VideoAuteur: Towards Long Narrative Video Generation

Junfei Xiao[1], Feng Cheng[2], Lu Qi[2], Liangke Gui[2], Yang Zhao[2], Shanchuan Lin[2]
Jiepeng Cen[2], Zhibei Ma[2] Alan Yuille[1], Lu Jiang[2]

[1]Johns Hopkins University [2]ByteDance

Project Page: https://videoauteur.github.io

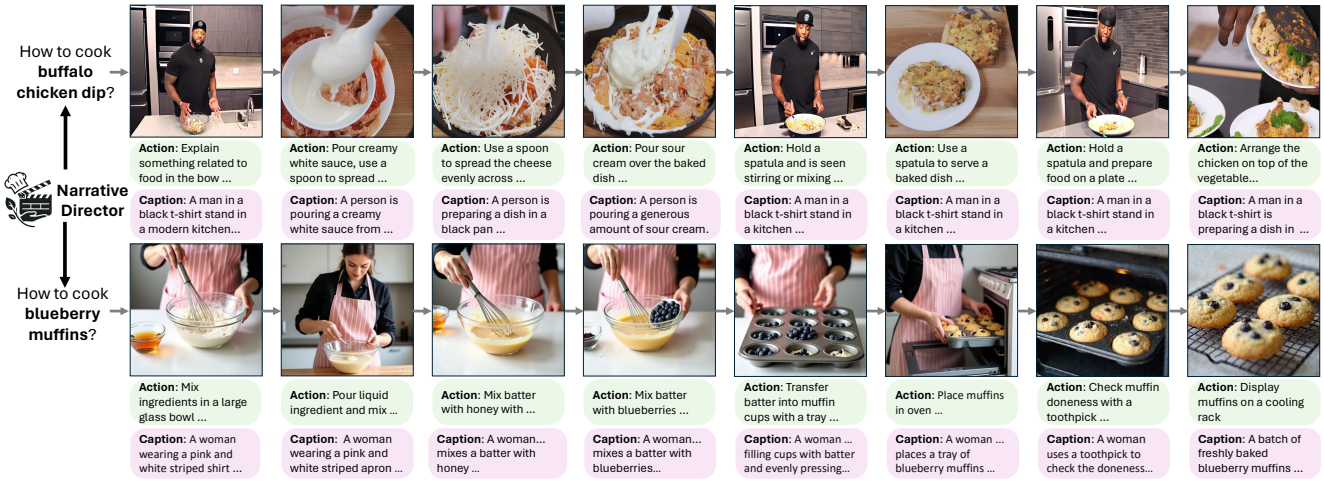

Figure 1. **Long Narrative Video Generation.** We curate a large-scale cooking video dataset to develop an interleaved auto-regressive model – **VideoAuteur**, which acts as a narrative director, sequentially generating actions, captions, and keyframes (two generated examples here). These elements condition a video generation model to create long narrative videos.

Abstract

Recent video generation models have shown promising results in producing high-quality video clips lasting several seconds. However, these models face challenges in generating long sequences that convey clear and informative events, limiting their ability to support coherent narrations. In this paper, we present a large-scale cooking video dataset designed to advance long-form narrative generation in the cooking domain. We validate the quality of our proposed dataset in terms of visual fidelity and textual caption accuracy using state-of-the-art Vision-Language Models (VLMs) and video generation models, respectively. We further introduce a Long Narrative Video Director to enhance both visual and semantic coherence in generated videos and emphasize the role of aligning visual embeddings to achieve improved overall video quality. Our method demonstrates substantial improvements in generating visually detailed and semantically aligned keyframes, supported by finetuning techniques that integrate text and image embeddings within the video generation process.

1. Introduction

Video generation [5, 6, 19, 20, 41, 51, 62] has witnessed remarkable advancements with diffusion [2, 21, 34, 57] and auto-regressive models [25, 43, 44, 53]. A primary objective is to generate video clips from text prompts and supports various downstream applications, such as image animation [8, 54], video editing [4, 11], video stylization [23].

With the maturity of generating high-fidelity short video clips, researchers begin setting their sights on the next north-star: creating videos capable of conveying a complete narrative which captures an account of events unfolding over time. The importance of narratives has been highlighted in the literature. For example, Bruner argues that narratives are essential tools for organizing experiences and memories [3]. The book *Sapiens: A Brief History of Humankind* emphasizes that the ability to share narratives (stories) has been pivotal in human development, setting humans apart from other animals [15].

Long **N**arrative **V**ideo **G**eneration (NVG) introduces several challenges. One particularly challenge is the scarcity of video data suitable for learning coherent narra-

tives in video. While our community has developed many video datasets, most are unsuitable for NVG. First, most videos are tagged with descriptions that are partially to NVG. Second, even for the relevant descriptions, these descriptions may be either too coarse or lack detailed actions needed for NVG. Finally, not all videos contain meaningful narratives suitable for learning and can be well evaluated.

Consequently, video data with clear, complete, and meaningful narratives is crucial not only for training but also for evaluating and comparing NVG methods. However, compared to story generation through a sequence of images [14, 24, 32, 56], progress in narrative video generation has been relatively slow, partly due to the absence of standardized training and evaluation benchmarks.

This paper contributes to advancing research in narrative video generation in two ways. First, we curate and annotate a large-scale video dataset on the cooking domain. The samples in our dataset are structured with clear narrative flows, each composed of sequential actions and visual states. Our dataset consists of approximately 200,000 video clips, with an average duration of 9.5 seconds per clip. We select cooking videos for their well-defined and less ambiguous narratives, making them more objective to evaluate consistently. To address video copyright concerns, we source videos from existing video datasets, YouCook2 [63] and HowTo100M [33]. We design various mechanisms to ensure high-quality videos and captions, organized in a structured storyboard format, as illustrated in Figure 1.

Additionally, we propose a new auto-regressive pipeline for long narrative video generation, comprising three main components: a long narrative director, a rolling-context conditioned keyframe renderer, and a visual-conditioned video generation model. The long narrative director produces a coherent narrative flow by generating a sequence of visual embeddings or keyframes that represent the story's logical progression. Building upon this, the rolling-context conditioned keyframe renderer utilizes a rolling history of reference images as contextual conditioning to generate high-quality keyframes with consistency. Finally, the visual-conditioned video generation model produces video clips based on these visual conditions to do narrative.

Extensive experiments on the large-scale collected dataset demonstrate the effectiveness of the proposed pipeline for long narrative video generation. To sum up, our contributions are as follows:

- We construct **CookGen**, a large-scale, structured dataset accompanied by an effective data pipeline to benchmark long-form narrative video generation. The dataset along with the necessary functionalities will be opensourced to advance future research in the area.
- We propose **VideoAuteur**, a novel approach for automated long video generation. It effectively bridges interleaved auto-regressive multimodal LLMs with pretrained DiTs, employing a rolling context strategy for enhanced generation quality and visual consistency.
- Extensive experimental results and ablation studies show that **VideoAuteur** achieves the state-of-the-art performance in long narrative video generation.

2. Related Works

Text-to-Image/Video Generation Text-to-image [7, 26, 35, 36, 38, 50, 58] and video generation [5, 6, 19, 20, 41, 51, 62] have made remarkable progress to generate high-fidelity video clip of 5-10 seconds. For example, latent design [38] has become mainstream, balancing effectiveness with efficiency. Building upon this design, diffusion-based models like DiT [34], Sora [2], and CogVideo [21, 57] leveraged larger datasets and explored refined architectures and loss functions to enhance performance. In contrast, auto-regressive models such as VideoPoet [25] and Emu series [43, 44, 53] sequentially predict image or video tokens. Instead, our work focuses on the model's ability to generate long narrative videos beyond a few seconds.

Interleaved Image-Text Modeling Interleaved image-text generation [1, 9, 13, 13, 45, 55] has garnered attention as a compelling research area that merges visual and textual modalities to produce rich outputs. Earlier approaches [29, 37, 37, 42] primarily relied on large-scale image-text paired datasets [12, 39] but were often confined to single-modality tasks, such as captioning or text-to-image generation. With the emergence of large language models [47], various vision-language models [28, 31, 52] have stepped in a new era of unified representations, leveraging well-curated datasets for interleaved generation. However, most existing works focus on the one-time generation and do not address the coherence of generated content, which is our focus.

Narrative Visual Generation Existing narrative visual generation primarily focuses on addressing challenges related to semantic and visual consistency. Recent approaches such as Narrative Visual Generation, VideoDirectorGPT [30], Vlogger [65], Animate-a-story [16], VideoTeris [46], IC Lora [22], Vlogger [65], and Animate-a-story [16] employ various methods to enhance semantic coherence and visual continuity. Unlike most prior methods that mainly focus on consistent image generation [22, 56, 64], our target is generating coherent narrative videos. While some works make efforts to be language-centric using text as conditions for video generation [54, 65] or appending with keyframes [61], different from these work, we propose an integrated approach that leverages multimodal large language models (LLMs) in conjunction with in-context diffusion transformer models to ensure global narrative coherence, subsequently conditioning the video generation model.

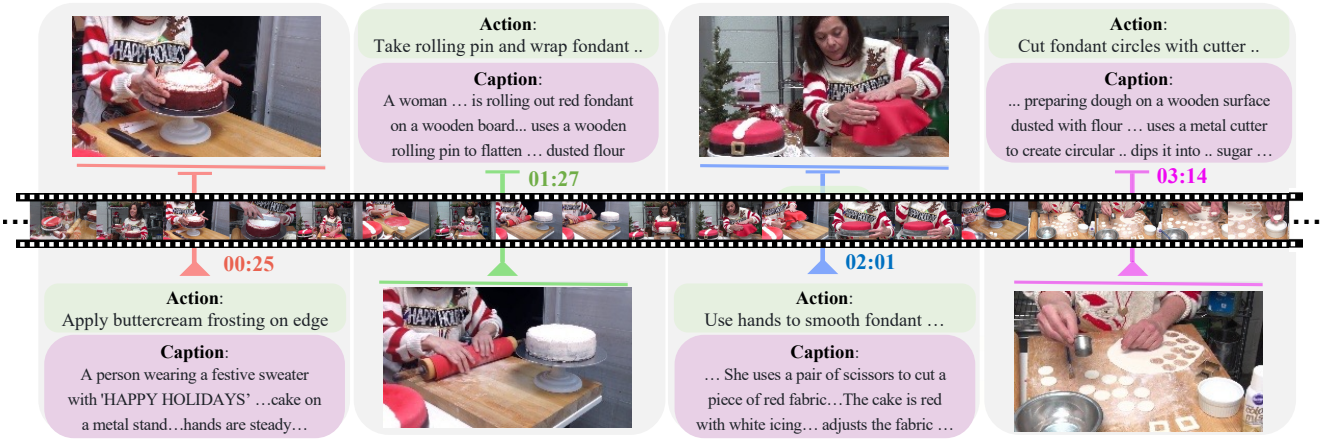

Figure 2. **CookGen contains long narrative videos annotated with actions and captions.** Each source video is cut into clips and matched with the labeled "actions". We use refined pseudo labels from ASR for Howto100M videos and use manual annotations for Youcook2 videos. We use state-of-the-art VLMs (*i.e.* GPT-4o and an expert captioner) to provide high-quality captions for all video clips.

3. CookGen: a Long Narrative Video Dataset

To the best of our knowledge, datasets for long narrative video generation research is extremely limited. To enable in-depth exploration and establish an experimental setting, we establish **CookGen**, a large video dataset with detailed annotations on captions, actions, and annotations. As the data example provided in Figure 2, our dataset focuses on cooking videos. We prioritize cooking over other video categories because each dish follows a pre-defined, strict sequence of action steps. These structured and unambiguous objectives in cooking videos are essential for learning and evaluating long video narrative generation.

3.1. Overview

We source over 30,000 raw videos about from two existing video datasets: YouCook2 [63] and HowTo100M [33]. Each video is filtered and cropped with processing to remove corruptions. Table 1 provides detailed information about the dataset statistics, video and clip details, and the train/val partitioning. Appendix B provides more details.

Table 2 compares our dataset with existing datasets most relevant to multimodal narrative generation. Unlike existing datasets that primarily focus on image-based comic story generation, our real-world narrative dataset offers several advantages. First, the videos in our dataset depict procedural activities (*i.e.*, cooking), providing unambiguous narratives that are easier to annotate and evaluate. Second, our dataset contains 150× the number of frames compared to the previous largest dataset, StoryStream. Third, we offer 5× denser textual descriptions, with an average of 763.8 words per video. These advantages make our dataset a better resource for narrative video generation.

Data Source	# Vid. (train/val)	# Clips	Clip Len.	# Clips / Vid.
YouCook2	1333 / 457	∼10K	19.6s	7.7
HowTo100M (subset)	30039 / 933	∼183K	9.5s	5.9

Table 1. **Long narrative dataset sources.** Our dataset is built upon Youcook2 and a cooking subset of Howto100M.

Datasets	Modality	Type	# Images	Text Length
Flintstones	Image	Comic	122k	86
Pororo	Image	Comic	74k	74
StorySalon	Image	Comic	160k	106
StoryStream	Image	Comic	258k	146
VIST	Image	Real world	210K	∼70
CookGen	Video	Real world	**39M**	**763.8**

Table 2. **Comparison with multi-modal narrative datasets.** Most existing datasets focus on image-based comic story generation. In contrast, our dataset consists of long narrative videos, containing 150× the number of frames and 5× the dense text annotations compared to the previous largest dataset, StoryStream.

3.2. Annotation and Processing

To ensure scalability and quality, we design an efficient annotation pipeline to support the annotation as below.

Captions. For open-source and scalability, we train a video captioner based on open-sourced VLM. Inspired by LLaVA-Hound [59], we begin by collecting a caption dataset using GPT-4o, with a focus on object attributes, subject-object interactions, and temporal dynamics. Subsequently, we fine-tune a captioning model based on LLaVA-NeXT [60] to optimize captioning performance.

Actions. We use HowTo100M ASR-based pseudo labels for 'actions' in each video, further refined by LLMs to provide enhanced annotations of the actions throughout the video [40]. This refinement improves the action quality to capture events and narrative context. However, the annotations are still noisy and sometimes not informative due to

the inherent errors in ASR scripts.

Caption-Action Matching and Filtering. To ensure alignment between captions and actions, we implement a matching process based on time intervals. Using Intersection-over-Union (IoU) as a metric, we evaluate whether the overlap between the captioned clip time and action time meets a threshold. An action is considered a match if the following conditions are met: the difference between the clip start time and the action start time (`start_diff`) is less than 5 seconds; the clip end time is later than the action end time; and the IoU between the clip and action time intervals is greater than 0.25, or if IoU>0.5. Here, `clip_time` and `action_time` represent the time intervals for the clip and action, respectively. Using this rule, we filter and match captions to actions, ensuring that each caption aligns with the relevant action. We found this step is important for creating narrative consistency throughout the video.

Annotation Quality Reverification. High-quality captions are essential for narrative visual generation. To verify the quality of our annotations, we build an evaluation pipline of inverse generation and visual understanding through VLM experts, which are detailed in Appendix §C.1 and §C.2.

4. Method

Given the text input, the task of long narrative video generation aims at generating a coherent long video $\mathcal{Y} \in \mathbb{R}^{H\times W\times F}$ that aligns with the progression of the text input sequentially. The H, W, and F are generated videos' height, width, and frame numbers. To achieve this, we propose **VideoAuteur**, which involves three main components: an interleaved long narrative video director, a rolling-context conditioned keyframe renderer, and a visual-conditioned video generation model. The long narrative video director creates a sequence of language states and visual embeddings to represent the narrative flow (§4.1). A pretrained DiT model then renders keyframes using a rolling history of reference images as contextual conditioning (§4.2). Finally, the video generation model produces video clips based on these visual conditions (§4.3).

4.1. Long Narrative Interleaved Director

As shown in Figure 3a, the long narrative video director generates a sequence of visual embeddings (or keyframes) that capture the narrative flow. The interleaved image-text director creates a sequence where text tokens and visual embeddings are interleaved, integrating narrative and visual content tightly. Using an auto-regressive model, it predicts the next token based on the accumulated context of both text and images. This helps maintain narrative coherence and align visuals with the text semantics.

Interleaved auto-regressive model. Our model performs next-token prediction for cross-modal generation, learning from sequences of interleaved image-text pairs with a context window size T. Each text token is supervised with cross-entropy loss, and the final visual embedding $\mathbf{z}_T$ is regressed using learnable query tokens, as illustrated in Figure 3b. The auto-regressive conditioning is given by:

$$p(\mathbf{y}_t \mid \mathbf{y}_{1:t-1}) = p(\mathbf{c}_t \mid \mathbf{c}_{1:t-1}) \cdot p(\mathbf{z}_t \mid \mathbf{c}_{1:t}, \mathbf{z}_{1:t-1}), \quad (1)$$

where $\mathbf{c}_t$ represents texts and $\mathbf{z}_t$ denotes visual embeddings.

Regression latent space. We utilize a CLIP-Diffusion visual autoencoder with a CLIP encoder E_{clip} and a diffusion decoder D_{diff} to encode raw images $\mathbf{x}$ to visual embeddings for auto-regressive generation:

$$\mathbf{z} = E_{\text{clip}}(\mathbf{x}), \quad \hat{\mathbf{x}} = D_{\text{diff}}(\mathbf{z}) \quad (2)$$

This setup generates language-aligned visual embeddings and reconstructs images from them.

Regression loss. To align the generated visual latents $\mathbf{z}_{\text{pred}}$ with the target latents $\mathbf{z}_{\text{target}}$, we use a combined loss:

$$L_{\text{reg}} = \alpha \left(1 - \frac{\mathbf{z}_{\text{pred}} \cdot \mathbf{z}_{\text{target}}}{\|\mathbf{z}_{\text{pred}}\| \|\mathbf{z}_{\text{target}}\|}\right) + \beta \frac{1}{N} \sum_{i=1}^{N} (\hat{z}_i - z_i)^2 \quad (3)$$

where α and β are hyper-parameters.

Narrative from "actions" to "visual states". The interleaved model generates a coherent narrative sequence by progressively conditioning each step on the cumulative context from previous steps, Figure 3b. At each time step t, the model generates an action $\mathbf{a}_t$, a caption $\mathbf{c}_t$, and a visual state $\mathbf{z}_t$, conditioned on the cumulative history $\mathcal{H}_{t-1}$:

$$\begin{aligned} \mathcal{H}_{t-1} &= \{\mathbf{a}_{1:t-1}, \mathbf{c}_{1:t-1}, \mathbf{z}_{1:t-1}\} \\ \mathbf{a}_t \mid \mathcal{H}_{t-1} \rightarrow \mathbf{c}_t \mid \{\mathcal{H}_{t-1}, \mathbf{a}_t\} &\rightarrow \mathbf{z}_t \mid \{\mathcal{H}_{t-1}, \mathbf{a}_t, \mathbf{c}_t\} \end{aligned} \quad (4)$$

This layered conditioning improves coherence across the sequence, aligning actions, language, and visuals.

4.2. Rolling Context Conditioned Render

While the interleaved auto-regressive director model can learn visual consistency, the CLIP representation space struggles to preserve fine visual details (*e.g.*, character features, clothing patterns), as demonstrated in Appendix Figure 16. To address this limitation and improve generation quality, we employ a pretrained Text-to-Image diffusion transformer model to render high-quality keyframes, conditioning on a rolling history of reference images. The context length can vary dynamically from 1 to 3, balancing flexibility and efficiency when generating keyframes.

As illustrated in Figure 3b, we use a rolling history of two reference images, I_t and I_{t-1}. This setup is further conditioned by the tiled global caption

$$\mathbf{c}_{\text{tiled}} = \text{tiled}(\mathbf{c}_{t-3}, \mathbf{c}_{t-2}, \mathbf{c}_{t-1}, \mathbf{c}_t), \quad (5)$$

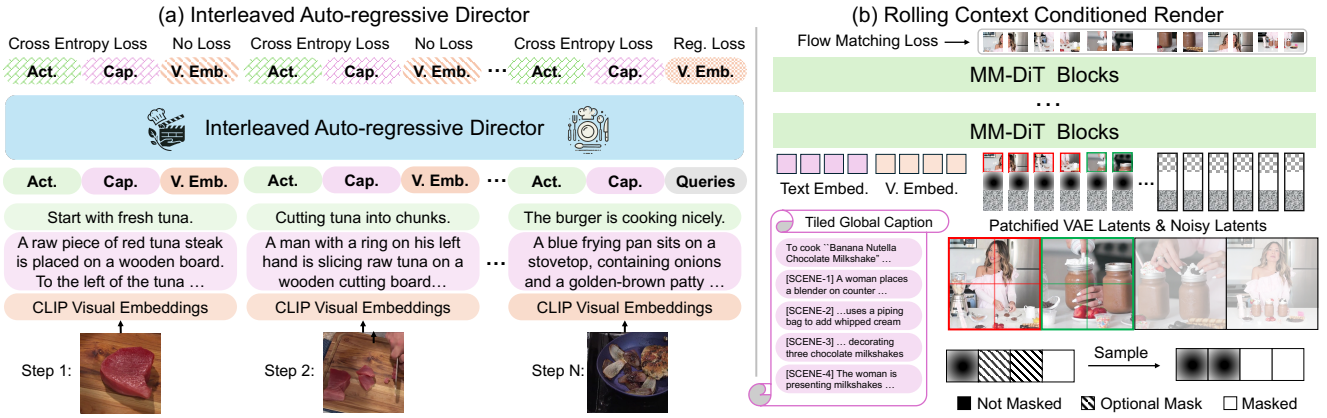

Figure 3. **Long Narrative Visual Condition Generation**. (a) *Interleaved Auto-regressive Director:* an auto-regressive vison-language model, takes a user query (e.g., "How to cook a tuna sandwich?") and an initial image-text pair as input. It then generates actions, captions, and visual states (*i.e.*, visual embeddings) step-by-step. (b) *Rolling Context Conditioned Render*: Apart from the semantics consistency through interleaved generation, we use a rolling of reference images as direct context conditions to further improve visual consistency with a diffusion transformer model. With them, a long narrative video can be created using these generated visual conditions (*i.e.*, visual embeddings and/or keyframes derived from the interleaved director and the keyframe render with rolling context conditioning.)

the predicted visual embeddings $\mathbf{z}_t$ and $\mathbf{z}_{t-1}$, as well as the reference images I_{t-3} and I_{t-2}.

$$\mathrm{D}(\mathbf{c}_{\text{tiled}},\, \mathbf{z}_{t-1},\, \mathbf{z}_t,\, I_{t-3},\, I_{t-2}) \rightarrow I_{t-1}, I_t, \tag{6}$$

where $\mathrm{D}(\cdot)$ denotes the diffusion model for synthesizing a new keyframe I_t by integrating the rolling context of images, captions, and visual embeddings. This layered conditioning improves coherence across frames.

Flow Matching Loss. We employ a flow matching loss that aligns the learned drift function f_θ with the ground-truth path from $\mathbf{x}_T$ to $\mathbf{x}_{T+1}$. We define:

$$\mathcal{L}_{\text{flow}}(\theta) \;=\; \mathbb{E}_{\mathbf{x}_T,\, \mathbf{x}_{T+1},\, T}\Big[\big\| f_\theta(\mathbf{x}_T, T) \;-\; \mathbf{v}(T)\big\|^2\Big], \tag{7}$$

where $\mathbf{v}(T)$ denotes the ideal drift path that transitions $\mathbf{x}_T$ towards $\mathbf{x}_{T+1}$. This objective enforces consistency across frames without relying on a separate diffusion loss.

4.3. Visual-Conditioned Video Generation

Using the sequence of actions $\mathbf{a}_t$, captions $\mathbf{c}_t$,visual states $\mathbf{z}_t$ and keyframe I_t generated by the interleaved director and rolling context conditioned render, we condition a video generation model to produce coherent long narrative videos. Unlike the classic Image-to-Video (I2V) pipeline that only uses an image as the starting frame, our approach leverages the regressed visual latents $\mathbf{z}_t$ as continuous conditions throughout the sequence (see §4.3.1). Furthermore, we improve the robustness and quality of the generated videos by adapting the model to handle noisy visual embeddings, since the regressed visual latents may not be perfect due to regression errors and keyframe uncertainty (see §4.3.2).

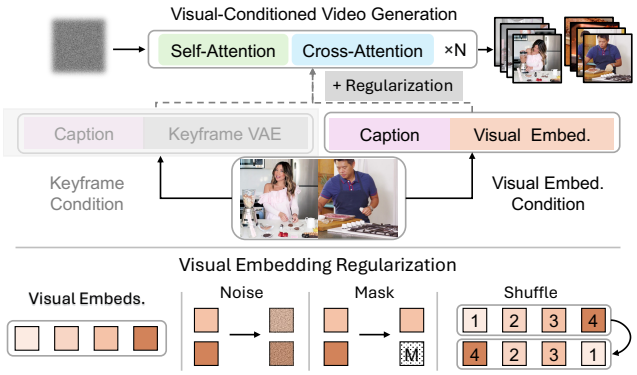

Figure 4. **Visual-conditioned video generation.** Our interleaved auto-regressive director and rolling context renderer generates both text and visual conditions, enabling the video generation process to be conditioned on keyframes (VAE embeddings) and CLIP latents. We apply Gaussian noise, random masking and random shuffling as regularization during the training process to improve robustness with the imperfect visual embeddings.

4.3.1 Visual Conditions Beyond Keyframes

Conventional visual-conditioned video generation typically uses initial keyframes to guide the model, where each frame $\mathbf{x}_t$ is generated as $\mathbf{x}_t = D_{\text{visual}}(I_t)$. Our interleaved auto-regressive director supports generating visual states $\mathbf{z}_t$ in a semantically aligned latent space, allowing direct conditioning from a pretrained visual decoder, as shown in Figure 4. By using these regressed visual latents $\mathbf{z}_t$ directly, each frame is generated as $\mathbf{x}_t = D_{\text{visual}}(\mathbf{z}_t)$. This follows the narrative and enhancing consistency by relying on narrative-aligned embeddings.

Figure 5. **Rolling Context Conditioned Render.** We integrate tiled global captions, predicted visual embeddings, and a rolling context of previous keyframes to render new keyframes throughout the narrative. By combining semantic conditioning from textual captions and CLIP embeddings with detailed information from VAE embeddings, the diffusion transformer maintains consistency in visual details such as clothing, food details, and character identities. Generated frames are highlighted with red edges.

4.3.2 Learning from Noisy Visual Conditions

To enhance the robustness over imperfect visual embeddings $\mathbf{z}_t$ from the auto-regressive director, we fine-tune the model using noisy embeddings $\mathbf{z}'_t$ defined by:

$$\mathbf{z}'_t = \mathcal{S}(\mathcal{M}(\mathbf{z}_t + \boldsymbol{\epsilon})) \tag{8}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 \mathbf{z}_t)$ represents Gaussian noise, $\mathcal{M}$ is a masking operator that sets a fraction of elements to zero, and $\mathcal{S}$ is a shuffling operator that permutes the order.

5. Experiments

5.1. Experimental Setup

Models. We initialize the auto-regressive model with [13], a pretrained 7B multi-modal LLM. We initialize the context conditioned render model with FLUX.1 Fill model [27]. For video generation, we employ a video generation model which has been pre-trained on large-scale video-text pairs and could accept both text and visual conditions.

Data. We use a total of ∼32K narrative videos for model training and another ∼1K videos for validation. All the videos are resized to 448 (short-side) and then center-cropped with 448×448 resolution.

Training & Evaluation. We train the interleaved auto-regressive director model for 5,000 steps by default. Training loss is a combination of cosine similarity loss and MSE loss for visual tokens and CrossEntropy loss for language tokens. For rolling context conditioned render, we use the flow matching loss following FLUX [27]. For visual-conditioned video generation, we use the diffusion loss following DiT [34] and Stable Diffusion 3 [10]. Narrative generation is mostly evaluated on the Youcook2 validation set because of the high-quality of action annotations and the Howto100M validation set is mostly used for data quality evaluation and I2V generation. Please refer to the appendix for implementation details.

Evaluation Metrics. The common metrics CLIP score [17] and FVD [49] are used to assess overall video quality, while the FID [18] score evaluates the quality of the generated keyframes. Additionally, when comparing to state-of-the-art baselines, human evaluation is used to assess generation aesthetics, realism, visual consistency across video clips, and the narrative score which reflects the coherence of the generated cooking steps, and if the cooking process has been successfully completed.

5.2. Rolling Context Conditioning

As detailed in Section 4.2, we leverage the in-context conditioning capabilities of the transformer architecture and adopt a rolling context conditioning strategy to enable DiT to render keyframes with superior visual consistency, while adhering to the extended narrative semantics produced by the interleaved auto-regressive director model. As shown in Figure 5, our keyframe renderer preserves fine visual details and exhibits high visual quality and aesthetics with the help of large-scale pretraining [27]. The reason is that the in-context conditioned VAE features could preserve visual details and the semantics are preserved through the auto-regressive model. Notably, the rolling context conditioning approach allows the renderer to strike a flexible balance between generation efficiency and visual consistency by dynamically adjusting the number of frames generated in each forward pass (*i.e.*, a dynamic number of frames).

5.3. Visual-Conditioned Video Generation

As detailed in Section 4.3, we fine-tune the model to be directly conditioned on the visual latents and generated by our interleaved director and keyframes generated by rolling-context renderer. Table 3 compares the keyframe-conditioned approach with our visual embedding-conditioned strategy. Our method improves CLIP-T [17] scores on both validation sets—from 25.9 to 26.4 on YouCook2 and from 26.6 to 27.3 on HowTo100M. Additionally, FVD scores decrease, indicating better video quality (557.7 vs. 512.6 on YouCook2, 541.1 vs. 520.7 on HowTo100M). Videos conditioned on visual embeddings

Figure 6. **Quality comparison on long narrative generation**. Here is a case with a narrative topic of *"Step-by-step guide to cooking blueberry muffins"*. Our interleaved director sequentially generates "actions," "captions," and image embeddings to construct a narrative on how to cook the dish step by step and then render keyframes. Our method shows state-of-the-art visual quality with superior consistency.

Visual Condition	YouCook2		HowTo100M	
	CLIP-T ↑	FVD ↓	CLIP-T ↑	FVD ↓
Keyframe	25.9	557.7	26.6	541.1
Embedding (w/o Reg.)	25.5	590.3	26.4	554.3
Embedding (w/ Reg.)	**26.4**	**512.6**	**27.3**	**520.7**

Table 3. **Visual-conditioned Video Generation with Regularization.** Evaluate CLIP-T and FVD scores for video generation conditioned on keyframes versus visual embeddings generated by our interleaved director with and without regularization.

Method	Prompting		Gen. Metric		Human Evaluation			
	Prompt Src.	Cond.	CLIP-T	FID	Aes.	Real.	Consist.	Narr.
SD-XL [35]	External	Text	27.1	-	4.0	2.9	3.3	N/A
FLUX.1-s [27]	External	Text	27.9	-	4.8	3.1	3.4	N/A
IC Lora [22]	External	Text	27.9	34.1	4.7	4.1	4.7	N/A
StoryDiffusion [64]	External	Text	25.9	36.4	3.9	2.9	3.7	N/A
Vlogger [65]	LLM	Text	25.5	45.5	4.0	2.4	3.1	3.7
Seed-Story [55]	Interleaved	V. Emb.	24.1	32.1	1.9	4.1	4.2	4.1
Ours (w.o RCC)	Interleaved	V. Emb.	26.1	**25.3**	2.1	4.3	4.5	4.4
Ours (w. RCC)	Interleaved	T+V. Emb.	**28.0**	29.4	**4.8**	**4.5**	**4.8**	**4.6**

Table 4. **Quantitative comparisons with metrics and human evaluation.** Each method is evaluated by both image generation metrics (CLIP-T and FID) and human ratings. Higher values indicate better performance for all human-evaluation metrics (5 tiers, from 1 to 5, higher is better). SD-XL and FLUX.1-s use narrative captions generated by our model and IC-Lora uses a tiled version. RCC: Rolling Context Conditioning. We use our generated narrative captions for the text-conditioned methods (row 1-5).

demonstrate higher semantic alignment and improved generation quality. We also provide qualitative samples on the demo page and in the appendix.

5.4. Comparisons of Long Narrative Generation

As most existing narrative generation methods [55, 64] only support image generation, we compare our model with state-of-the-art methods on the task of long narrative keyframe generation. We provide both quantitative comparisons in (§5.4.1) and qualitative comparisons (§5.4.2).

5.4.1 Long Narrative Keyframe Generation

We compare our method with leading narrative keyframe generation approaches, including IC Lora [22], StoryDiffusion [64], Vlogger [65], and Seed-Story [55], as well as a language-centric strategy that relies solely on captions (using models such as SD-XL [35] and FLUX.1-schnell [27]). Except for IC Lora and Seed-Story, which are fine-tuned on our CookGen dataset for two epochs, all other methods follow their official inference guidelines with the official checkpoints. As shown in Table 4, our approach achieves the highest generation scores, with a CLIP-Text score of **28.0** and an FID score of **25.3**. We also conduct a human evaluation (Table 4) using a five-tier rating scale, where higher is better. Our method attains top performance in aesthetics (**4.8** vs. **4.7**, IC Lora), realism (**4.5** vs. **4.1**, Seed-Story), and visual consistency (**4.8** vs. **4.7**, IC Lora), as well as the highest narrative score of **4.6**. These results demonstrate that our method achieves state-of-the-art performance

Loss Type		Training		Validation	
MSE	Cosine	L2 Dist ↓	Cosine Sim. ↑	CLIP-T ↑	FID ↓
✓	✗	**0.41**	0.82	23.6	31.9
✗	✓	1.1	0.82	24.1	32.1
✓	✓	**0.41**	**0.83**	**25.1**	**30.1**

Table 5. **Both scale and direction matter.** We track the training convergence and evaluate models with the CLIP-T and FID metrics on the validation set. The combination of both MSE loss and Cosine Similarity loss performs best on the validation metrics.

for long narrative generation.

5.4.2 Qualitative Comparisons

In Figure 6, we compare our method with state-of-the-art long narrative keyframe generation approaches, including StoryDiffusion, Vlogger, and Seed-Story, and observe that our results maintain superior visual quality and consistency. In particular, our keyframes balance realism with appealing aesthetics while preserving character identities and smooth transitions. In contrast, competing methods often exhibit color inconsistencies or lose track of concepts—Vlogger occasionally produces uneven color schemes between frames, StoryDiffusion can introduce visual confusion, and Seed-Story sometimes generates mismatched clothing across different scenes. This comparison aligns with the human evaluation results in Table 4, demonstrating our method achieves state-of-the-art performance for long narrative visual generation. The generated keyframes can be extended into full video clips with consistent visuals and coherent storytelling.

5.5. Ablation Studies

In this section, we ablate important designs in VideoAuteur, which improve the interleaved auto-regressive model and the visual-conditioned video generation model for interleaved narrative visual generation.

Latent scale and direction matter. To determine an effective supervision strategy for visual embeddings, we firstly test the robustness of the latents to pseudo regression errors by rescaling (multiplying by a factor) and adding random Gaussian noise. Figure 17 indicates that both scale and direction are critical in latent regression. Notably, rescaling primarily affects object shape while preserving key semantic information (*i.e*. object type and location), whereas adding noise drastically impacts reconstruction quality. As shown in Section 5.5, combining MSE loss (for scale) and cosine similarity (for direction) leads to the best generation quality, improving CLIP-T by 1.5 points and reducing FID by 1.8 points compared to using MSE alone.

From "Actions" to "Visual States". We also explore how different regression tasks influence the director's capability in narrative visual generation. Specifically, we compare various reasoning settings for the interleaved director, examining transitions from sequential actions to language states, and ultimately to visual embeddings. As shown in Table 6, a chain of reasoning that progresses from actions to language states and then to visual states proves effective for long narrative visual generation. This approach enhances both training convergence, achieving a lower L2 distance (**0.41** vs. **0.43**), and generation quality, reflected in a superior FID score of 25.3 (an improvement of **+0.8**).

Regression Task	Training		Validation	
	L2 Dist ↓	Cosine Sim. ↑	CLIP-T ↑	FID ↓
Action → Vis. Embed.	0.43	0.82	22.7	27.9
Caption → Vis. Embed.	0.41	0.82	25.7	26.1
Action → Caption → Vis. Embed.	**0.41**	**0.83**	**26.1**	**25.3**

Table 6. **From "Actions" to "Visual States".** We report the L2 distance and cosine similarity scores for tracking the training convergence and evaluate the generation images with CLIP score and FID score. Models are trained and evaluated on the collected Howto100M subset. SEED-X latent is used for visual regression.

Regularization Setting	CLIP-T ↑	FVD ↓
Naive Baseline	26.4	554.3
+Random Masking	26.9	539.7
+Random Gaussian. Noise	27.2	522.1
+Random Shuffling	**27.3**	**520.7**

Table 7. **Learn from Noisy Visual Conditions.** Our training regularization strategy enhances the robustness of the visual-conditioned video generation model. Specifically, we apply random masking and shuffling at a rate of 25%, and introduce Gaussian noise with 0.5 std of the embeddings of two thousand samples.

Learn from noisy visual conditions. Table 7 presents an ablation study examining the effect of robustness regularization on the visual-conditioned video generation model. We evaluate the generated videos using CLIP-T and FVD. The progressively improved results from 26.4 to 27.3 on CLIP-T and 554.3 to 520.7 on FVD demonstrate the effectiveness of our regularization strategy, which combines random masking, Gaussian noise, and shuffling.

6. Conclusion

In this paper, we tackle the challenges of generating long-form narrative videos and empirically evaluate its efficacy in the cooking domain. We curate and annotate a large-scale cooking video dataset, capturing clear and high-quality narratives essential for training and evaluation. Our proposed two-stage auto-regressive pipeline, which includes a long narrative director, a rolling context conditioned keyframe renderer and a visual-conditioned video generation model, demonstrates promising improvements in semantic and visual consistency in generated long narrative videos with an unified pipeline. Through experiments on our dataset, we observe enhancements in spatial and temporal coherence across video sequences. We hope our work can facilitate further research in long narrative video generation.

Acknowledgment This project is partially supported by ONR N00014-23-1-2641. We also thank Kelly Zhang, Ziyan Zhang and Jiaming Han for their fruitful discussion.

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. In *NeurIPS*, 2022. 2

[2] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators, 2024. 1, 2

[3] Jerome Bruner. The narrative construction of reality. *Critical inquiry*, 18(1):1–21, 1991. 1

[4] Duygu Ceylan, Chun-Hao P Huang, and Niloy J Mitra. Pix2video: Video editing using image diffusion. In *CVPR*, 2023. 1

[5] 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. In *arXiv*, 2023. 1, 2

[6] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *CVPR*, 2024. 1, 2

[7] Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization. In *CVPR*, 2024. 2

[8] Xi Chen, Zhiheng Liu, Mengting Chen, Yutong Feng, Yu Liu, Yujun Shen, and Hengshuang Zhao. Livephoto: Real image animation with text-guided motion control. In *ECCV*, 2025. 1

[9] 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. In *arXiv*, 2023. 2

[10] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*, pages 12606–12633. PMLR, 2024. 6

[11] Ruoyu Feng, Wenming Weng, Yanhui Wang, Yuhui Yuan, Jianmin Bao, Chong Luo, Zhibo Chen, and Baining Guo. Ccedit: Creative and controllable video editing via diffusion models. In *CVPR*, 2024. 1

[12] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. In *NeurIPS*, 2024. 2

[13] 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. In *arXiv*, 2024. 2, 6

[14] Tanmay Gupta, Dustin Schwenk, Ali Farhadi, Derek Hoiem, and Aniruddha Kembhavi. Imagine this! scripts to compositions to videos. In *Proceedings of the European conference on computer vision (ECCV)*, pages 598–613, 2018. 2

[15] Yuval Noah Harari. *Sapiens: A brief history of humankind*. Random House, 2014. 1

[16] Yingqing He, Menghan Xia, Haoxin Chen, Xiaodong Cun, Yuan Gong, Jinbo Xing, Yong Zhang, Xintao Wang, Chao Weng, Ying Shan, et al. Animate-a-story: Storytelling with retrieval-augmented video generation. *arXiv preprint arXiv:2307.06940*, 2023. 2

[17] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 6

[18] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 6

[19] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. In *arXiv*, 2022. 1, 2

[20] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. In *NeurIPS*, 2022. 1, 2

[21] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. In *arXiv*, 2022. 1, 2

[22] Lianghua Huang, Wei Wang, Zhi-Fan Wu, Yupeng Shi, Huanzhang Dou, Chen Liang, Yutong Feng, Yu Liu, and Jingren Zhou. In-context lora for diffusion transformers. *arXiv preprint arXiv:2410.23775*, 2024. 2, 7

[23] Nisha Huang, Yuxin Zhang, and Weiming Dong. Style-a-video: Agile diffusion for arbitrary text-based video style transfer. *IEEE Signal Processing Letters*, 2024. 1

[24] Ting-Hao Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. Visual storytelling. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 1233–1239, 2016. 2

[25] Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Grant Schindler, Rachel Hornung, Vighnesh Birodkar, Jimmy Yan, Ming-Chang Chiu, et al. Videopoet: A large language model for zero-shot video generation. In *ICML*, 2024. 1, 2

[26] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *CVPR*, 2023. 2

[27] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 6, 7

[28] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with

frozen image encoders and large language models. In *ICML*, 2023. 2

[29] Yanghao Li, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, and Kaiming He. Scaling language-image pre-training via masking. In *CVPR*, 2023. 2

[30] Han Lin, Abhay Zala, Jaemin Cho, and Mohit Bansal. Videodirectorgpt: Consistent multi-scene video generation via llm-guided planning. In *COLM*, 2024. 2

[31] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2024. 2

[32] Adyasha Maharana and Mohit Bansal. Integrating visuospatial, linguistic and commonsense structure into story visualization. *arXiv preprint arXiv:2110.10834*, 2021. 2

[33] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2630–2640, 2019. 2, 3

[34] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023. 1, 2, 6

[35] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 2, 7

[36] Lu Qi, Lehan Yang, Weidong Guo, Yu Xu, Bo Du, Varun Jampani, and Ming-Hsuan Yang. Unigs: Unified representation for image generation and segmentation. In *CVPR*, 2024. 2

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 2

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

[39] 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. In *NeurIPS*, 2022. 2

[40] Nina Shvetsova, Anna Kukleva, Xudong Hong, Christian Rupprecht, Bernt Schiele, and Hilde Kuehne. Howtocaption: Prompting llms to transform video annotations at scale. In *European Conference on Computer Vision*, pages 1–18. Springer, 2025. 3

[41] 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. In *arXiv*, 2022. 1, 2

[42] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. In *arXiv*, 2023. 2

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

[44] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. In *CVPR*, 2024. 1, 2, 23

[45] Changyao Tian, Xizhou Zhu, Yuwen Xiong, Weiyun Wang, Zhe Chen, Wenhai Wang, Yuntao Chen, Lewei Lu, Tong Lu, Jie Zhou, et al. Mm-interleaved: Interleaved image-text generative modeling via multi-modal feature synchronizer. In *arXiv*, 2024. 2

[46] Ye Tian, Ling Yang, Haotian Yang, Yuan Gao, Yufan Deng, Jingmin Chen, Xintao Wang, Zhaochen Yu, Xin Tao, Pengfei Wan, et al. Videotetris: Towards compositional text-to-video generation. *arXiv preprint arXiv:2406.04277*, 2024. 2

[47] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 2

[48] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 17

[49] 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:1812.01717*, 2018. 6, 18

[50] Chaoyang Wang, Xiangtai Li, Lu Qi, Henghui Ding, Yunhai Tong, and Ming-Hsuan Yang. Semflow: Binding semantic segmentation and image synthesis via rectified flow. In *NeurIPS*, 2024. 2

[51] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. In *arXiv*, 2023. 1, 2

[52] Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. In *Advances in Neural Information Processing Systems*, 2024. 2

[53] 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. In *arXiv*, 2024. 1, 2

[54] Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. Magicanimate: Temporally consistent human image animation using diffusion model. In *CVPR*, 2024. 1, 2

[55] 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. In *arXiv*, 2024. 2, 7

[56] 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

[57] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. In *arXiv*, 2024. 1, 2

[58] Xuanyu Yi, Zike Wu, Qingshan Xu, Pan Zhou, Joo-Hwee Lim, and Hanwang Zhang. Diffusion time-step curriculum for one image to 3d generation. In *CVPR*, 2024. 2

[59] Ruohong Zhang, Liangke Gui, Zhiqing Sun, Yihao Feng, Keyang Xu, Yuanhan Zhang, Di Fu, Chunyuan Li, Alexander Hauptmann, Yonatan Bisk, et al. Direct preference optimization of video large multimodal models from language model reward. *arXiv preprint arXiv:2404.01258*, 2024. 3, 18, 20

[60] Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model, 2024. 3

[61] Canyu Zhao, Mingyu Liu, Wen Wang, Jianlong Yuan, Hao Chen, Bo Zhang, and Chunhua Shen. Moviedreamer: Hierarchical generation for coherent long visual sequence. *arXiv preprint arXiv:2407.16655*, 2024. 2

[62] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models. In *arXiv*, 2022. 1, 2

[63] Luowei Zhou, Chenliang Xu, and Jason J Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018. 2, 3

[64] Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. Storydiffusion: Consistent self-attention for long-range image and video generation. *arXiv preprint arXiv:2405.01434*, 2024. 2, 7

[65] Shaobin Zhuang, Kunchang Li, Xinyuan Chen, Yaohui Wang, Ziwei Liu, Yu Qiao, and Yali Wang. Vlogger: Make your dream a vlog. In *CVPR*, 2024. 2, 7