This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Divot: Diffusion Powers Video Tokenizer for Comprehension and Generation

Yuying Ge **Yizhuo Li** **Yixiao Ge** **Ying Shan**
ARC Lab, Tencent PCG
https://github.com/TencentARC/Divot

Abstract

*In recent years, there has been a significant surge of interest in unifying image comprehension and generation within Large Language Models (LLMs). This growing interest has prompted us to explore extending this unification to videos. The core challenge lies in developing a versatile video tokenizer that captures both the spatial characteristics and temporal dynamics of videos to obtain representations for LLMs, and the representations can be further decoded into realistic video clips to enable video generation. In this work, we introduce Divot, a **Di**ffusion-Powered **V**ideo **T**okenizer, which leverages the diffusion process for self-supervised video representation learning. We posit that if a video diffusion model can effectively de-noise video clips by taking the features of a video tokenizer as the condition, then the tokenizer has successfully captured robust spatial and temporal information. Additionally, the video diffusion model inherently functions as a de-tokenizer, decoding videos from their representations. Building upon the Divot tokenizer, we present Divot-LLM through video-to-text autoregression and text-to-video generation by modeling the distributions of continuous-valued Divot features with a Gaussian Mixture Model. Experimental results demonstrate that our diffusion-based video tokenizer, when integrated with a pre-trained LLM, achieves competitive performance across various video comprehension and generation benchmarks. The instruction tuned Divot-LLM also excels in video storytelling, generating interleaved narratives and corresponding videos. Models and codes are available at https://github.com/TencentARC/Divot.*

1. Introduction

In recent years, the rapid evolution of Multimodal Large Language Models (MLLMs) [12, 15–17, 55–57, 77, 86, 91, 93] has demonstrated significant progresses in unified image understanding and generation, which empowers LLMs [6, 11, 60] with the ability to generate images beyond texts. While these work primarily focus on image-text data, the extension of this unification to the video domain remains relatively under-explored. Achieving unified video comprehension and generation is essential for the development of more sophisticated artificial intelligence (AI) systems that are capable of understanding and creating dynamic visual content in the real world.

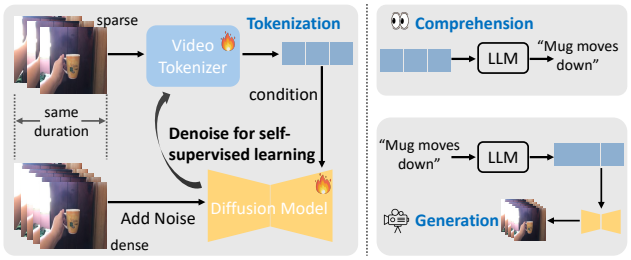


Figure 1. We utilize the **diffusion procedure** to learn a video tokenizer in a self-supervised manner for unified comprehension and generation, where the spatiotemporal representations serve as the condition of a diffusion model to de-noise video clips. Additionally, the proxy diffusion model functions as a de-tokenizer to decode realistic video clips from the video representations.

The primary challenge of achieving unified video comprehension and generation lies in developing a versatile video tokenizer that can effectively address the complexities inherent in video data. This tokenizer should be able to obtain robust video representations that serve as inputs of MLLMs for video comprehension, and these representations can be further decoded into realistic video clips to enable video generation. Unlike static images, videos encompass both spatial characteristics and temporal dynamics, making their representation significantly more complex. Recent pioneering work [26, 40, 74] adopt a discrete video tokenizer for unifying video comprehension and generation, where a video is represented as a sequence of discrete frame tokens, or keyframe tokens followed by discrete motion tokens. This approach eases video generation with a LLM through an autoregressive next-token prediction mechanism, but sacrifices the performance of multimodal understanding, as pointed out by recent work [77]. In this work, we aim to investigate an alternative approach by *utilizing continuous video representations* to unify video comprehension and generation.

To this end, we introduce Divot, a **Di**ffusion-Powered

Vide**o** **T**okenizer that leverages the diffusion process [50] for self-supervised video representation learning as shown in Fig. 1. The core premise is that if a diffusion model can effectively predict the noise added to the Variational Autoencoder (VAE) latents [28] of video clips, when conditioned on the features produced by the video tokenizer, it demonstrates that *the tokenizer has successfully captured robust spatial and temporal information* inherent in the video data. This capability is crucial for representing the intricate dynamics present in videos. Furthermore, in addition to being a proxy module for learning the tokenizer, the diffusion model can act as a de-tokenizer to effectively decode realistic videos from their learned representations. This dual functionality facilitates a seamless integration of understanding and creating video content within a LLM.

Specifically, the Divot tokenizer is composed of a pre-trained Vision Transformer (ViT) encoder [13], a Spatial-Temporal transformer, and a Perceiver Resampler [1] to obtain video representations from video frames sampled at low frame rate (fps) considering the semantic redundancy between adjacent frames. The video representations serve as the condition of a pre-trained video diffusion model, DynamiCrafter [78] (without the concatenation of a conditional image with initial noise), to predict the noise added to the VAE latents of video frames. After training, the video diffusion model can generate realistic video clips from noise by taking the video representations provided by the Divot tokenizer as the condition.

We further present Divot-LLM by equipping the pre-trained Mistral-7B[1] [24] with the Divot tokenizer. Divot-LLM is pre-trained with a next-word prediction objective on video-caption data by taking the spatiotemporal representations as inputs for video comprehension. The challenge then arises in *modeling the continuous video representations with the LLM* for video generation. We empirically find that simply minimizing the distance between the LLM output and video representations using mean squared error (MSE) loss achieves unsatisfactory results, since the deterministic regression regularizes the LLM to learn overly averaged features of videos. To address this, inspired by recent work [33], we shift our focus from deterministic modeling to **probabilistic modeling** by modeling the distributions of video features with a Gaussian Mixture Model (GMM). Specifically, we train the LLM to predict GMM parameters, including means, variances, and mixture probabilities through minimizing the discrepancy between the predicted GMM distribution and the actual video representations using negative log-likelihood (NLL) loss. During inference, we draw samples from the predicted GMM distribution as the condition of the video de-tokenizer to decode video clips.

[1]We do not explore more advanced LLMs because we want to ensure that our superiority stems from the improved visual representations, rather than from the capabilities of a more sophisticated foundation model.

We benchmark Divot-LLM on a broad range of video comprehension tasks and zero-shot video generation, achieving competitive performance through pre-training on 5 million video-text pairs. By leveraging the generality of the video tokenizer, our Divot-LLM also enables video storytelling, which generates interleaved narratives and corresponding videos that are temporally coherent through fine-tuning on a specific animation dataset.

Our contributions are three-fold. (1) We introduce Divot, an advanced video tokenizer that leverages a diffusion procedure for self-supervised video representation learning, aiming to unify video comprehension and generation. (2) We present Divot-LLM, composed of a pre-trained LLM and the Divot tokenizer to enable understanding and generating video content within a single framework. We investigate effective approaches for fitting continuous video representations using the LLM with probabilistic modeling for video generation. (3) We conduct extensive experiments to demonstrate Divot-LLM's competitive performance on existing video comprehension and generation benchmarks, as well as video storytelling. All models and code are released.

2. Related Work

MLLMs for Comprehension and Generation. With the rapid development of Multimodal Large Language Models (MLLM), recent studies have been working on unified MLLMs [12, 15–17, 25, 26, 36, 40, 41, 54, 55, 57, 67, 72–74, 77, 83, 84, 86, 89, 91, 93] that are capable of multimodal comprehension and generation. To empower LLMs with the capability to generate visual content, existing work primarily employs the following three approaches: (1) utilizing a pre-trained stable diffusion model to generate images conditioned on LLM output (either continuous features or discrete tokens); (2) employing a Vector Quantized (VQ) [63] based decoder to generate visual content from the discrete codes predicted by LLMs; (3) using LLMs to de-noise Gaussian noise through a diffusion process. While most work predominantly focus on the unification of images and texts, some pioneering studies [26, 40, 67, 74] further advance the integration of video comprehension and generation within an LLM through generating videos from discrete codes using a VQ-based decoder, which falls into the second approach. In this work, we adopt the first approach, which involves leveraging a diffusion model to achieve unified video understanding and generation from continuous representations.

Video Tokenizer in MLLMs. Previous work on video generation with LLMs predominantly employs a discrete video tokenizer to convert video signals into a sequence of quantized tokens. For example, LWM [40] and VILA-U [74] utilize a frame-level tokenizer to discretize each frame into a sequence of codes. VideoPoet [29], Loong [69] and

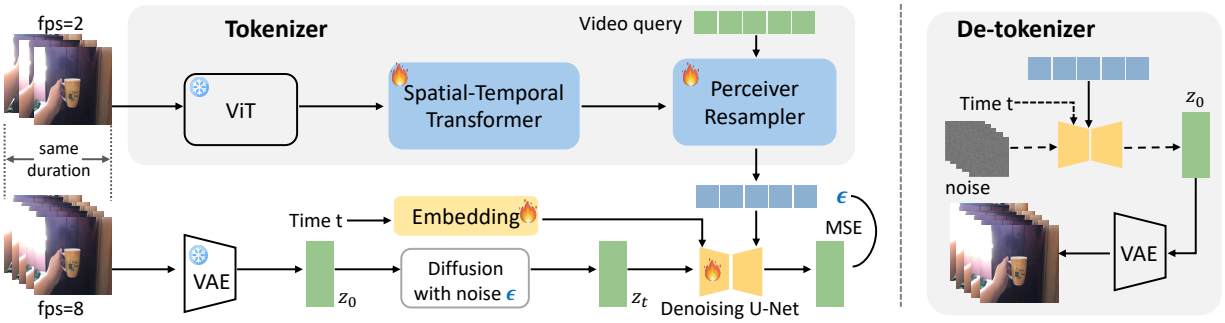


Figure 2. Overview of **Divot tokenization and de-tokenization**. During training, sparsely sampled video frames are fed into the tokenizer to obtain spatiotemporal representations. These representations serve as the conditions for a U-Net, which is trained to de-noise the noisy VAE latents of densely sampled video frames. During inference, the video representations from the Divot tokenizer can be decoded into realistic video clips with the U-Net.

Emu3 [67] leverage a 3D CNN architecture, where the encoded spatial-temporal features are quantized into discrete tokens. Video-LaVIT [26] represents video clips as a keyframe followed by extracted motion vectors, obtaining the respective discrete codes. By converting continuous visual signals into discrete tokens, the original next-token prediction mechanism can be adopted to facilitate video generation with an LLM. However, recent work [77] observes a significant performance degradation in multimodal comprehension tasks when discrete representations are used instead of continuous representations. In this work, we introduce a video tokenizer with continuous representations through leveraging the diffusion [50] procedure, enabling it to be effectively integrated with a LLM for unified comprehension and generation.

Diffusion for Representation Learning. The diffusion process has been explored as a criterion for representation learning. Some works [3, 75, 80, 90] leverage the intermediate activations of pre-trained diffusion models for downstream tasks including classification, segmentation and depth estimation. Other works [23, 65, 70] employ the diffusion model as a proxy module for self-supervised learning, where noisy inputs are de-noised by conditioning on the image representations. This approach encourages the emergence of informative representations that capture key properties and semantics of the images. In this work, to the best of our knowledge, we **for the first time** leverage diffusion for video representation learning, where a video diffusion model is trained to de-noise video clips through taking the spatiotemporal representations as conditions, thereby encouraging the capture of spatial characteristics and temporal dynamics.

3. Method

3.1. Divot Tokenizer

We introduce Divot, a diffusion-powered tokenizer that leverage diffusion procedure for video representation learning.

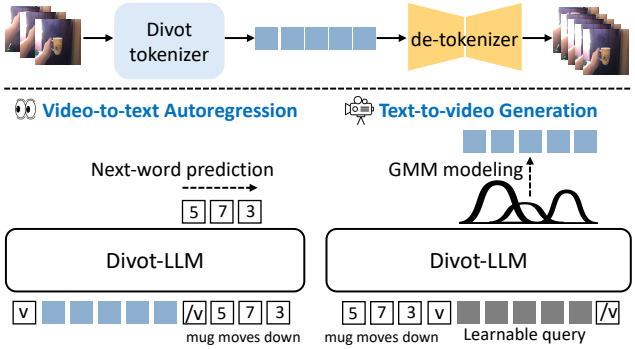


Figure 3. Overview of **Divot-LLM**. Video features from the Divot tokenizer are fed into the LLM to perform next-word prediction for video comprehension, while learnable queries are input into the LLM to model the distributions of Divot features using a **Gaussian Mixture Model** (GMM) for video generation. During inference, video features are sampled from the predicted GMM distribution to decode videos using the de-tokenizer.

Additionally, the proxy diffusion model used for training the tokenizer can serve as a de-tokenizer to decode realistic video clips from their spatiotemporal representations.

3.1.1. Preliminary: Video Diffusion Model.

Diffusion models [20, 50] learns to model a probability distribution by reversing a process that progressively adds noise to the data. Specifically, given data $\mathbf{x}_0 \sim p(\mathbf{x})$, the forward process gradually adds random Gaussian noise $\epsilon_\mathbf{t} \in \mathcal{N}(\mathbf{0}, \mathbf{I})$ to the data sample $\mathbf{x}_0$ with a total of T timesteps to yield $\mathbf{x}_t$ through a parameterization trick. The denoising process predicts $\epsilon_\mathbf{t}$ in the forward diffusion process with a denoising network $\epsilon_\theta\left(\mathbf{x}_t, t\right)$, which is trained by the objective below,

$$\min_\theta \mathbb{E}_{t,\mathbf{x}\sim p,\epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\|\epsilon - \epsilon_\theta\left(\mathbf{x}_t, t\right)\|_2^2, \tag{1}$$

where ϵ is the sampled Gaussian noise and θ indicates the parameters of the denoising network. During inference, we can perform iterative denoising from a random Gaussian

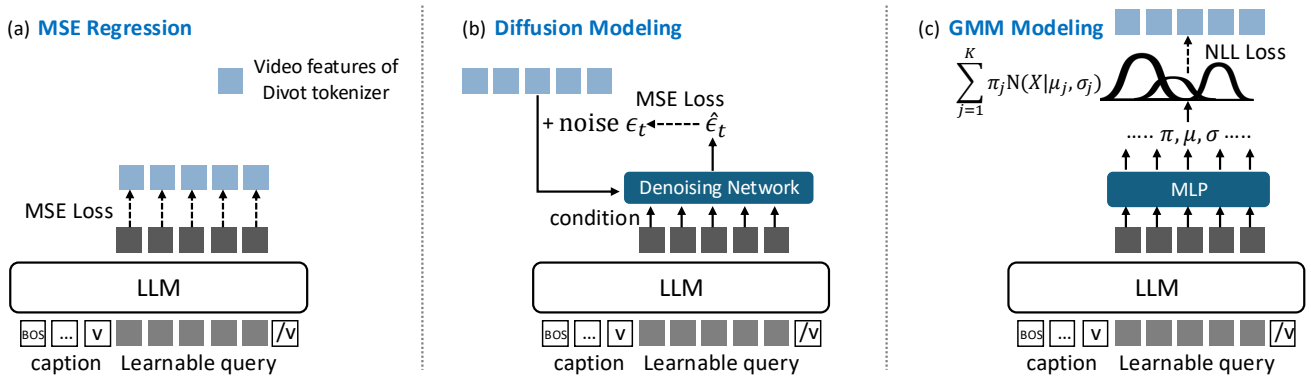


Figure 4. Paradigms for **modeling video representations** from the Divot tokenizer with a LLM for video generation. (a) **MSE Regression**, where the LLM output is trained to minimize its distance with video features using Mean Squared Error (MSE) loss; (b) **Diffusion Modeling**, where the LLM output is fed into a denoising network as the condition to predict the noise added to video features; (c) **GMM Modeling**, where the LLM output is trained to predict the parameters of a Gaussian Mixture Model (GMM) for modeling video feature distributions.

noise for the denoised data $\mathbf{x}_0$

For video diffusion models [8, 78], given a video $\mathbf{x}$, a latent representation $\mathbf{z} = \mathcal{E}(\mathbf{x})$ is first encoded to reduce the computational complexity. Then the forward diffusion process and backward denoise process are performed in this latent space with a denoising network $\epsilon_\theta(\mathbf{x}_t, \mathbf{c}, t)$, where $\mathbf{c}$ denotes denoising conditions like text or visual prompts.

3.1.2. Training Pipeline

As illustrated in Fig. 2, given a video clip, we separately sample sparse frames at 2 fps to obtain the video representations from the tokenizer, and sample dense frames at 8 fps to obtain latent representations z_0 from the frozen VAE [28] encoder. Sparse frames are sampled as the input of the video tokenizer considering the semantic redundancy between adjacent frames. The forward diffusion process gradually adds Gaussian noise θ to z_0 for producing the noisy input z_t. At each backward step t, a denoising U-Net is trained to predict the noise added from the previous step to the current step by taking the time embedding and video representations as the condition. Specifically, the video representations interact with the denoising U-Net intermediate features through cross-attention layers, where each noisy latent attends to all video tokens. By constraining the U-Net to reconstruct fine-grained spatial and temporal information of video clips through relying on video features, the Divot tokenizer is optimized to capture both spatial characteristics and temporal dynamics for robust video representations. The Divot tokenizer is trained on pure videos of a subset of WebVid-10M [2] and Panda-70M [9], totaling 10M videos.

After training the Divot tokenizer, the proxy denoising U-Net (employed to implement the parameterized loss function) can serve as an effective video de-tokenizer, which is able to decode semantically aligned video clips from the learned spatiotemporal representations as shown in Fig. 5.

3.1.3. Model Architecture

As shown in Fig. 2, the Divot tokenizer is composed of a pre-trained ViT encoder to extract frame-level features, a transformer for spatial and temporal fusion, and a Perceiver Resampler [1] to produce a fixed number of video tokens. The Perceiver Resampler is adopted for two reasons: (1) to reduce the number of video tokens that a LLM need to predict for generation, and (2) to transform the patch-position dependent features into a sequence of high-level features without 3D positional dependencies, which we empirically find easier for an LLM to fit (See Sec. 4.3). Specifically, given a video clip with a duration of two seconds, we sample 5 frames at 2 fps, resulting in a total of 64 video tokens. We adopt the de-noising U-Net in DynamiCrafter [78], but reduce the input channel of the 3D convolution from 8 to 4 since we remove the original concatenation of a conditional image with noisy latents.

3.2. Video Representation Modeling with LLM

The core challenge of generating videos using a LLM with the Divot tokenizer lies in effectively modeling the continuous video features. The most straightforward solution is to minimize the distance between the LLM output and the video representations using mean squared error (MSE) loss following previous work [17, 54] for image generation as illustrated in Fig. 4 (a). However, we empirically find that this approach is not effective for modeling continuous video features, as the generated videos tend to exhibit repeating patterns. We analyze that the deterministic regression regularizes the LLM to learn overly averaged representations, which is particularly catastrophic in video generation as videos must ensure both spatial and temporal diversity.

Inspired by recent work MAR [33], instead of deterministic regression, we aim to model the probability distribution of video representations using the LLM. As shown in Fig. 4,

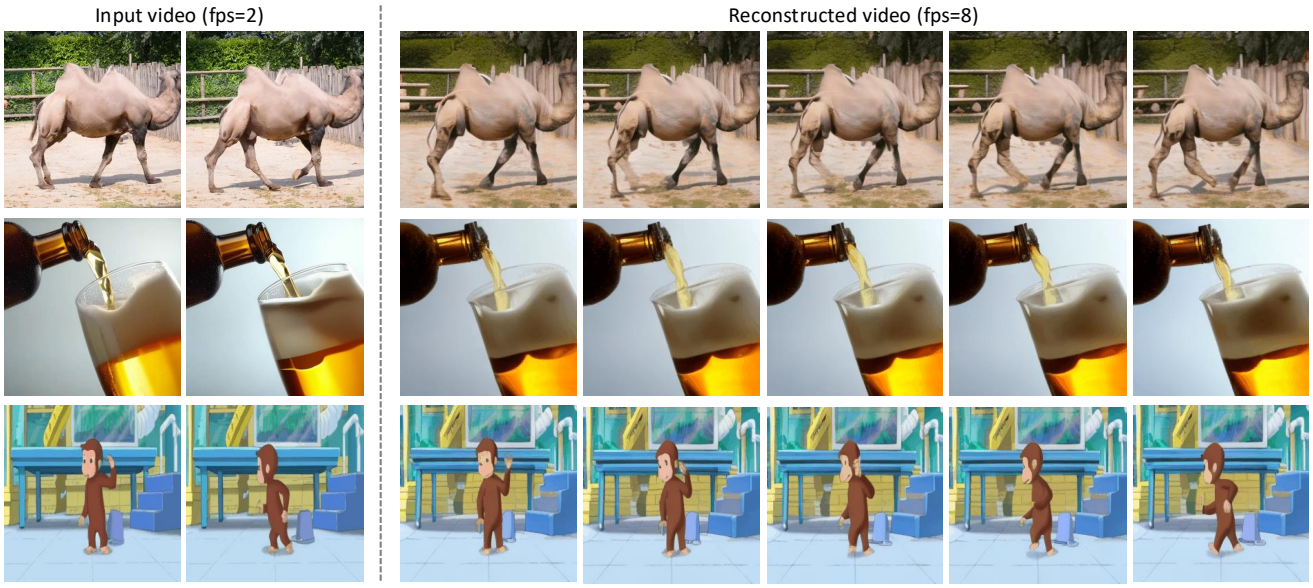


Figure 5. Reconstructed videos, where the Divot tokenizer obtains spatiotemporal representations of sparsely sampled video frames and the de-tokenizer decodes these representations into semantically aligned and temporally coherent video clips.

we explore two approaches including (b) Diffusion Modeling [33] and (c) GMM Modeling [61]. Specifically, for the diffusion modeling, given continuous-valued video features to be predicted, the LLM produces output, which serves as the condition of a denoising network (a small MLP) to predict the Gaussian noise added to the video features. The diffusion model is trained for representing the distribution of video features. For GMM modeling, we use a Gaussian Mixture Model (GMM) to model the distribution of the video features, and train the LLM to predict $2kd + k$ parameters per video token (kd mean and kd variance parameters for the mixture components, and k mixture probabilities). We optimize the LLM by minimizing the discrepancy between the predicted GMM distribution and the video representations with negative log-likelihood (NLL) loss.

During inference, in diffusion modeling, the denoising network denoise the final video features from Gaussian noise gradually by taking the LLM output as the condition. In GMM modeling, we draw samples from the predicted GMM distribution as the final video representations. To empirically investigate the effectiveness of the approaches above, we train the LLM with MSR-VTT [79], and evaluates text-to-video generation on test set with FVD [62] and similarity score [49] as the metric following previous work [26, 82]. As listed in Tab. 7, GMM modeling achieves better performance than diffusion modeling and MSE regression in video generation. We speculate that high-level features obtained by the Divot tokenizer are more sensitive to Gaussian noise compared to the VAE latents used by MAR, making training more challenging and resulting in suboptimal results. Therefore, we adopt GMM modeling to train Divot-LLM.

Table 2. Datasets used for training the tokenizer and Divot-LLM.

Stage	Type	Dataset
Tokenize	Pure Video	WebVid-10M [2], Panda-70M [9]
Pre-train	Video-text	WebVid-10M [2]
	Image-text	CC3M [52], CapsFusion [87], LAION-COCO [51]
SFT	Classification	Kinetics-710 [27], SSV2 [18]
	VQA	TGIF [34], NextQA [76], CLEVRER [85], YouCook2 [92], PerceptionTest[48], EgoQA [19], ActivityNetQA[88]
	Instruction	Video-ChatGPT[43], LLaVA-mixed[39], Valley [42], LLaVA-Video-178K[37]
	Generation	WebVid-10M [2]
	StoryTelling	In-house data

3.3. Pre-training and Instruction Tuning

3.3.1. Training Stage I: Multimodal Pre-training

As shown in Fig 3, Divot-LLM adopts next-word prediction and GMM modeling on video-text data for video comprehension and generation. Specifically, the video features from the Divot tokenizer, the special tokens indicating the start and end of video features, along with the text tokens of the caption are fed into the pre-trained Mistral-7B [24] for next token prediction trained with cross-entropy loss. Text tokens of the caption and N learnable queries are input into the LLM, where the output of the learnable queries are trained via bidirectional attention to model a GMM distribution for the video features using NLL loss. During inference, we draw samples from the predicted GMM distribution as condition of the denoising U-Net to decode videos. We pre-train Divot-LLM from the pre-trained Mistral-7B model using

Table 3. Comparison for video comprehension with MLLMs. "Video-Gen" denotes whether the model can generate videos besides texts. The evaluation metric is accuracy. The best results are bold and the second best results are underlined.

Model	LLM size	Video-Gen	EgoSchema	Perception-Test	MVBench	MSVD	ActivityNet
Gemini 1.0 Pro [58]	-	×	55.7	51.1	-	-	49.8
Gemini 1.5 Pro [59]	-	×	63.2	-	-	-	56.7
GPT4-V [46]	-	×	55.6	-	43.7	-	59.5
GPT4-O [47]	-	×	**72.2**	-	-	-	<u>61.9</u>
LLaMA-VID [35]	7B	×	38.5	44.6	41.9	69.7	47.4
Video-ChatGPT [43]	7B	×	-	-	-	64.9	35.2
Video-LLaVA [37]	7B	×	38.4	44.3	41.0	70.7	45.3
VideoChat2 [31]	7B	×	42.2	47.3	51.1	70.0	49.1
LLaVA-NeXT-Video [38]	7B	×	43.9	48.8	46.5	67.8	53.5
LLaVA-NeXT-Video [38]	32B	×	60.9	-	-	-	54.3
PLLaVA [81]	34B	×	-	58.1	-	-	60.9
LLaVA-OneVision [30]	72B	×	62.0	-	-	-	**62.3**
VideoLLaMA2 [10]	7B	×	51.7	51.4	<u>54.6</u>	70.9	50.2
VideoLLaMA2 [10]	72B	×	<u>63.9</u>	<u>57.5</u>	**62.0**	71.0	55.2
LWM [40]	7B	✓	-	-	-	55.9	-
Video-LaVIT [26]	7B	✓	37.3	47.9	-	73.2	50.1
VILA-U [74]	7B	✓	-	-	-	<u>75.3</u>	52.7
Divot-LLM	7B	✓	46.5	**58.3**	52.1	**76.4**	55.8

LoRA [22] on a subset of WebVid-10M [2] data (filtered for temporal dynamics in captions) and image-text data.

3.3.2. Training Stage II: Multimodal Instruction Tuning

We perform multimodal instruction tuning on Divot-LLM to align it with human instructions through supervised fine-tuning on public datasets as listed in Tab. 2 with a LoRA module. We further fine-tune the pretrained Divot-LLM on an animated series called "Curious George" to achieve video storytelling, which generates storyline and corresponding video clips in an interleaved manner.

4. Experiment

4.1. Quantitative Evaluation

Video Comprehension. We conduct extensive evaluations on video comprehension benchmarks including *Multi-choice Video Question Answering* (MC-VQA) on EgoSchema [44], Perception-Test [48], MVBench [32], and *Open-Ended Video Question Answering* (OE-VQA) on MSVD [7], ActivityNet [88]. Following VideoLLaMA 2 [10], we utilize GPT-3.5 to assess the quality of the generated answers of OE-VQA by determining whether the answers match the ground truth, and we report the percentage of "Yes" as Accuracy.

For each testing video, we sample a maximum of 20 clips, each containing 5 frames. The evaluation results are reported in Tab. 3. Divot-LLM outperforms the baseline models that can generate both texts and videos, which demonstrates that our model effectively achieves video comprehension within a unified framework. Compared to VideoLLMs specifically designed for video comprehension of the same model size of LLM, Divot-LLM achieves competitive results with significantly fewer video-caption pairs for training (4.8M vs. 100M in VideoLLaMA 2). By utilizing diffusion procedure for video representation learning, our Divot tokenizer effectively captures robust spatiotemporal representations, enhancing the comprehension capabilities.

Video Generation. We evaluate zero-shot text-to-video generation on MSR-VTT [79]. We randomly sample one caption for each testing video and generate 16 frames in 256 x 256px resolution. We adopt the CLIP similarity (CLIPSIM) [71] and Fréchet video distance (FVD) [62] as the evaluation metric following Loong [69]. As listed in Tab. 4, Divot-LLM achieves performance comparable to existing video generation models in terms of visual quality and semantic alignment with captions using only 4.8 million video-text pairs for training.

4.2. Qualitative Evaluation

Text-to-video Generation. We perform qualitative comparison of text-to-video generation with baseline MLLMs that are capable of unified video comprehension and generation. As illustrated in Fig. 6, through modeling the distributions of Divot features with a predicted GMM, our Divot-LLM can generate videos that are both semantically aligned with text prompts and temporally coherent within frames.

Video StoryTelling. We fine-tune the pre-trained Divot-LLM on an animated series called "Curious George" for video storytelling. As shown in Fig. 7, given a brief story instruction, our Divot-LLM can generate a sequence of multimodal stories with rich narrative text and contextually relevant videos that are temporally coherent. Since we only

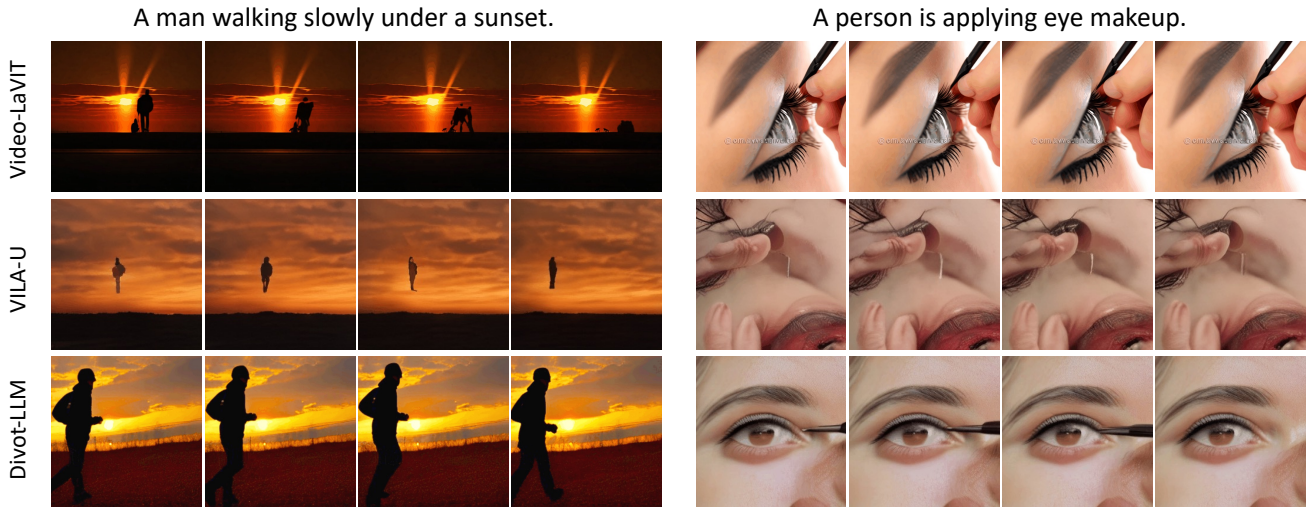


Figure 6. Qualitative comparison of text-to-video generation with MLLMs that are capable of unified video comprehension and generation. Divot-LLM effectively generates videos that are both semantically aligned with text prompts and temporally coherent within frames.

fine-tune the de-tokenizer for adaptation to the new domain, it demonstrates the generalizability of our Divot tokenizer for obtaining robust video representations.

Table 4. Comparison for zero-shot text-to-video generation. "Data size" refers to the number of training video data, and "Unified" denotes if the model enables video comprehension and generation. The best results are bold and the second best results are underlined.

Model	Data size	Unified	MSR-VTT	
			CLIPSIM (↑)	FVD (↓)
CogVideo [21]	5.4M	×	0.2631	1294
Video LDM [5]	10M	×	0.2929	-
VideoComposer [66]	10M	×	0.2932	580
InternVid [68]	28M	×	0.2951	-
Make-A-Video [53]	20M	×	**0.3049**	-
VideoPoet [29]	270M	×	**0.3049**	213
PYoCo [14]	22.5M	×	-	-
SVD [4]	152M	×	-	-
Video-LavIT [26]	10M	✓	<u>0.3012</u>	<u>188.36</u>
Loong [69]	16M	×	0.2903	274
Snap Video [45]	-	×	0.2793	**110.4**
VILA-U [74]	1M	✓	0.2937	499.06
Divot-LLM	4.8M	✓	0.2938	301.4

4.3. Ablation Study

Diffusion for Video Comprehension. We design two baselines to validate the effectiveness of the diffusion procedure to learn spatiotemporal representations for VideoLLMs. As shown in Tab. 5, both models are pre-trained on Valley [42] and instruction tuned on Video-ChatGPT [43]. The model with diffusion loss employs our Divot tokenizer, while the model with caption loss adopts the same architecture but its tokenizer is pre-trained using captioning loss with a frozen LLM on Valley. The model that employs a video tokenizer trained with diffusion loss achieves better performance in video comprehension benchmarks, demonstrating that the diffusion process can effectively learn robust video representations in a self-supervised manner, without the need for paired caption annotations.

Table 6. Ablation study on the training objective of the video tokenizer. The evaluation metric is accuracy.

Loss Type	MV-Bench	MSVD	ActivityNet
Caption	30.8	66.1	43.2
Diffusion	**33.2**	**68.9**	**44.3**

Video Generation with LLM. We perform various ablation studies to explore an effective approach for generating videos with a LLM through training on MSR-VTT training set and evaluating text-to-video generation on test set. We use ViT-G/14 to calculate CLIPSIM for better discrimination.

Q1: Which type of video representations is easier? We investigate two types of video representations, patch-position dependent features obtained from a spatial-temporal transformer and patch-position independent features after a Perceiver Resample with learnable queries. As listed in Tab. 7, fitting features without 3D positional dependencies achieves higher performance, which is also observed in recent work [64]. We also experiment with training the video tokenizer in the VAE manner, which involves predicting the means and variances of a normal distribution and sampling video representations using the re-parametrization trick following GIVT [61]. However, we observe that it is difficult for the LLM to converge and the video de-tokenizer achieves unsatisfactory reconstruction results. We conclude that *introducing variances during tokenization for high-level video features may not be appropriate*.

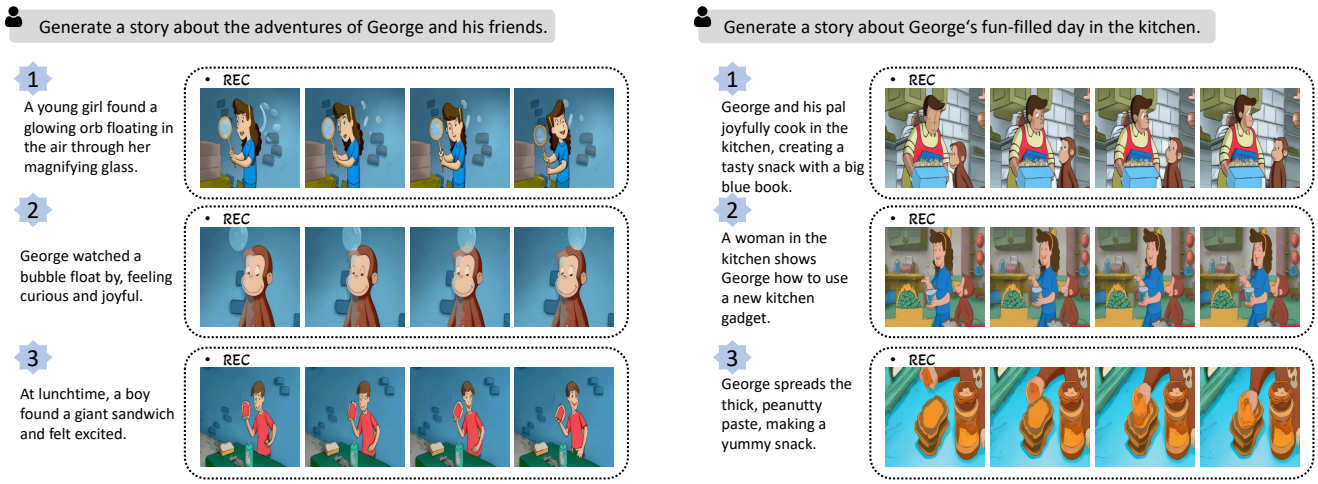


Figure 7. Qualitative examples of video storytelling by Divot-LLM. Given a story instruction, Divot-LLM can generate rich textual narratives along with corresponding video clips that are temporally coherent in an interleaved manner.

Table 8. Ablation study on video representation modeling with LLMs for generation. We evaluate text-to-video generation on MSR-VTT.

	Representation		**Objective**				**Mechanism**		
	patch-position dependent	patch-position independent	MSE	Diffusion		GMM	AR	Query	
				ϵ-pred	v-pred			causal	bidirectional
CLIPSIM (↑)	0.3192	**0.3265**	0.3168	0.2811	0.2842	**0.3265**	0.2386	0.3080	**0.3265**
FVD (↓)	378.50	**366.60**	438.94	418.19	377.17	**366.60**	447.88	416.60	**366.60**

Q2: Which training objective is more suitable? As introduce in Sec. 3, we explore MSE regression, Diffusion modeling and GMM modeling to fit high-level continuous features with a LLM. As listed in Tab. 7, simply aligning the LLM output with video features using MSE loss yields the lowest generation quality, suggesting that deterministic regression is inadequate for modeling spatiotemporal representations. Training a denoising network to denoise the noisy video features by taking the LLM output as the condition also achieves inferior performance with both ϵ prediction and v prediction. Different from MAR [33] that denoises low-level VAE latents, our goal is to denoise high-level video features. We speculate that these features are more sensitive to Gaussian noise, making them more challenging to denoise. Training the LLM to model the distribution of high-level video features using a GMM model achieves the best generation quality and semantic alignment with captions.

Q3: Which LLM mechanism is more effective? We train the LLM to fit the video representations with GMM modeling using both an autoregressive approach and a query-based approach, with the latter exploring causal attention and bidirectional attention within the LLM. Predicting the features of each video token in an autoregressive manner results in the worst performance due to the accumulation of errors, particularly when the features of the pervious token are sampled from a GMM distribution for predicting the distribution of the current token. The query-based approach achieves better results with bidirectional attention, as it enables each query to attend to all tokens for predictions.

5. Conclusion

In this work, we introduce Divot, a diffusion-powered video tokenizer learned in a self-supervised manner for unified comprehension and generation. We further investigate effective approaches for modeling continuous video representations with the LLM and present Divot-LLM to understand and generate video content in a single framework. Divot-LLM achieves competitive performance in video comprehension and generation benchmarks, and enables video storytelling effectively. We hope our work will draw increased attention to unifying video comprehension and generation through the design of sophisticated tokenizers.

Limitation. As we primarily focus on exploring effective representations and approaches for video generation with a unified LLM, the current model is trained to predict video representations for only a single clip and does not generate longer videos, which will be explored in our future work.

References

[1] 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. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022. 2, 4

[2] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1728–1738, 2021. 4, 5, 6

[3] Dmitry Baranchuk, Ivan Rubachev, Andrey Voynov, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. *arXiv preprint arXiv:2112.03126*, 2021. 3

[4] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 7

[5] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023. 7

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 1

[7] David Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 190–200, 2011. 6

[8] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023. 4

[9] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, et al. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13320–13331, 2024. 4, 5

[10] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024. 6

[11] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022. 1

[12] Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, et al. Dreamllm: Synergistic multimodal comprehension and creation. *arXiv preprint arXiv:2309.11499*, 2023. 1, 2

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[14] Songwei Ge, Seungjun Nah, Guilin Liu, Tyler Poon, Andrew Tao, Bryan Catanzaro, David Jacobs, Jia-Bin Huang, Ming-Yu Liu, and Yogesh Balaji. Preserve your own correlation: A noise prior for video diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22930–22941, 2023. 7

[15] Yuying Ge, Yixiao Ge, Ziyun Zeng, Xintao Wang, and Ying Shan. Planting a seed of vision in large language model. *arXiv preprint arXiv:2307.08041*, 2023. 1, 2

[16] Yuying Ge, Sijie Zhao, Ziyun Zeng, Yixiao Ge, Chen Li, Xintao Wang, and Ying Shan. Making llama see and draw with seed tokenizer. *arXiv preprint arXiv:2310.01218*, 2023.

[17] Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying Shan. Seed-x: Multimodal models with unified multi-granularity comprehension and generation. *arXiv preprint arXiv:2404.14396*, 2024. 1, 2, 4

[18] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017. 5

[19] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022. 5

[20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3

[21] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022. 7

[22] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 6

[23] Drew A Hudson, Daniel Zoran, Mateusz Malinowski, Andrew K Lampinen, Andrew Jaegle, James L McClelland, Loic Matthey, Felix Hill, and Alexander Lerchner. Soda: Bottle-

neck diffusion models for representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23115–23127, 2024. 3

[24] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023. 2, 5

[25] Yang Jin, Kun Xu, Liwei Chen, Chao Liao, Jianchao Tan, Bin Chen, Chenyi Lei, An Liu, Chengru Song, Xiaoqiang Lei, et al. Unified language-vision pretraining with dynamic discrete visual tokenization. *arXiv preprint arXiv:2309.04669*, 2023. 2

[26] 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. *arXiv preprint arXiv:2402.03161*, 2024. 1, 2, 3, 5, 6, 7

[27] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 5

[28] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2, 4

[29] 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. *arXiv preprint arXiv:2312.14125*, 2023. 2, 7

[30] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 6

[31] 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. *arXiv preprint arXiv:2311.17005*, 2023. 6

[32] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206, 2024. 6

[33] Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. *arXiv preprint arXiv:2406.11838*, 2024. 2, 4, 5, 8

[34] Yuncheng Li, Yale Song, Liangliang Cao, Joel Tetreault, Larry Goldberg, Alejandro Jaimes, and Jiebo Luo. Tgif: A new dataset and benchmark on animated gif description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4641–4650, 2016. 5

[35] Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. *arXiv preprint arXiv:2311.17043*, 2023. 6

[36] Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. Mini-gemini: Mining the potential of multi-modality vision language models. *arXiv preprint arXiv:2403.18814*, 2024. 2

[37] Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023. 5, 6

[38] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 6

[39] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 5

[40] Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. World model on million-length video and language with ringattention. *arXiv preprint arXiv:2402.08268*, 2024. 1, 2, 6

[41] Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, and Aniruddha Kembhavi. Unified-io 2: Scaling autoregressive multimodal models with vision, language, audio, and action. *arXiv preprint arXiv:2312.17172*, 2023. 2

[42] Ruipu Luo, Ziwang Zhao, Min Yang, Junwei Dong, Da Li, Pengcheng Lu, Tao Wang, Linmei Hu, Minghui Qiu, and Zhongyu Wei. Valley: Video assistant with large language model enhanced ability. *arXiv preprint arXiv:2306.07207*, 2023. 5, 7

[43] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023. 5, 6, 7

[44] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244, 2023. 6

[45] Willi Menapace, Aliaksandr Siarohin, Ivan Skorokhodov, Ekaterina Deyneka, Tsai-Shien Chen, Anil Kag, Yuwei Fang, Aleksei Stoliar, Elisa Ricci, Jian Ren, et al. Snap video: Scaled spatiotemporal transformers for text-to-video synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7038–7048, 2024. 7

[46] OpenAI. Gpt-4v(ision) system card, 2023. 6

[47] OpenAI. Gpt-4o system card, 2024. 6

[48] Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Mateusz Malinowski, Yi Yang, Carl Doersch, et al. Perception test: A diagnostic benchmark for multimodal video models. *Advances in Neural Information Processing Systems*, 36, 2024. 5, 6

[49] 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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 5

[50] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 3

[51] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 5

[52] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018. 5

[53] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 7

[54] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Zhengxiong Luo, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, et al. Generative multimodal models are in-context learners. *arXiv preprint arXiv:2312.13286*, 2023. 2, 4

[55] Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative pretraining in multimodality. *arXiv preprint arXiv:2307.05222*, 2023. 1, 2

[56] Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative pretraining in multimodality. *arXiv preprint arXiv:2307.05222*, 2023.

[57] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024. 1, 2

[58] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 6

[59] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 6

[60] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 1

[61] Michael Tschannen, Cian Eastwood, and Fabian Mentzer. Givt: Generative infinite-vocabulary transformers. In *European Conference on Computer Vision*, pages 292–309. Springer, 2025. 5, 7

[62] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. 5, 6

[63] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 2

[64] Hanyu Wang, Saksham Suri, Yixuan Ren, Hao Chen, and Abhinav Shrivastava. Larp: Tokenizing videos with a learned autoregressive generative prior. *arXiv preprint arXiv:2410.21264*, 2024. 7

[65] Wenxuan Wang, Quan Sun, Fan Zhang, Yepeng Tang, Jing Liu, and Xinlong Wang. Diffusion feedback helps clip see better. *arXiv preprint arXiv:2407.20171*, 2024. 3

[66] 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. *Advances in Neural Information Processing Systems*, 36, 2024. 7

[67] 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*, 2024. 2, 3

[68] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*, 2023. 7

[69] 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*, 2024. 2, 6, 7

[70] Chen Wei, Karttikeya Mangalam, Po-Yao Huang, Yanghao Li, Haoqi Fan, Hu Xu, Huiyu Wang, Cihang Xie, Alan Yuille, and Christoph Feichtenhofer. Diffusion models as masked autoencoders. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16284–16294, 2023. 3

[71] Chenfei Wu, Lun Huang, Qianxi Zhang, Binyang Li, Lei Ji, Fan Yang, Guillermo Sapiro, and Nan Duan. Godiva: Generating open-domain videos from natural descriptions. *arXiv preprint arXiv:2104.14806*, 2021. 6

[72] Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. *arXiv preprint arXiv:2410.13848*, 2024. 2

[73] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519*, 2023.

[74] 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*, 2024. 1, 2, 6, 7

[75] Weilai Xiang, Hongyu Yang, Di Huang, and Yunhong Wang. Denoising diffusion autoencoders are unified self-supervised

learners. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15802–15812, 2023. 3

[76] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9777–9786, 2021. 5

[77] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024. 1, 2, 3

[78] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Gongye Liu, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Dynamicrafter: Animating open-domain images with video diffusion priors. In *European Conference on Computer Vision*, pages 399–417. Springer, 2025. 2, 4

[79] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016. 5, 6

[80] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2955–2966, 2023. 3

[81] Lin Xu, Yilin Zhao, Daquan Zhou, Zhijie Lin, See Kiong Ng, and Jiashi Feng. Pllava: Parameter-free llava extension from images to videos for video dense captioning. *arXiv preprint arXiv:2404.16994*, 2024. 6

[82] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*, 2021. 5

[83] Jian Yang, Dacheng Yin, Yizhou Zhou, Fengyun Rao, Wei Zhai, Yang Cao, and Zheng-Jun Zha. Mmar: Towards lossless multi-modal auto-regressive prababilistic modeling. *arXiv preprint arXiv:2410.10798*, 2024. 2

[84] Shuai Yang, Yuying Ge, Yang Li, Yukang Chen, Yixiao Ge, Ying Shan, and Yingcong Chen. Seed-story: Multimodal long story generation with large language model. *arXiv preprint arXiv:2407.08683*, 2024. 2

[85] Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. Clevrer: Collision events for video representation and reasoning. *arXiv preprint arXiv:1910.01442*, 2019. 5

[86] Lili Yu, Bowen Shi, Ramakanth Pasunuru, Benjamin Muller, Olga Golovneva, Tianlu Wang, Arun Babu, Binh Tang, Brian Karrer, Shelly Sheynin, et al. Scaling autoregressive multimodal models: Pretraining and instruction tuning. *arXiv preprint arXiv:2309.02591*, 2023. 1, 2

[87] Qiying Yu, Quan Sun, Xiaosong Zhang, Yufeng Cui, Fan Zhang, Yue Cao, Xinlong Wang, and Jingjing Liu. Capsfusion: Rethinking image-text data at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14022–14032, 2024. 5

[88] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9127–9134, 2019. 5, 6

[89] Chuyang Zhao, Yuxing Song, Wenhao Wang, Haocheng Feng, Errui Ding, Yifan Sun, Xinyan Xiao, and Jingdong Wang. Monoformer: One transformer for both diffusion and autoregression. *arXiv preprint arXiv:2409.16280*, 2024. 2

[90] Wenliang Zhao, Yongming Rao, Zuyan Liu, Benlin Liu, Jie Zhou, and Jiwen Lu. Unleashing text-to-image diffusion models for visual perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5729–5739, 2023. 3

[91] Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. *arXiv preprint arXiv:2408.11039*, 2024. 1, 2

[92] Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018. 5

[93] Jinguo Zhu, Xiaohan Ding, Yixiao Ge, Yuying Ge, Sijie Zhao, Hengshuang Zhao, Xiaohua Wang, and Ying Shan. Vl-gpt: A generative pre-trained transformer for vision and language understanding and generation. *arXiv preprint arXiv:2312.09251*, 2023. 1, 2