Generating Long-take Videos via Effective Keyframes and Guidance

Hsin-Ping Huang[1,2] Yu-Chuan Su[1] Ming-Hsuan Yang[1,2]
[1]Google DeepMind [2]University of California, Merced

Abstract

We tackle the challenge of generating long-take videos encompassing multiple non-repetitive yet coherent events. Existing approaches generate long videos conditioned on single input guidance, often leading to repetitive content. To address this problem, we develop a framework that uses multiple guidance sources to enhance long video generation. The main idea of our approach is to decouple video generation into keyframe generation and frame interpolation. In this process, keyframe generation focuses on creating multiple coherent events, while the frame interpolation stage generates smooth intermediate frames between keyframes using existing video generation models. A novel mask attention module is further introduced to improve coherence and efficiency. Experiments on challenging real-world videos demonstrate that the proposed method outperforms prior methods by up to 9.5% in objective metrics.

1. Introduction

Recent advancements in text-to-video generation have shown significant progress in enhancing visual fidelity, including improvements in resolution, frame rate, and video length [3, 9, 10, 26, 28, 29, 38, 46, 71–73, 75, 80, 90]. These models excel in generating impressive visuals within a scene but face challenges when it comes to constructing complex sequences involving multiple events, due to several limitations. First, they produce video clips of a predefined fixed length, which is significantly shorter than real-world videos due to substantial computational and memory overhead. Second, the input conditions, such as a class label [13, 22, 49, 51, 53, 54, 61, 63, 64, 69] or a text prompt describing the entire video [4, 25, 31, 32, 34, 39, 56, 66, 77, 78, 92], often lack complexity and richness. However, real-world videos encompass diverse content that surpasses the capability of encoding within such simplistic conditions. Thus, optimizing the video generation model alone cannot resolve these limitations and close the gap between generated and real-world videos. For example, generating a real-world video as shown in Fig. 1 requires creating a meaningful storyline with different events, such as cutting food, moving a plate, and putting it into the fridge. To achieve this goal, it is necessary to develop a video generation system that creates the high-level contents of multiple events and employs existing models to create detailed frames for each event.

Several methods address the challenge of generating long videos. Autoregressive approaches [24, 85] create video clips sequentially, with each frame building upon previously generated ones. Latent-based methods [57, 91] utilize a shared latent content vector across all frames. While successfully generating lengthy videos, these techniques often produce homogeneous content featuring natural scenes and recurring human actions. Video generation conditioned on multiple guidance sources [2, 24, 30, 84, 89] generate diverse events but are typically limited to synthetic environments and cartoons. Text-to-video models successfully handle transitions within videos [11, 52], but efficiently prompting the model to generate long videos across multiple clips remains an open problem, often requiring extensive trial-and-error and manual post-processing. Recently, multi-shot video generation [43, 44, 93] has been introduced to synthesize videos consisting of multiple distinct snippets lacking direct temporal continuity between shots. However, generating a *long-take single-shot* video featuring *multiple semantically different events*, as illustrated in Fig. 1, remains a challenging problem that requires further exploration.

We tackle the challenges of generating long-take videos encompassing multiple real-world events. To achieve this, guidance signals provided at multiple timesteps are essential for effectively controlling the generation process. We utilize various levels of guidance, including object labels as high-level guidance that describe objects in the generated video and image layouts as low-level guidance. The core concept of our approach is to divide the video generation process into keyframe generation and frame interpolation. Keyframe generation focuses on creating multiple coherent frames with varied video content, each representing key events specified by the guidance. The frame interpolation stage then generates smooth intermediate frames to connect these keyframes using existing video models. Our method is orthogonal to existing efforts to optimize the generation quality of video models and is compatible with different video generation models by incorporating them into our system. To ensure temporal consistency in the gener-

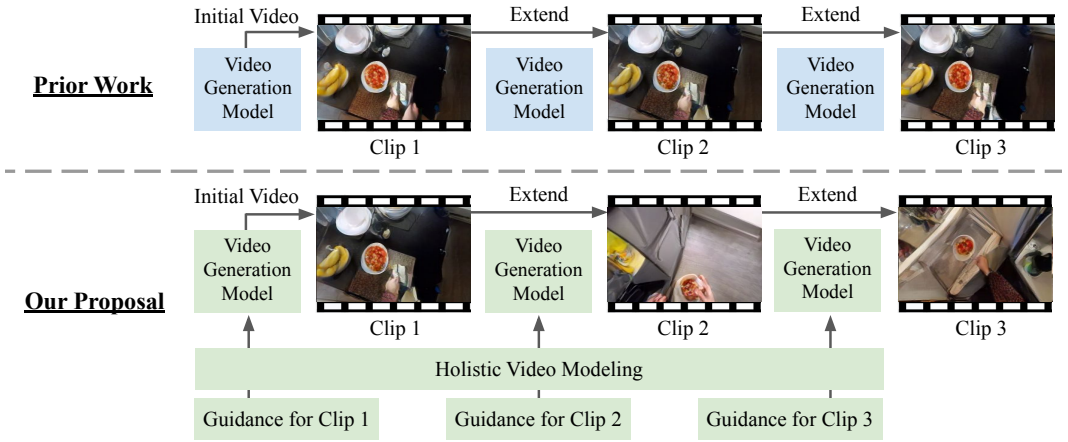

Figure 1. Existing schemes [24] generate long videos by iteratively extending the generated frames, often leading to videos with repetitive patterns. In contrast, we generate *long-take videos encompassing multiple non-repetitive yet coherent events* using multiple guidances.

ated videos, we introduce a novel masked attention module to enhance efficiency and object coherence during long-take video generation. We conduct extensive experiments on real-world datasets, demonstrating that our approach outperforms state-of-the-art video generation models by 9.5% on the FVD metric. Our main contributions are as follows:

- We address the problem of generating long-take videos with multiple non-repetitive yet coherent events.
- We propose a multi-stage approach involving the generation and the connection of keyframes.
- We introduce a masked attention module to enhance continuity and efficiency for long-take videos.

2. Related Work

Video Generation. GAN-based methods [6, 13, 53, 54, 58, 62, 63, 69] have demonstrated early success in generating short video clips, while the generation quality decreases significantly when applied to long videos. Recently, diffusion models [33, 68, 83] have shown promising visual quality in video generation. Auto-regressive models [1, 36, 76] have benefited from developments in vector quantization [21, 65] and transformer [17] models, allowing them to predict discrete tokens in the latent space [8, 24, 41, 50, 82, 87] with impressive visual quality. Despite the recent success in video generation, these models mostly focus on generating videos with a predefined length, often shorter than real videos, or are limited to synthesizing long videos with repetitive contents [24, 67], such as human actions [5, 55, 60] and sky timelapse [81] under the setting of unconditional video generation [6, 57, 85, 91]. Few works explore video generation conditioned on input guidance at multiple timesteps [2, 24, 30, 84, 89], but they are limited to synthetic environments. In contrast, our work aims to generate real-world videos. Most recently, text-to-video models [3, 4, 9, 10, 26, 28, 28, 29, 31, 32, 34, 38, 46, 56, 71–73, 75, 75, 77, 79, 80, 90] have been developed to generate videos given natural language inputs. However, the generated videos are limited in length and complexity due to computation constraints and limitations in text description. Instead of trying to increase the capacity of existing models, we focus on an alternative approach to improving video generation, namely generating videos with multiple events. Our approach is generic, and off-the-shelf video generation models can be adopted as a component in our pipeline. Concurrent video generation works address different topics such as video-to-video translation [14, 27, 70, 86], generating style or static-dynamic transitions [52], and creating artificial transitions between distinct scenes [11]. In contrast, we focus on long-take video generation from high-level guidance, connecting keyframes by natural transitions to model semantic changes (e.g., moving to different rooms, picking up objects) in real videos.

Story visualization. Story visualization [42, 47, 48, 59] aims to synthesize a sequence of images that depict a story composed of multiple sentences. GeNeVA [15, 19, 23, 45] is a conditional text-to-image generation task developed on the CoDraw [40] dataset, which addresses the problem of constructing a scene iteratively based on a sequence of descriptions. However, this line of work primarily focuses on experiments conducted with synthetic data and targets the generation of images. In contrast, our focus lies in generating videos using real-world data.

3. Method

3.1. Overview

We first provide an overview of our long-take video generation system and then describe the three stages in detail. Given a series of N guidance signals and a reference frame I_0 as input, our objective is to synthesize a video V that encompasses the content specified by the guidance, starting from the initial frame I_0. The input guidance consists of

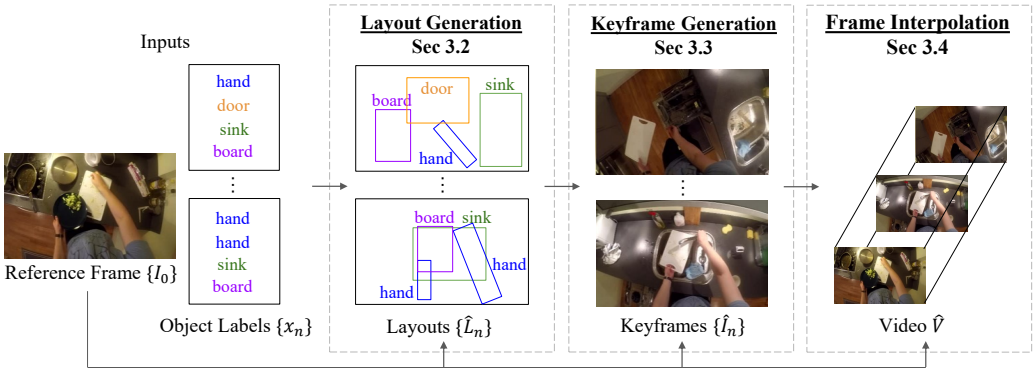

Figure 2. **Approach overview.** The proposed method consists of three stages: 1) *Layout generation* predicts a series of layouts from the object label sets. 2) *Keyframe generation* predicts a sequence of keyframes for the video using the layout sequences and the reference frame as inputs. 3) *Frame interpolation* synthesizes intermediate frames based on the keyframe sequence to obtain the complete video.

object labels $x_1, x_2, \cdots, x_N$. Each set x_n contains k^n object labels, which may vary across timesteps n, indicating the objects appearing in the video. To address the challenge of generating videos covering multiple events, our approach decouples the video generation process into layout generation, keyframe generation, and frame interpolation, as shown in Fig. 2. Layout generation explicitly predicts the structures of keyframes from object labels, addressing the deterioration of structure prediction over time during long video generation. This layout can also be provided and edited by users. Keyframe generation focuses on jointly modeling the entire video to create multiple coherent events, with each keyframe representing core events specified by the guidance. Using existing video models, the frame interpolation stage then synthesizes smooth intermediate frames between keyframes. Specifically, our system produces a series of N keyframes in the video based on the object label sets. Each keyframe serves as the initial and final frames of short video clips. An existing video generation model predicts intermediate frames between keyframes to complete the video.

Our approach offers several advantages: First, it simplifies generating long videos by predicting keyframes that encapsulate high-level event concepts. Second, it leverages advanced video generation models to generate intermediate frames, enhancing flexibility in video synthesis. We utilize generative transformer models capable of handling various modalities by converting them into a unified vocabulary. This capability allows for incorporating different levels of input guidance (e.g., labels and layouts) in our task. Transformers are adept at tasks such as interpolation and frame prediction, making them versatile across our multi-stage approach. They also offer significant speed advantages and achieve competitive output quality compared to diffusion models [88].

3.2. Layout Generation

We first generate a series of layouts from the object label sets to explicitly constrain the structures of keyframes. Given a series of N object label sets $\{x_n\}$ and a reference frame I_0, we synthesize a series of layouts $\{L_n\}$ representing the layouts of the N keyframes in the video. Since the reference first frame I_0 is given, we assume L_0 is known and can be used as a reference layout. We define the layout L as a set of bounding boxes with a variable length k^n, i.e., $\{b_1, b_2, \cdots, b_{k^n}\}$. The bounding box attributes include its object label, x-coordinate, y-coordinate, width, and height, i.e., $b = \{c, x, y, w, h\}$. We apply tokenization to encode the layouts $\{L_n\}$ and the object label set $\{x_n\}$ into discrete token sequences, where the values of the bounding box attributes are discretized by uniform quantization.

Given the tokens of the reference layout L_0 and the object labels c as inputs, we replace the other attributes of the bounding boxes $\{x, y, w, h\}$ with a `[MASK]` token and train the model to predict the masked bounding box attributes in the layout series. Our layout generation module predicts multiple layouts jointly to preserve the temporal consistency of the generated layout sequence. Although few datasets contain annotations of labels and layouts, we utilize an off-the-shelf segmentation model to obtain annotations through an automatic pipeline, which is readily scalable to general-domain real-world datasets.

3.3. Keyframe Generation

Next, we generate a sequence of keyframes from the predicted layout sequences. Given a reference first frame I_0 and a series of layouts $\{\hat{L}_n\}$ either provided by users or synthesized in the previous stage, our goal is to generate a sequence of keyframes $\{\hat{I}_n\}$. Specifically, the keyframe sequence $\{I_n\}$ is encoded into visual tokens, $\{e_n\}$, by a VQVAE [21] model. During training, the tokens of layouts and keyframes are concatenated as $s = \{L_0, \hat{L}_1, \cdots, \hat{L}_N, e_0, e_1, \cdots, e_N\}$ in dataset $\mathcal{D}$. We ran-

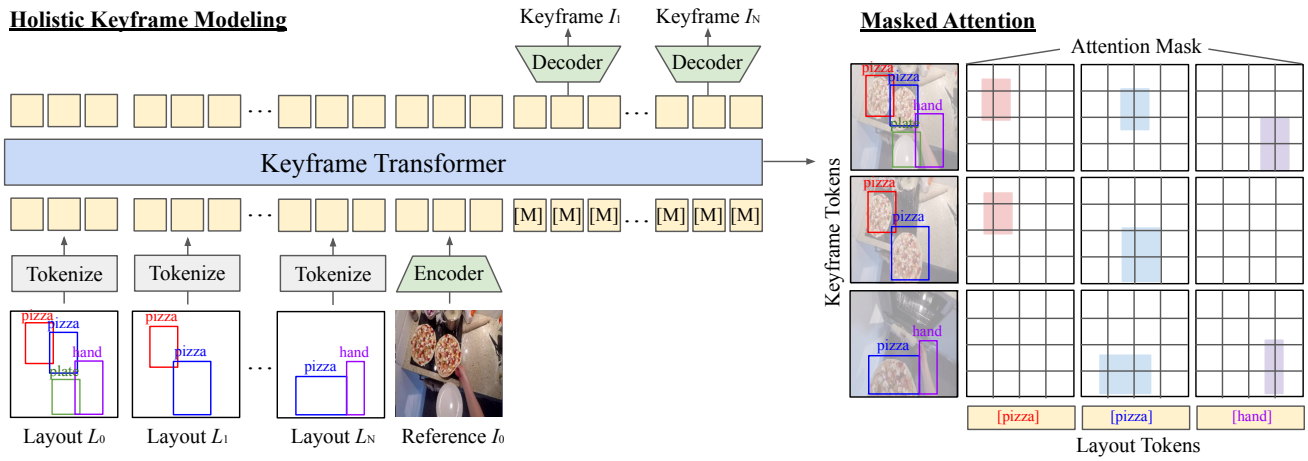

Figure 3. **Keyframe generation.** Our key ideas are: 1) *Holistic keyframe modeling*: the model predicts multiple keyframes jointly to maintain coherence, using the layouts and the reference frame as inputs. 2) *Masked attention*: the attention mask between irrelevant keyframe and layout tokens is set to zero, encouraging the model to focus on relevant tokens, thus improving coherence and efficiency. Tokens of a specific color in a bounding box are associated with layout tokens of the same color, indicating they represent the same object.

domly replace tokens in the sequence with the `[MASK]` token, obtaining the masked sequence s_M. The generative transformer model is trained to predict the masked tokens of the target keyframes $s_t, t \in M$, by minimizing the negative log-likelihood:

$$\mathcal{L} = - \mathop{\mathbb{E}}_{s \in \mathcal{D}} \sum_{t \in M} \log p(s_t \mid s_M) \tag{1}$$

At test time, given the tokens of the layout sequence and the first frame $\{\hat{L}_n\}$ and e_0, the model predicts the tokens of the subsequent keyframes $\{\hat{e}_n\}$, which are then reconstructed into raw images $\{\hat{I}_n\}$ by the decoder of VQVAE, as depicted in Fig. 3. Our keyframe generator predicts all frames holistically throughout the entire video, enhancing consistency across keyframes and the final long video.

Masked attention. We further propose a novel masked attention module to improve the efficiency and coherence of keyframe generation. Since our input to the keyframe generator is a concatenation of layout and keyframe tokens, each token attends to every other token without considering the relationship between the layout and the keyframe. However, this approach may be redundant because the visual tokens within a specific bounding box in the keyframe ($\hat{e}_n$) are highly related to the layout tokens of the corresponding bounding box in $\hat{L}_n$. Moreover, since we generate a series of keyframes jointly, the tokens of each bounding box in a single layout $\hat{L}_n$ are correlated with every relevant visual token in all the keyframes $\{e_n\}$, as long as the bounding box represents the same instance in the sequence. We encode this prior knowledge into our model through the attention mask, setting irrelevant attention to zero to guide the model to focus on relevant tokens. This prior knowledge helps the model respect the relationship between layout and keyframe to improve their alignment and temporal coherence for the same instance. Additionally, masked attention reduces computational costs and improves efficiency in handling long keyframe sequences, as shown in Tab. 4.

3.4. Frame Interpolation

Finally, given the reference frame I_0 and a sequence of generated keyframes $\{\hat{I}_n\}$, we input the keyframes into an existing video generation model [4, 24, 87] trained for infilling tasks to generate intermediate frames between the keyframes. Specifically, we use a model that predicts intermediate frames based on the initial and final keyframes as input. We first convert the video into discrete tokens using a 3D-VQVAE model. Then, a transformer model is employed to predict masked video tokens for the intermediate frames. Finally, the interpolated video tokens are decoded back into raw videos by the 3D-VQVAE decoder.

During inference, given two consecutive keyframes $\hat{I}_{n-1}$ and $\hat{I}_n$, the model predicts video token sequences that connect these keyframes, denoted as $\hat{z}_n$. N clips of video tokens $\{\hat{z}_n\}$ are predicted and concatenated into a sequence to generate the long video using the 3D decoder. Since our keyframe generation module has already generated consistent objects over a longer duration, it facilitates the frame interpolation stage to maintain object consistency and temporal coherence when predicting intermediate frames. Our frame interpolation module can also be adapted to models trained on open-domain data, such as [4, 28, 32], to enhance the quality and diversity of generated videos.

4. Experiments

Dataset. We experiment with two real-world video datasets: 1) `EPIC Kitchen` [16] comprises egocentric

Table 1. **Quantitative results of video generation.** Metrics are averaged across frames. Our method consistently outperforms state-of-the-art video generation models.

(a) EPIC Kitchen

Methods	FVD ↓	LPIPS ↓	AP ↑	CLIP ↑	Cons. ↑
TATS	737.6	0.615	0.072	0.2452	0.9971
StyleGAN-V	581.8	0.550	0.074	0.2257	0.9960
MAGVIT (frame pred.)	291.4	0.475	0.133	0.2785	0.9965
MAGVIT (class cond.)	285.6	0.469	0.140	0.2788	0.9968
Ours	**258.4**	**0.421**	**0.192**	0.2855	**0.9973**
Ours (GT layouts)	214.7	0.346	0.390	**0.2866**	0.9970
Ours (GT keyframes)	174.1	0.242	0.451	0.2865	0.9972

(b) nuScenes

Methods	FVD ↓	LPIPS ↓	AP ↑
MAGVIT (frame pred.)	171.1	0.478	0.044
Ours	**158.9**	**0.448**	**0.059**
Ours (GT layouts)	106.0	0.384	0.137
Ours (GT keyframes)	84.5	0.306	0.219

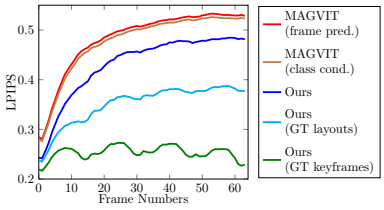

Figure 4. **Per-frame results.** The results of MAGVIT degrade over time, while our method shows slower quality degradation.

Table 2. **User study.** We report the percentage of raters who prefer our method based on video quality and fidelity to the ground truth.

Methods	Quality	Reproduction
Ours *vs.* MAGVIT (frame pred.)	76.3%	82.1%
Ours *vs.* MAGVIT (class cond.)	68.4%	66.7%

videos of kitchen activities. Unlike commonly used datasets such as [18, 37, 60], EPIC Kitchen videos feature complex and dynamic scenes. Foreground objects move in and out of the camera's field of view, while background scenes and camera viewpoints frequently change, including transitions between different rooms. To synthesize such videos, the video generation model needs to generate multiple non-repetitive events within a short time window, such as various actions performed by the subject. 2) `nuScenes` [7] consists of driving videos that capture both object motion (e.g., moving cars) and dynamics caused by egocentric motion, where background scenes change over time.

Evaluation metrics. We evaluate the generated videos using the following metrics: 1) `FVD` and `FID` assess the quality of generated videos by measuring the feature distribution compared to real videos. 2) `LPIPS` assesses the similarity between the generated and ground truth video frames. 3) `AP` measures the semantic correctness by comparing the detected objects between the generated and ground truth frames. 4) `CLIP` computes the CLIP similarity of the generated frame and a text prompt constructed from the input labels. 5) `Cons.` assesses frame consistency using the CLIP image similarity between two frames [20, 74].

Implementation details. We utilize a segmentation model [12], not specifically trained on our datasets, to automatically extract sparse labels without human effort. The model covers common objects and is readily extendable to various real-world video datasets in general domains. These labels describe both foreground objects and dynamic background scenes (e.g., walls, tables, floors). We set the maximum number of objects to 14 and pad sequences if fewer than 14 objects are detected. We sample one keyframe every 16 frames, resulting in 16×N frames per video. Our quantitative evaluation uses 64-frame videos at 128×128 resolution, with longer videos of up to 512 frames included in visual results. Inference costs are 0.01 seconds for layout generation, 1.5 seconds for keyframe generation, and 37 fps for frame interpolation, achieving an inference speed of 25 fps, which is 10 times faster [87] than autoregressive and diffusion models.

4.1. Video Generation

We evaluate video generation results to verify the effectiveness of additional guidance and the capability of generating multiple coherent events in the videos.

Baselines. We compare with the following state-of-the-art video generation methods: 1) `MAGVIT (frame pred.)` [87]: given the reference frame as input, we apply MAGVIT to generate a 16-frame clip. We then use the last predicted frames iteratively to generate the entire video. This follows the sliding window approach for long-take video generation. 2) `MAGVIT (class cond.)`: we condition MAGVIT on the reference frame and the object label. This extends the sliding window approach to incorporate additional guidance, similar to our method. 3) `StyleGAN-V` [57]: we train their unconditional model to generate a 64-frame video and apply the GAN projection method to condition on the reference frame. 4) `TATS` [24]: similar to MAGVIT, we generate a 16-frame clip and use a sliding window approach. 5) `Ours (GT layouts)` and `Ours (GT keyframes)`: to understand how the quality of layout and keyframe generation affects video results, we compare two variants of our method that use the ground truth layouts and keyframes as inputs.

Quantitative results. Tab. 1 shows our method achieves substantial improvements of 9.5% and 7.6% over baselines in FVD. The videos closely resemble ground truth, confirmed by LPIPS. AP and CLIP scores validate our model's success in generating videos with specified object inputs. Consistency scores verify maintained coherency, demonstrating our approach's efficacy. `MAGVIT (class cond.)` outperforms `MAGVIT (frame pred.)`, highlighting the benefits of additional guidance. Further, our approach surpasses `MAGVIT (class cond.)`, showing the combined effectiveness of additional guidance and our system, which generates videos by predicting

Figure 5. **Video generation on EPIC Kitchen (top) and nuScenes (bottom).** TATS, StyleGAN-V, and MAGVIT predict relatively static videos with repetitive patterns and quality degradation when inferring videos beyond the model's output length. In contrast, our method generates videos with non-recurrent events, including scene changes, object insertion, and deletion, maintaining meaningful content.

keyframes and infilling the intermediate frames. `Ours (GT layouts)` and `Ours (GT keyframes)` significantly enhance video quality, suggesting room for improvement in long-take video generation despite using the same video generation model. This underscores the importance of developing video generation systems alongside enhancing base model quality. Our multi-stage approach, generating videos via keyframes and guidance, complements existing efforts and improves addressing real-world video generation challenges. Our method also enables user refinement of intermediate representations, like detailed layouts, to enhance generated videos interactively.

The per-frame LPIPS scores are shown in Fig. 4. The X-axis represents the frame number, ranging from 0 to 64, while the Y-axis shows the LPIPS score averaged across all test videos for a specific frame. Image quality degrades rapidly in MAGVIT, particularly when generating videos beyond the length of the training clip (i.e., 16 frames). In contrast, our approach exhibits slower quality degradation, highlighting its effectiveness in generating videos covering diverse content.

User study. We conduct a user study to complement the quantitative evaluation. In this study, participants are presented with two videos generated by different methods alongside the ground truth video. They are then asked to evaluate the videos based on two criteria: 1) `Quality`: which video has better visual quality, and 2) `Reproduction`: which video better reproduces the con-

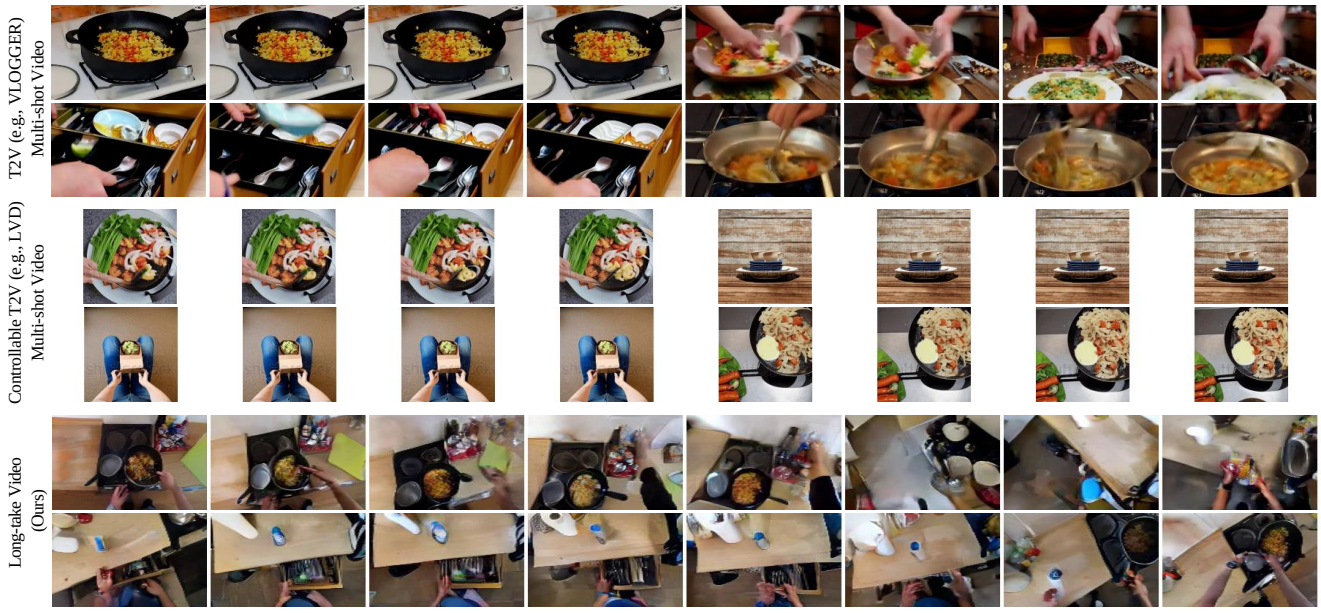

Figure 6. **Comparison to concurrent works.** Our method generates a single-shot long-take video with multiple non-recurrent events, such as cooking, obtaining condiments, and opening drawers. Video frames are presented sequentially from left to right and top to bottom. In contrast, multi-scene video generation VLOGGER and LVD [43, 44, 93], creates multiple snippets without direct temporal continuity.

tent of the ground truth video. The study includes 40 videos and 11 participants using the EPIC Kitchen dataset. The results, presented in Tab. 2, align with the quantitative findings and further validate the significant performance gain of our method compared to the baselines.

Qualitative results. In Fig. 5, `TATS`, `StyleGAN-V`, and `MAGVIT` tend to produce videos with homogeneous content, with quality deteriorating for long sequences. In contrast, our method generates videos featuring scene changes (e.g., from wall to table), object deletions (e.g., plate at the bottom), and object insertions (e.g., left/right hand), demonstrating the model's ability to generate videos with multiple events. Additionally, `Ours` generates videos that match the label sets but with different layouts than `Ours (GT layouts)`, validating our model's capability to generate videos satisfying different levels of input guidance, i.e., object label sets or (more constrained) layouts. Overall, we achieve satisfying results for the challenging video generation problem, where objects move in and out of camera views with significant motion and changing scenes.

Comparison to concurrent works. T2V and controllable T2V models create impressive visuals within a scene but typically produce short clips (e.g., 32 frames) and struggle with complex sequences. VLOGGER [93] and VideoDirectorGPT [44] extend T2V models to generate longer videos, but focus on generating multiple snippets without direct temporal continuity in the adjacent shots. We address the challenge of generating single-shot, long-take videos up to 512 frames. By applying LVD [43] and VLOGGER [93], we generate long videos using a series of text prompts. In Fig. 6, LVD and VLOGGER create multiple distinct shots, while our model produces a long-take video with non-recurrent events (e.g., cooking, fetching condiments, opening drawers). Our method complements multi-scene T2V models and can be integrated into VLOGGER's pipeline to create long takes as part of multiple snippets.

Table 3. **Quantitative results of keyframe generation.** Metrics are averaged across the keyframes. Our method consistently outperforms alternative approaches.

Methods	FID ↓	LPIPS ↓
MaskGIT	46.9	0.633
HCSS	50.2	0.653
Ours	**29.9**	**0.548**
Ours (GT)	24.2	0.416
Ours single (GT)	27.5	0.480

Limitations. The maximum video length our system generates is constrained by the number of keyframes, and the output length of the interpolation model. To generate longer videos, our approach can be integrated with multi-step inference processes to iteratively generate keyframes conditioned on preceding keyframes. As shown in Fig. 6, up to 512 frames are generated using 32 keyframes. In addition, the video tokenizer from MAGVIT limits the proposed method, and using a better tokenizer such as MAGVITv2 could reduce flickering artifacts. We plan to train our models on caption-video data and swap the backbone to T2V models to generate videos in broader domains.

Figure 7. **Qualitative results of keyframe generation.** MaskGIT predicts similar content, while HCSS fails to generate consistent frames. Our superior results validate the importance of providing additional input guidance and modeling the keyframes holistically.

4.2. Keyframes Generation

We evaluate our keyframe generation model to assess the importance of jointly generating all keyframes and the effectiveness of masked attention and layout generation.

Baselines. We compare the following methods: 1) `MaskGIT` [8]: the model takes the reference as input and iteratively predicts the next keyframe, representing keyframe generation without input guidance. 2) `HCSS` [35]: the model takes a single layout as input and generates a single keyframe, representing independent keyframe generation without full video modeling. 3) `Ours`: our keyframe generation given the predicted layouts as inputs. 4) `Ours (GT)`: our keyframe generation using the ground truth layouts (upper bound). 5) `Ours single (GT)`: an iterative approach that predicts each keyframe conditioned on the previous keyframe and the ground truth layout, representing generation without full video modeling.

Quantitative results. In Tab. 3, our method consistently outperforms `HCSS`, confirming the importance of joint prediction for all keyframes. `MaskGIT` tends to predict repetitive keyframes with minimal changes across frames, suggesting that the resulting video may appear static and unsuitable for video generation, highlighting the importance of guidance. We also compare different variants of our method. The superior performance of `Ours (GT)` over `Ours single (GT)` further verifies that a model considering the entire video jointly leads to better keyframe generation.

Qualitative results. Fig. 7 shows that `MaskGIT` tends to generate keyframes with similar content, validating the importance of providing input guidance at multiple timesteps.

Table 4. **Ablation study.** We validate the effectiveness of the proposed masked attention module and layout generation stage in synthesizing high-quality and accurate keyframes.

(a) Masked attention

Methods	5 objects FID ↓	5 objects LPIPS ↓	5 objects AP ↑	14 objects FID ↓	14 objects LPIPS ↓	14 objects AP ↑
Ours w/o mask	33.6	0.584	0.183	34.8	0.575	0.216
Ours	**31.3**	**0.565**	**0.192**	**29.9**	**0.548**	**0.223**

(b) Layout generation

Methods	FID ↓	LPIPS ↓	AP ↑
Ours w/o layout	60.3	0.672	0.095
Ours	**29.9**	**0.548**	**0.223**

In contrast, `HCSS` fails to generate consistent results across keyframes, as evidenced by variations in sink styles. Similarly, the iterative approach `Ours single (GT)` exhibits inconsistencies compared to `Ours (GT)`. These examples emphasize the significance of joint modeling for the entire video.

Ablation study. We conduct ablation studies on the masked attention module and the layout generation process. The proposed masked attention module enhances coherence and efficiency in generating long videos. We analyze the module's impact using different numbers of input objects. As shown in Tab. 4 (a), removing the masked attention module and allowing each layout token to attend to every keyframe token leads to a decrease in performance. With masked attention, our model scales effectively to videos with up to 14 objects, generating keyframes that better align with the reference frame and contain more details, resulting in improved FID and LPIPS scores. The improvements in AP score further confirm its effectiveness in concentrating generation within specific bounding box areas.

The layout generation process helps preserve the 2D frame structure, which can easily deteriorate over time during long video generation. In Tab. 4 (b), we conduct an ablation study using object labels instead of the generated layouts as inputs to the keyframe generator. The FID and LPIPS scores increase without the layouts, validating the importance of the layout generation process.

5. Conclusions

This work addresses the challenge of generating long-take videos that encompass multiple events. We introduce a video generation system that utilizes multiple guidance inputs to control the generation process. A multi-stage approach has been developed to generate core events through keyframes and connect these events using existing video generation models. Empirical results validate the effectiveness of our approach, emphasizing the importance of guidance in improving video generation quality.

References

[1] Mohammad Babaeizadeh, Mohammad Taghi Saffar, Suraj Nair, Sergey Levine, Chelsea Finn, and Dumitru Erhan. Fitvid: Overfitting in pixel-level video prediction. *arXiv preprint arXiv:2106.13195*, 2020. 2

[2] Amir Bar, Roei Herzig, Xiaolong Wang, Gal Chechik, Trevor Darrell, and Amir Globerson. Compositional video synthesis with action graphs. In *ICML*, 2021. 1, 2

[3] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 1, 2

[4] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *CVPR*, 2023. 1, 2, 4

[5] Navaneeth Bodla, Gaurav Shrivastava, Rama Chellappa, and Abhinav Shrivastava. Hierarchical video prediction using relational layouts for human-object interactions. In *CVPR*, 2021. 2

[6] Tim Brooks, Janne Hellsten, Miika Aittala, Ting-Chun Wang, Timo Aila, Jaakko Lehtinen, Ming-Yu Liu, Alexei A Efros, and Tero Karras. Generating long videos of dynamic scenes. In *NeurIPS*, 2022. 2

[7] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. *arXiv preprint arXiv:1903.11027*, 2019. 5

[8] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T. Freeman. Maskgit: Masked generative image transformer. In *CVPR*, 2022. 2, 8

[9] 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. *arXiv preprint arXiv:2401.09047*, 2024. 1, 2

[10] Shoufa Chen, Mengmeng Xu, Jiawei Ren, Yuren Cong, Sen He, Yanping Xie, Animesh Sinha, Ping Luo, Tao Xiang, and Juan-Manuel Perez-Rua. Gentron: Delving deep into diffusion transformers for image and video generation. *arXiv preprint arXiv:2312.04557*, 2023. 1, 2

[11] Xinyuan Chen, Yaohui Wang, Lingjun Zhang, Shaobin Zhuang, Xin Ma, Jiashuo Yu, Yali Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Seine: Short-to-long video diffusion model for generative transition and prediction. *arXiv preprint arXiv:2310.20700*, 2023. 1, 2

[12] Bowen Cheng, Alexander G. Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. In *NeurIPS*, 2021. 5

[13] Aidan Clark, Jeff Donahue, and Karen Simonyan. Adversarial video generation on complex datasets. *arXiv preprint arXiv:1907.06571*, 2019. 1, 2

[14] Nathaniel Cohen, Vladimir Kulikov, Matan Kleiner, Inbar Huberman-Spiegelglas, and Tomer Michaeli. Slicedit: Zero-shot video editing with text-to-image diffusion models using spatio-temporal slices. In *ICML*, 2024. 2

[15] Gaoxiang Cong, Liang Li, Zhenhuan Liu, Yunbin Tu, Weijun Qin, Shenyuan Zhang, Chengang Yan, Wenyu Wang, and Bin Jiang. Ls-gan: Iterative language-based image manipulation via long and short term consistency reasoning. In *ACM MM*, 2022. 2

[16] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *IJCV*, 130:33–55, 2022. 4

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019. 2

[18] Frederik Ebert, Chelsea Finn, Alex X. Lee, and Sergey Levine. Self-supervised visual planning with temporal skip connections. In *CoRL*, 2017. 5

[19] Alaaeldin El-Nouby, Shikhar Sharma, Hannes Schulz, Devon Hjelm, Layla El Asri, Samira Ebrahimi Kahou, Yoshua Bengio, and Graham W. Taylor. Tell, draw, and repeat: Generating and modifying images based on continual linguistic instruction. In *ICCV*, 2019. 2

[20] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *ICCV*, 2023. 5

[21] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, 2021. 2, 3

[22] Chelsea Finn, Ian Goodfellow, and Sergey Levine. Unsupervised learning for physical interaction through video prediction. In *NeurIPS*, 2016. 1

[23] Tsu-Jui Fu, Xin Wang, Scott Grafton, Miguel Eckstein, and William Yang Wang. SSCR: Iterative language-based image editing via self-supervised counterfactual reasoning. In *EMNLP*, 2020. 2

[24] Songwei Ge, Thomas Hayes, Harry Yang, Xi Yin, Guan Pang, David Jacobs, Jia-Bin Huang, and Devi Parikh. Long video generation with time-agnostic vqgan and time-sensitive transformer. In *ECCV*, 2022. 1, 2, 4, 5

[25] 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 *ICCV*, 2023. 1

[26] 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. *arXiv preprint arXiv:2305.10474*, 2024. 1, 2

[27] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. In *ICLR*, 2024. 2

[28] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 1, 2, 4

[29] Agrim Gupta, Lijun Yu, Kihyuk Sohn, Xiuye Gu, Meera Hahn, Li Fei-Fei, Irfan Essa, Lu Jiang, and José Lezama.

Photorealistic video generation with diffusion models. *arXiv preprint arXiv:2312.06662*, 2023. 1, 2

[30] William Harvey, Saeid Naderiparizi, Vaden Masrani, Christian Weilbach, and Frank Wood. Flexible diffusion modeling of long videos. In *NeurIPS*, 2022. 1, 2

[31] 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. 1, 2

[32] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 1, 2, 4

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

[34] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. In *ICLR*, 2023. 1, 2

[35] Manuel Jahn, Robin Rombach, and Björn Ommer. High-resolution complex scene synthesis with transformers. In *CVPRW*, 2021. 8

[36] Nal Kalchbrenner, Aäron van den Oord, Karen Simonyan, Ivo Danihelka, Oriol Vinyals, Alex Graves, and Koray Kavukcuoglu. Video pixel networks. In *ICML*, 2017. 2

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

[38] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. *ICCV*, 2023. 1, 2

[39] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *ICCV*, 2023. 1

[40] Jin-Hwa Kim, Nikita Kitaev, Xinlei Chen, Marcus Rohrbach, Byoung-Tak Zhang, Yuandong Tian, Dhruv Batra, and Devi Parikh. CoDraw: Collaborative drawing as a testbed for grounded goal-driven communication. In *ACL*, 2019. 2

[41] Xiang Kong, Lu Jiang, Huiwen Chang, Han Zhang, Yuan Hao, Haifeng Gong, and Irfan Essa. Blt: Bidirectional layout transformer for controllable layout generation. In *ECCV*, 2022. 2

[42] Yitong Li, Zhe Gan, Yelong Shen, Jingjing Liu, Yu Cheng, Yuexin Wu, Lawrence Carin, David Carlson, and Jianfeng Gao. Storygan: A sequential conditional gan for story visualization. In *CVPR*, 2019. 2

[43] Long Lian, Baifeng Shi, Adam Yala, Trevor Darrell, and Boyi Li. Llm-grounded video diffusion models. In *ICLR*, 2024. 1, 7

[44] Han Lin, Abhay Zala, Jaemin Cho, and Mohit Bansal. Videodirectorgpt: Consistent multi-scene video generation via llm-guided planning. *arXiv preprint arXiv:2309.15091*, 2023. 1, 7

[45] Zhenhuan Liu, Jincan Deng, Liang Li, Shaofei Cai, Qianqian Xu, Shuhui Wang, and Qingming Huang. Ir-gan: Image manipulation with linguistic instruction by increment reasoning. In *ACM MM*, 2020. 2

[46] Zhengxiong Luo, Dayou Chen, Yingya Zhang, Yan Huang, Liang Wang, Yujun Shen, Deli Zhao, Jingren Zhou, and Tieniu Tan. Videofusion: Decomposed diffusion models for high-quality video generation. *arXiv preprint arXiv:2303.08320*, 2023. 1, 2

[47] Adyasha Maharana and Mohit Bansal. Integrating visuospatial, linguistic and commonsense structure into story visualization. In *EMNLP*, 2021. 2

[48] Adyasha Maharana, Darryl Hannan, and Mohit Bansal. Improving generation and evaluation of visual stories via semantic consistency. In *NAACL*, 2021. 2

[49] Michael Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. In *ICLR*, 2015. 1

[50] Guillaume Le Moing, Jean Ponce, and Cordelia Schmid. CCVS: Context-aware controllable video synthesis. In *NeurIPS*, 2021. 2

[51] Junhyuk Oh, Xiaoxiao Guo, Honglak Lee, Richard Lewis, and Satinder Singh. Action-conditional video prediction using deep networks in atari games. In *NeurIPS*, 2015. 1

[52] Haonan Qiu, Menghan Xia, Yong Zhang, Yingqing He, Xintao Wang, Ying Shan, and Ziwei Liu. Freenoise: Tuning-free longer video diffusion via noise rescheduling. *arXiv preprint arXiv:2310.15169*, 2023. 1, 2

[53] Masaki Saito, Eiichi Matsumoto, and Shunta Saito. Temporal generative adversarial nets with singular value clipping. In *ICCV*, 2017. 1, 2

[54] Masaki Saito, Shunta Saito, Masanori Koyama, and Sosuke Kobayashi. Train sparsely, generate densely: Memory-efficient unsupervised training of high-resolution temporal gan. *IJCV*, 2020. 1, 2

[55] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. In *NeurIPS*, 2019. 2

[56] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data. In *ICLR*, 2023. 1, 2

[57] Ivan Skorokhodov, Sergey Tulyakov, and Mohamed Elhoseiny. Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2. In *CVPR*, 2022. 1, 2, 5

[58] Ivan Skorokhodov, Sergey Tulyakov, and Mohamed Elhoseiny. Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2. In *CVPR*, 2022. 2

[59] Yun-Zhu Song, Zhi-Rui Tam, Hung-Jen Chen, Huiao-Han Lu, and Hong-Han Shuai. Character-preserving coherent story visualization. In *ECCV*, 2020. 2

[60] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 2, 5

[61] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhutdinov. Unsupervised learning of video representations using LSTMs. In *ICML*, 2015. 1

[62] Yu Tian, Jian Ren, Menglei Chai, Kyle Olszewski, Xi Peng, Dimitris N. Metaxas, and Sergey Tulyakov. A good image generator is what you need for high-resolution video synthesis. In *ICLR*, 2021. 2

[63] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. MoCoGAN: Decomposing motion and content for video generation. In *CVPR*, 2018. 1, 2

[64] Joost van Amersfoort, Anitha Kannan, Marc'Aurelio Ranzato, Arthur Szlam, Du Tran, and Soumith Chintala. Transformation-based models of video sequences. *arXiv preprint arXiv:1701.08435*, 2017. 1

[65] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *NeurIPS*, 2017. 2

[66] Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual description. In *ICLR*, 2023. 1

[67] Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and D. Erhan. Phenaki: Variable length video generation from open domain textual description. In *ICLR*, 2023. 2

[68] Vikram Voleti, Alexia Jolicoeur-Martineau, and Christopher Pal. Mcvd: Masked conditional video diffusion for prediction, generation, and interpolation. In *NeurIPS*, 2022. 2

[69] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. In *NeurIPS*, 2016. 1, 2

[70] Fu-Yun Wang, Wenshuo Chen, Guanglu Song, Han-Jia Ye, Yu Liu, and Hongsheng Li. Gen-l-video: Multi-text to long video generation via temporal co-denoising. *arXiv preprint arXiv:2305.18264*, 2023. 2

[71] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023. 1, 2

[72] Weimin Wang, Jiawei Liu, Zhijie Lin, Jiangqiao Yan, Shuo Chen, Chetwin Low, Tuyen Hoang, Jie Wu, Jun Hao Liew, Hanshu Yan, Daquan Zhou, and Jiashi Feng. Magicvideo-v2: Multi-stage high-aesthetic video generation. *arXiv preprint arXiv:2401.04468*, 2024. 1, 2

[73] Wenjing Wang, Huan Yang, Zixi Tuo, Huiguo He, Junchen Zhu, Jianlong Fu, and Jiaying Liu. Videofactory: Swap attention in spatiotemporal diffusions for text-to-video generation. *arXiv preprint arXiv:2305.10874*, 2024. 1, 2

[74] 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. In *NeurIPS*, 2023. 5

[75] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *arXiv preprint arXiv:2309.15103*, 2023. 1, 2

[76] Dirk Weissenborn, Oscar Täckström, and Jakob Uszkoreit. Scaling autoregressive video models. In *ICLR*, 2020. 2

[77] 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. 1, 2

[78] Chenfei Wu, Jian Liang, Lei Ji, Fan Yang, Yuejian Fang, Daxin Jiang, and Nan Duan. Nüwa: Visual synthesis pre-training for neural visual world creation. In *ECCV*, 2022. 1

[79] Chenfei Wu, Jian Liang, Lei Ji, Fan Yang, Yuejian Fang, Daxin Jiang, and Nan Duan. Nüwa: Visual synthesis pre-training for neural visual world creation. In *ECCV*, 2022. 2

[80] Zhen Xing, Qi Dai, Han Hu, Zuxuan Wu, and Yu-Gang Jiang. Simda: Simple diffusion adapter for efficient video generation. In *CVPR*, 2024. 1, 2

[81] Wei Xiong, Wenhan Luo, Lin Ma, Wei Liu, and Jiebo Luo. Learning to generate time-lapse videos using multi-stage dynamic generative adversarial networks. In *CVPR*, 2018. 2

[82] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*, 2021. 2

[83] Ruihan Yang, Prakhar Srivastava, and Stephan Mandt. Diffusion probabilistic modeling for video generation. *arXiv preprint arXiv:2203.09481*, 2022. 2

[84] Shengming Yin, Chenfei Wu, Huan Yang, Jianfeng Wang, Xiaodong Wang, Minheng Ni, Zhengyuan Yang, Linjie Li, Shuguang Liu, Fan Yang, Jianlong Fu, Ming Gong, Lijuan Wang, Zicheng Liu, Houqiang Li, and Nan Duan. NUWA-XL: Diffusion over diffusion for eXtremely long video generation. In *ACL*, 2023. 1, 2

[85] Jaehoon Yoo, Semin Kim, Doyup Lee, Chiheon Kim, and Seunghoon Hong. Towards end-to-end generative modeling of long videos with memory-efficient bidirectional transformers. In *CVPR*, 2023. 1, 2

[86] Sunjae Yoon, Gwanhyeong Koo, Geonwoo Kim, and Chang D. Yoo. Frag: Frequency adapting group for diffusion video editing. In *ICML*, 2024. 2

[87] Lijun Yu, Yong Cheng, Kihyuk Sohn, José Lezama, Han Zhang, Huiwen Chang, Alexander G. Hauptmann, Ming-Hsuan Yang, Yuan Hao, Irfan Essa, and Lu Jiang. Magvit: Masked generative video transformer. *arXiv preprint arXiv:2212.05199*, 2022. 2, 4, 5

[88] Lijun Yu, José Lezama, Nitesh B. Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Agrim Gupta, Xiuye Gu, Alexander G. Hauptmann, Boqing Gong, Ming-Hsuan Yang, Irfan Essa, David A. Ross, and Lu Jiang. Language model beats diffusion – tokenizer is key to visual generation. *arXiv preprint arXiv:2310.05737*, 2023. 3

[89] Wei Yu, Wenxin Chen, Songhenh Yin, Steve Easterbrook, and Animesh Garg. Modular action concept grounding in semantic video prediction. In *CVPR*, 2022. 1, 2

[90] David Junhao Zhang, Jay Zhangjie Wu, Jia-Wei Liu, Rui Zhao, Lingmin Ran, Yuchao Gu, Difei Gao, and

Mike Zheng Shou. Show-1: Marrying pixel and latent diffusion models for text-to-video generation. *arXiv preprint arXiv:2309.15818*, 2023. 1, 2

[91] Qihang Zhang, Ceyuan Yang, Yujun Shen, Yinghao Xu, and Bolei Zhou. Towards smooth video composition. In *ICLR*, 2023. 1, 2

[92] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*, 2022. 1

[93] Shaobin Zhuang, Kunchang Li, Xinyuan Chen, Yaohui Wang, Ziwei Liu, Yu Qiao, and Yali Wang. Vlogger: Make your dream a vlog. *arXiv preprint arXiv:2401.09414*, 2024. 1, 7