effectiveness of our proposed conditioning strategy.

B.2. Multimodal Understanding

This section presents qualitative results of Video-LaVIT for image and video understanding. First, Table 11 showcases our performance in image question answering using the famous test example from GPT-4 (OpenAI, 2023a). As can be seen, our model produces a reasonable answer with a good number of correct details (e.g. the type of the vehicle being SUV) and even a friendly safety warning. In comparison, one of the recent multimodal LLMs, LLaVA (Liu et al., 2023b), produces a roughly correct answer with some inaccurate detail (mistaking the vehicle type as "minivan or van").

For video question answering, Tables 12 and 13 compares our method to Video-LLaVA (Lin et al., 2023) and Video-ChatGPT (Maaz et al., 2023) based on the video clips from Video-ChatGPT. In the first example of Table 12 which asks to explain why a video is funny, our model yields the most concise answer among the video-language models compared, and at the same time contains a salient point that the other models failed to mention. The next example in Table 12, on the other hand, shows that our method produces fewer hallucinations than Video-LLaVA and Video-ChatGPT, as the latter two models tend to generate overly detailed action descriptions that have no basis in the video. And lastly, in the example of Table 13, our model follows the instruction prompt by producing a beautiful fairy tale with both conciseness and a moral lesson ("love can conquer all"). To summarize, our method demonstrates reasonably good multimodal understanding capabilities across different test cases, in line with the previous quantitative comparison on multiple benchmarks.

B.3. Ablation Studies

Impact of Motion Tokens on Image Comprehension. Video-LaVIT indiscriminately treat all the modalities (video, image, and text) as 1D discrete tokens fed into LLMs. The impact of incorporating motion tokens on image comprehension is reported in Table 9. As observed, including motion tokens hardly affects the understanding performance of the image, which demonstrated the effectiveness of the proposed decoupled visual-motional tokenization. Video-LaVIT is capable of modeling video, image, and text data in a unified framework.

Impact of Weight Initialization. We re-trained the detokenizer of Video-LaVIT from scratch without svd-img2vid-xt initialization on the WebVid-10M dataset and found that our model can still achieve comparable video generation performance. The detailed text-to-video generation results are reported in Table 10. As observed, training the detokenizer from scratch had little impact on the final results.

C. Limitations

Our proposed model cannot generate very long videos due to its limited context window (4096) and dataset restriction. This work only used the public WebVid-10M as the video pre-training data. In WebVid, the video durations are relatively short (about 15s on average) and the video scenes barely change, which results in our model generating similar keyframes in different clips. On the other hand, a general concern is that our training cost is still too high to scale to web-scale video data, which may require further optimization through joint exploitation of spatial and temporal redundancies in video.
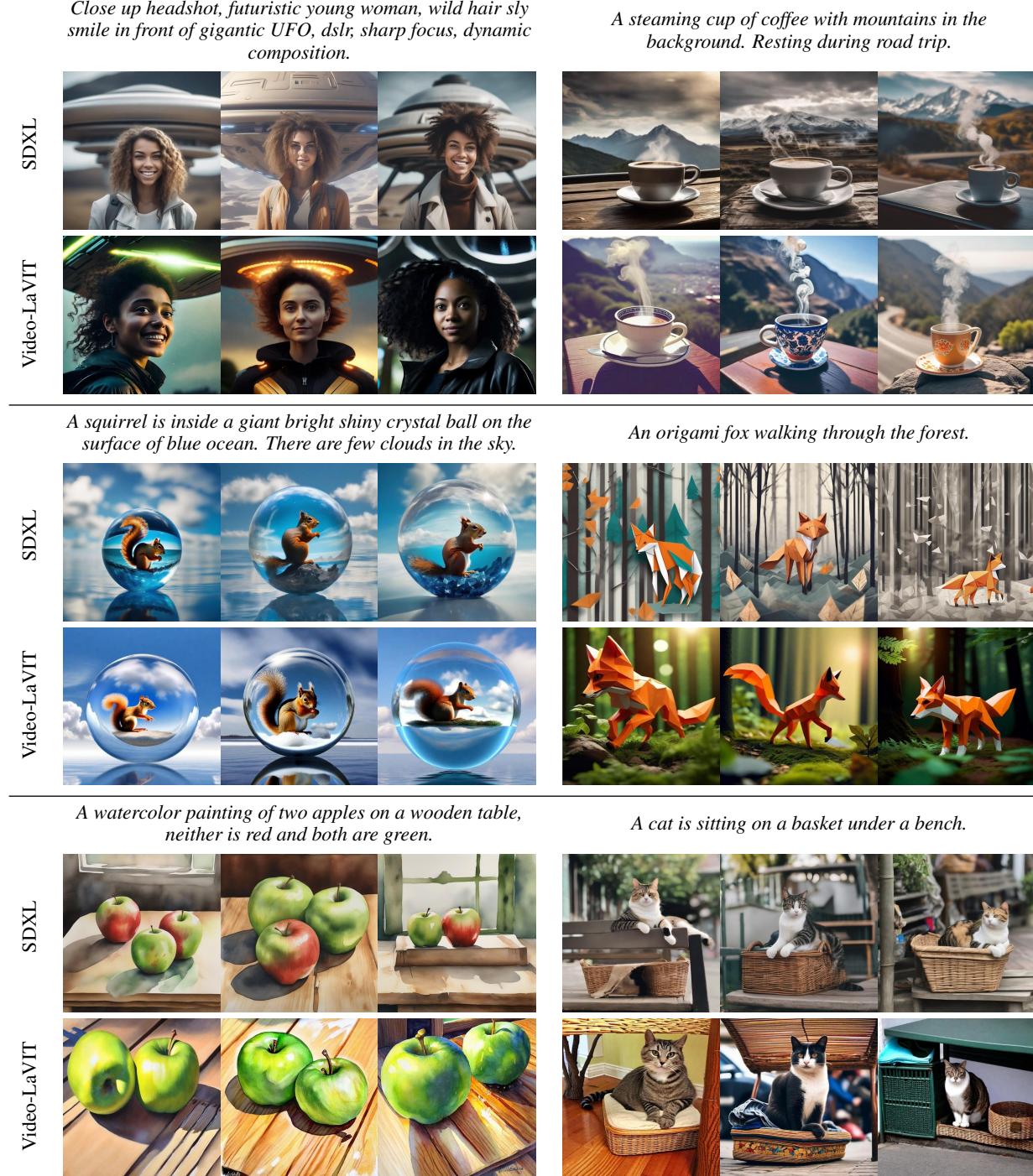
Figure 6. Text-to-image generation comparison with SDXL (Podell et al., 2024). Prompts are from SDXL, CM3Leon (Aghajanyan et al., 2022), Imagen (Saharia et al., 2022), VideoPoet (Kondratyuk et al., 2023), LMD (Lian et al., 2023), and LayoutGPT (Feng et al., 2024). Our model provides comparable visual quality while showing better logical and spatial reasoning abilities (see the last two cases).
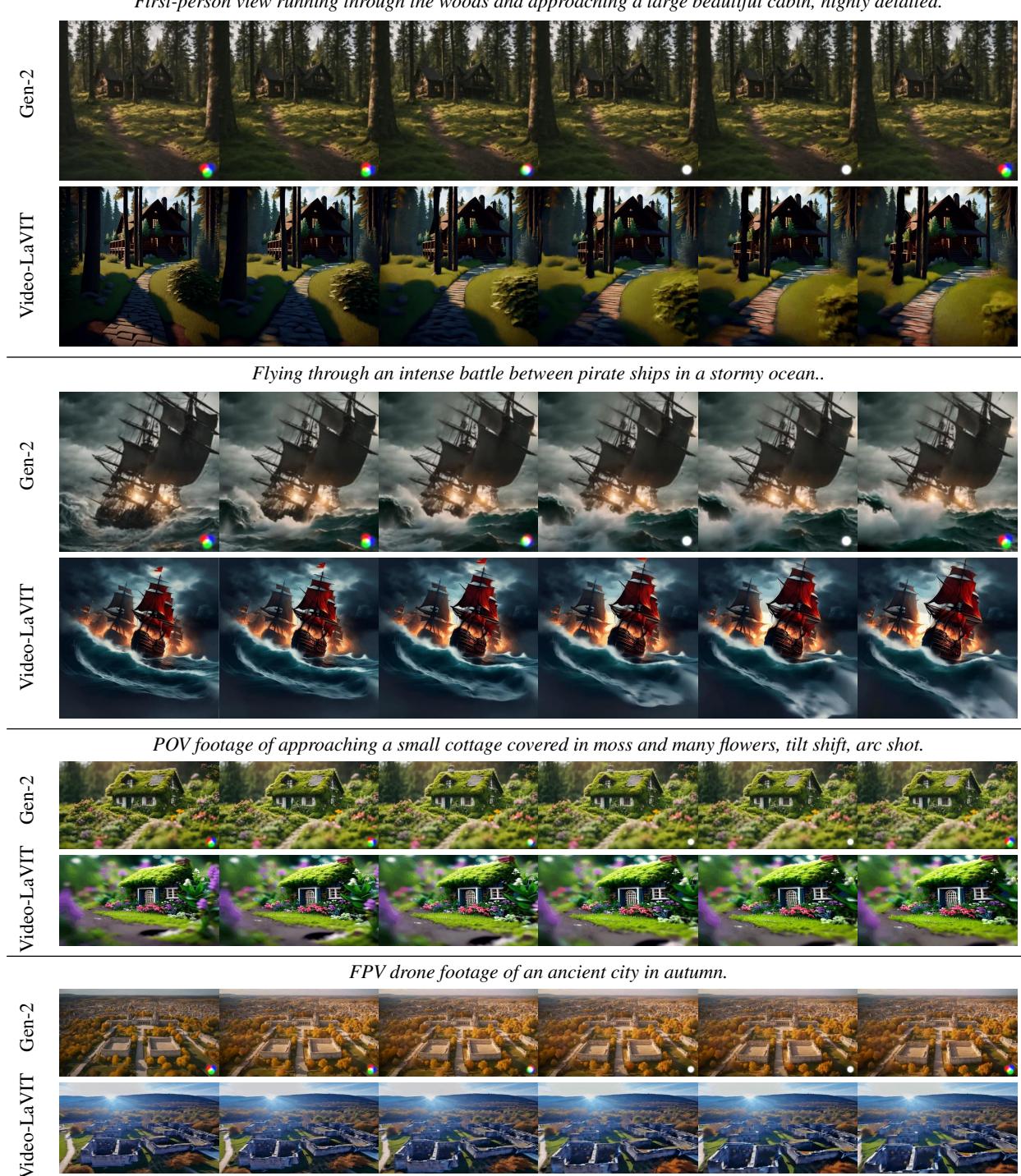
Figure 7. Text-to-video generation comparison with Gen-2 (Runaway, 2023) using default parameters. Prompts are from VideoPoet (Kondratyuk et al., 2023) and PixelDance (Zeng et al., 2023). Our model provides a similarly high visual quality (in the bottom two cases) while following the text prompt better (including "running" in the first example and "pirate ships" in the second examples).
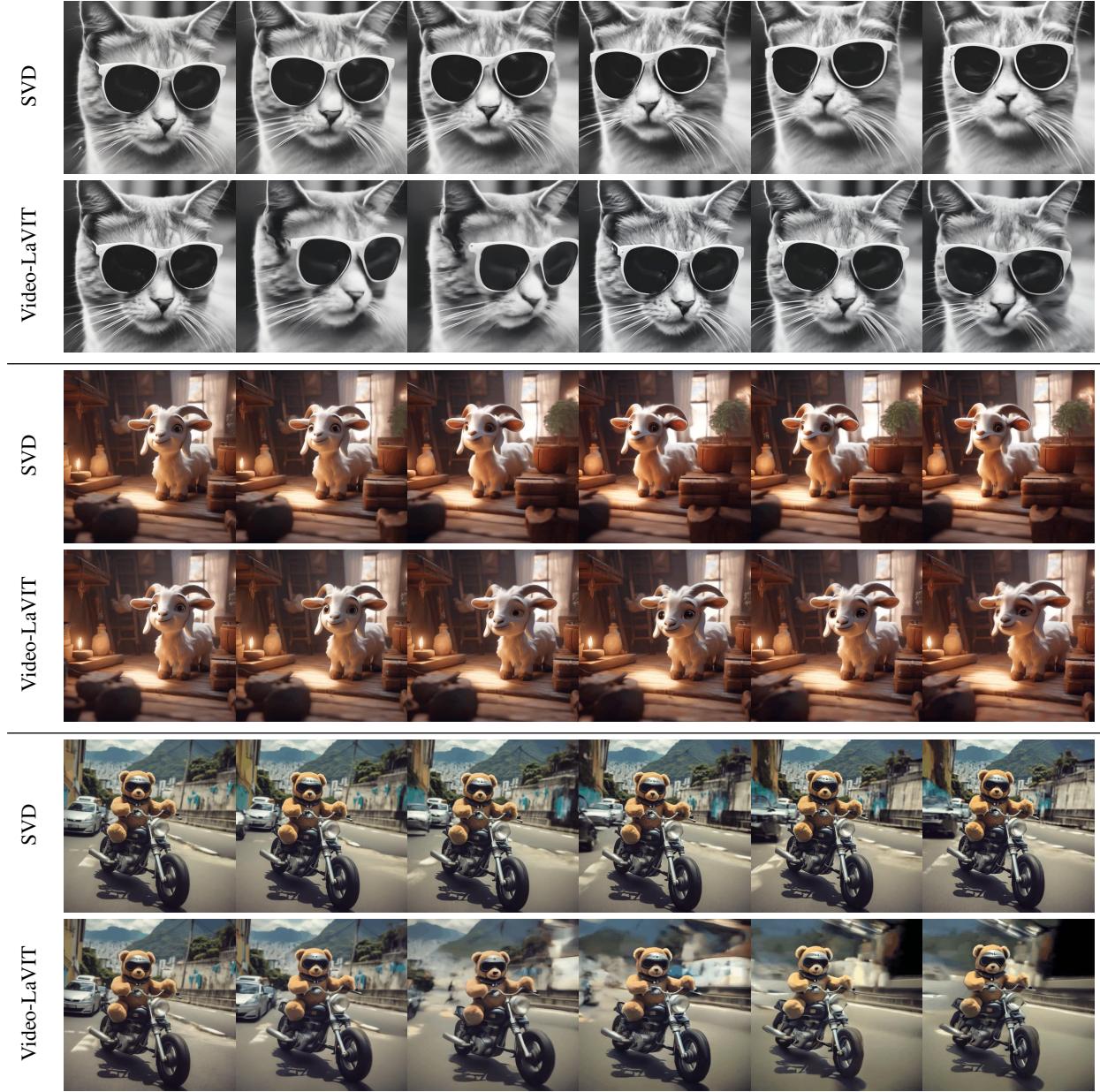
Figure 8. Image-to-video generation comparison with SVD (Blattmann et al., 2023a) using the stable-video-diffusion-img2vid-xt version. The generation is conditioned on the leftmost frame. Our model can produce more sophisticated animal motions (see the top two cases) while not violating the physical rules (e.g., in the second last row, the motorcycle is not moving in the direction of its tire).

A dog in the sun.

A jeep car is moving on the beach.

First-person view running through the woods and approaching a large beautiful cabin, highly detailed.
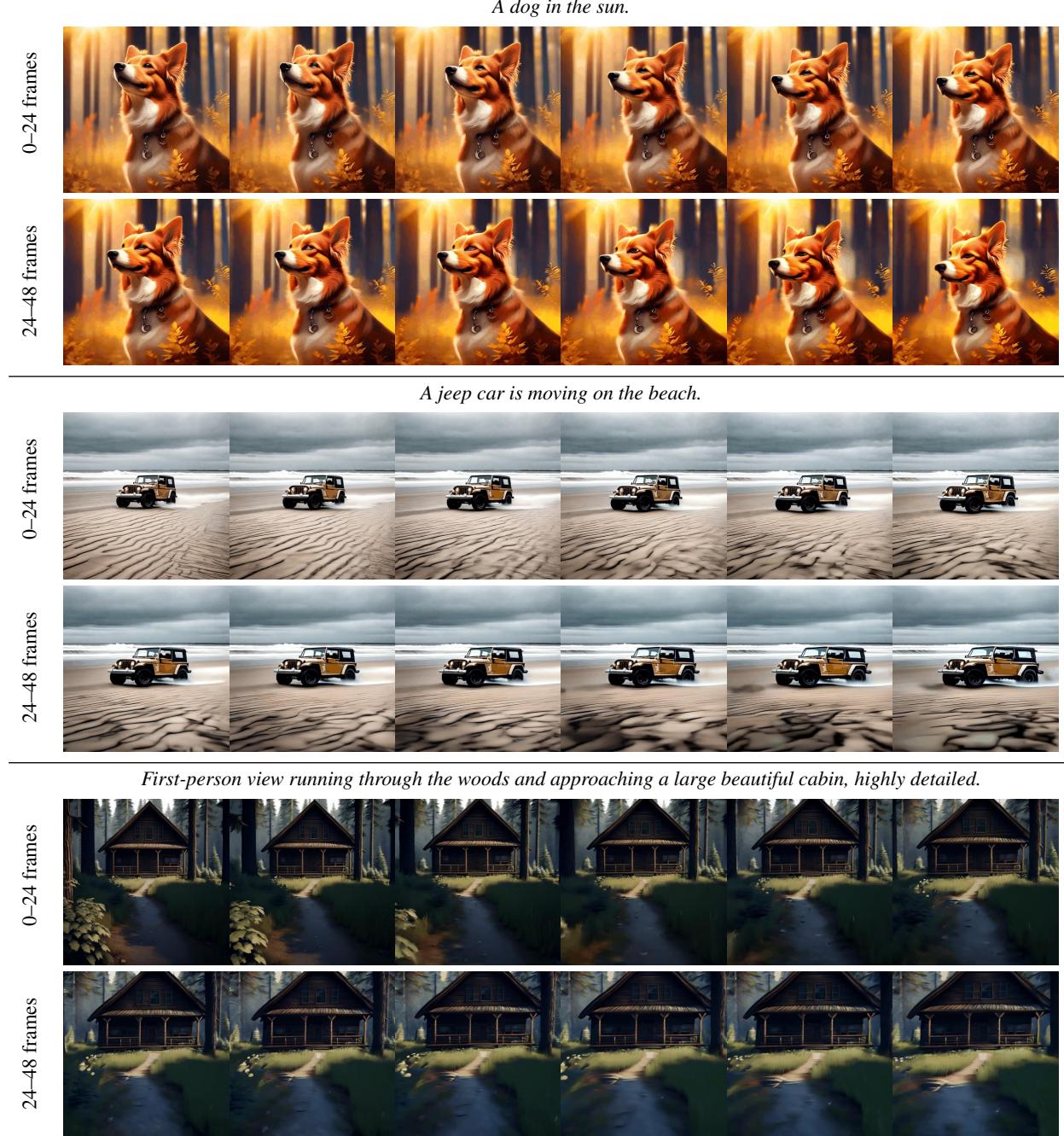
Figure 9. Long video generation examples. Prompts are Gen-L-Video (Wang et al., 2023a) and VideoPoet (Kondratyuk et al., 2023). Our generated videos are temporally coherent even across different decoded clips, thanks to our proposed explicit noise constraint.
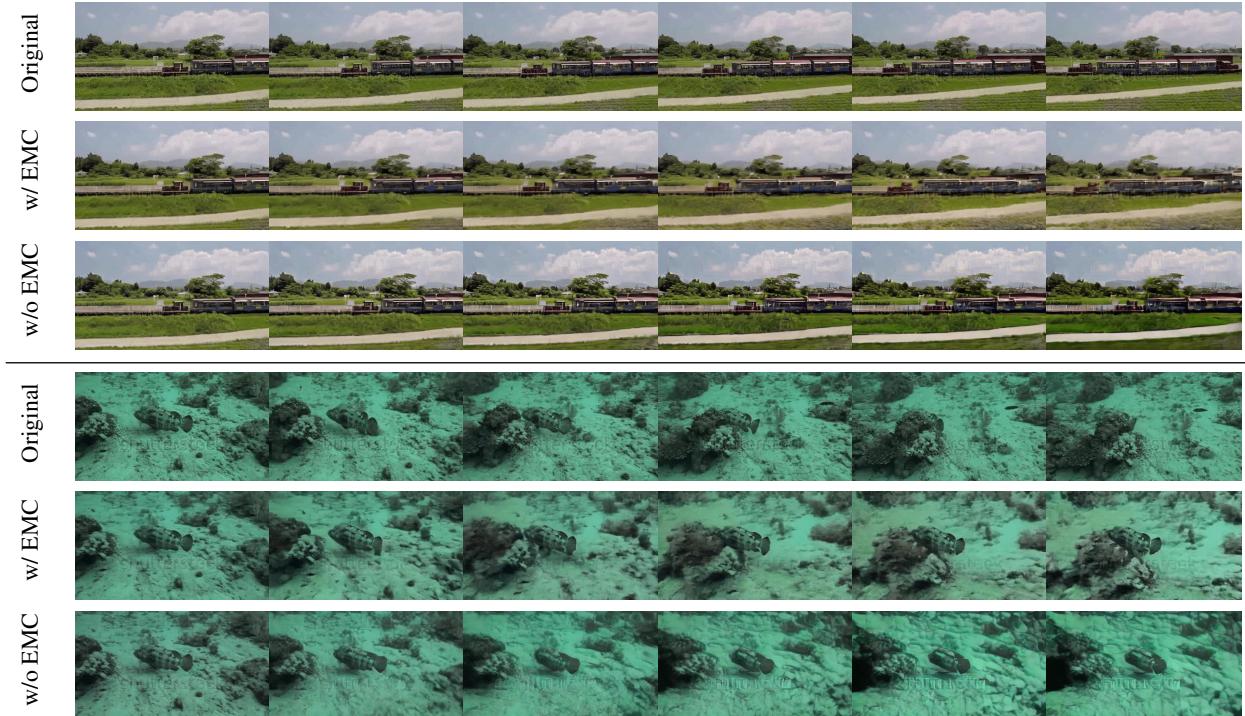
Figure 10. Ablation study of enhanced motion conditioning (EMC) for video reconstruction. The last row (i.e., w/o EMC) indicates only motion vectors as the input condition for training 3D U-Net. As illustrated, incorporating EMC can improve the fidelity of the reconstructed videos. The video samples are taken from WebVid-10M (Bain et al., 2021).

Table 11. Image question answering comparison with LLaVA (Liu et al., 2023b), GPT-4 (OpenAI, 2023a), BLIP-2 (Li et al., 2023c), and OpenFlamingo (Awadalla et al., 2023). Our model answers correctly like the other multimodal LLMs, but without missing details such as the vehicle type (it is an SUV instead of a van or minivan). The table format follows GPT-4.

Visual input example, Extreme Ironing:

Source: https://www.barnorama.com/wp-content/uploads/2016/12/03-Confusing-Pictures.jpg

User	What is unusual about this image?
Video-LaVIT	The unusual aspect of this image is that a man is ironing clothes on an ironing board placed on the back of a yellow SUV while it is driving down the street. This is not a typical scenario, as ironing clothes is usually done indoors, in a home or a laundry room, and not while the vehicle is in motion. The man's actions pose a risk to himself and others on the road, as ironing on a moving vehicle can lead to accidents or distractions for the driver.
	[*Start a new conversation, and clear the history*]
User	What's happening in the scene?
Video-LaVIT	The image depicts a man ironing clothes on an ironing board placed on the back of a yellow SUV. The man is standing on the back of the vehicle, attentively ironing his clothes. The ironing board is positioned in the middle of the SUV, with the man standing on it to reach the clothes. In the background, there are two cars visible, one on the left side and another on the right side of the scene. The man appears to be focused on his task, despite the unusual location for ironing clothes.

User	What is unusual about this image?
LLaVA (Liu et al., 2023b)	The unusual aspect of this image is a man ironing clothes on the back of a minivan or van. This is not a typical place to perform this activity, as one would usually iron clothes in a more stationary and safe location, such as a home, using a regular ironing board. The scene depicted in the image is peculiar as it involves a makeshift ironing setup on a vehicle, which can be both unsafe and unconventional. Additionally, it is not clear how the man is able to maintain balance and stability while ironing clothes in such an unstable environment.
	[*Start a new conversation, and clear the history*]
User	What's happening in the scene?
LLaVA (Liu et al., 2023b)	The image depicts a man standing on top of a yellow SUV in a busy city street. He is holding a portable ladder, seemingly ironing clothes while standing on the vehicle. Around the scene, there are other cars, a traffic light, and a fire hydrant. Additionally, there are two other people in the vicinity, one close to the man on the car and another further away. The man on the ladder appears to be attracting attention due to his unconventional choice of ironing his clothes on top of a moving car. The city street around him is bustling with activity, adding to the unique nature of the scene.

User	What is unusual about this image?
GPT-4 (OpenAI, 2023a)	The unusual thing about this image is that a man is ironing clothes on an ironing board attached to the roof of a moving taxi.

User	Can you explain this meme in detail?
BLIP-2 (Li et al., 2023c)	a man is sitting on the back of a yellow cab
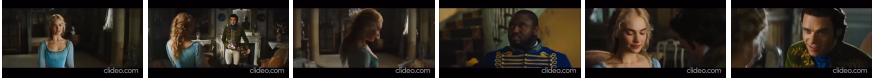
User	Can you explain this meme in detail?
OpenFlamingo (Awadalla et al., 2023)	The man is drying his clothes on the hood of his car.

Table 12. Video question answering comparison with Video-LLaVA (Lin et al., 2023) and Video-ChatGPT (Maaz et al., 2023). Video examples are from Video-ChatGPT. In the first example, our model directly points to a salient explanation for why the video is funny. In the second example, our model responses accurately without hallucinating about the video (the woman only swung the ax once).

User Video-LaVIT	Why is this video funny? The video is funny because the baby is wearing glasses and trying to read a book, but she is not able to do so. The baby's attempts to read the book are adorable and amusing to watch.
User Video-LLaVA (Lin et al., 2023)	Why is this video funny? The video is funny because the baby is trying to read a book while wearing glasses, which is a humorous sight. The baby's attempts to read the book are also amusing, as they seem to be struggling with the task. Additionally, the baby's facial expressions and movements add to the overall humor of the video.
User Video-ChatGPT (Maaz et al., 2023)	Why is this video funny? The video is funny because it shows a baby girl who is learning to read. She is sitting on a bed and looking at a book, and she is trying to read it. The baby girl is wearing glasses, and she is holding the book with both hands. The baby girl is trying to read the book, and her actions are adorable and endearing. The video captures the innocence and curiosity of a young child as she learns to read, which is a joyful and amusing sight for the viewer.

User Video-LaVIT	What does the woman use to split the logs and how does she do it? The woman uses an ax to split the logs. She picks up the ax and swings it down with force, splitting the logs in half.
User Video-LLaVA (Lin et al., 2023)	What does the woman use to split the logs and how does she do it? The woman uses an ax to split the logs. She holds the ax in her hand and swings it down onto the log. She repeats this process several times, splitting the log into smaller pieces.
User Video-ChatGPT (Maaz et al., 2023)	What does the woman use to split the logs and how does she do it? The woman uses a large axe to split the logs. She starts by placing the axes on the log and then swings it down with a powerful motion. The axe slices through the log with ease, and the woman continues to swing it until the log is split in half.

24

Table 13. Video question answering comparison with Video-LLaVA (Lin et al., 2023) and Video-ChatGPT (Maaz et al., 2023). Video examples are from Video-ChatGPT. As shown in this example, our model captures the underlying context of this video, which is a love story. The fairy tale told, which implies that love will prevail, is naturally inspired by the video without introducing an unmentioned third party (jealous sister or stepmother) as in the other models. The use of the phrase "beautiful glass slipper" adds a nice detail to the story.

User Video-LaVIT	Create a short fairy tale with a moral lesson inspired by the video. Once upon a time, there was a beautiful princess who lived in a grand castle. She was known for her kindness and compassion towards others. One day, a handsome prince visited her castle and fell deeply in love with her. However, the princess was hesitant to accept his love due to her fear of losing her independence and being trapped in a loveless marriage. The prince, determined to win her heart, gave her a beautiful glass slipper as a gift. The princess tried it on and danced with the prince, realizing that true love transcends all boundaries. From that day on, the princess and the prince lived happily ever after, proving that love can conquer all.
User Video-LLaVA (Lin et al., 2023)	Create a short fairy tale with a moral lesson inspired by the video. Once upon a time, in a kingdom far away, there lived a beautiful princess named Rose. She was known for her kindness and beauty, and everyone loved her. One day, a handsome prince named Leo came to visit the kingdom. He was fascinated by Rose's beauty and decided to stay in the kingdom. However, Rose's jealous sister, Lily, did not like the attention Leo was giving her sister. She tried to harm Rose, but Leo saved her and they fell in love. The kingdom celebrated their love and they lived happily ever after. The moral of the story is that true love conquers all and that jealousy can lead to harmful consequences.
User Video-ChatGPT (Maaz et al., 2023)	Create a short fairy tale with a moral lesson inspired by the video. Sure, here's a short fairy tale with a moral lesson inspired by the video: Once upon a time, there was a beautiful princess named Cinderella who lived in a castle with her cruel stepmother and two stepsisters. She was known for her kindness and her love for animals, especially birds. One day, while she was out in the garden, she met a handsome prince who was lost in the woods. He was in need of help, and Cinderella offered to guide him back to the castle. As they walked together, the prince noticed that Cinderella's stepsisters were cruel to her and her and stepmother was always mean to her. He asked her about her life and how she managed to stay so kind and gentle despite the hardships she faced. Cinderella told him that she found happiness in the simple things in life, such as the beauty of nature and the love of animals. The prince was touched by her story and decided to help her in any way he could. He introduced her to the king and queen, who were impressed by her kindness and grace. They arranged for her to attend a ball at the castle, where she met the prince once again. They fell in love and got married, living happily ever after. The moral of the story is that even in difficult situations, one should find happiness in the simple things in life and never lose sight of their values and kindness.