

MEVG: Multi-event Video Generation with Text-to-Video Models

Gyeongrok Oh¹, Jaehwan Jeong¹, Sieun Kim¹, Wonmin Byeon²,
Jinkyu Kim¹, Sungwoong Kim¹, and Sangpil Kim¹

¹Korea University ²NVIDIA

Abstract. We introduce a novel diffusion-based video generation method, generating a video showing multiple events given multiple individual sentences from the user. Our method does not require a large-scale video dataset since our method uses a pre-trained diffusion-based text-to-video generative model without a fine-tuning process. Specifically, we propose a last frame-aware diffusion process to preserve visual coherence between consecutive videos where each video consists of different events by initializing the latent and simultaneously adjusting noise in the latent to enhance the motion dynamic in a generated video. Furthermore, we find that the iterative update of latent vectors by referring to all the preceding frames maintains the global appearance across the frames in a video clip. To handle dynamic text input for video generation, we utilize a novel prompt generator that transfers coarse text messages from the user into the multiple optimal prompts for the text-to-video diffusion model. Extensive experiments and user studies show that our proposed method is superior to other video-generative models in terms of temporal coherency of content and semantics. Video examples are available on our project page: <https://kuai-lab.github.io/eccv2024mevg>.

Keywords: Multi-event video generation · Training-free · Diffusion model

1 Introduction

Deep generative models in the computer vision community gain a significant spotlight due to their unprecedented performance. Especially, text-to-image generation model (T2I) [7, 13, 23, 31–33] has successfully produced high-quality images with complex text descriptions. However, the complexity of spatial-temporal relations for modeling motion dynamics, light conditions, and scene transitions for video generation with deep generative models requires huge computational resources and large-scale text-video paired datasets. Despite these challenges, recent methods [2, 15, 20–22, 24, 34] for text-to-video generation achieve data and cost-efficient training by leveraging the pre-trained text-to-image generative models [7, 32]. Although spatial-temporal modeling aided by prior knowledge from text-image pairs helps to generate high-quality frames and capture semantically complex descriptions, it falls short in addressing real-world video comprehensively.

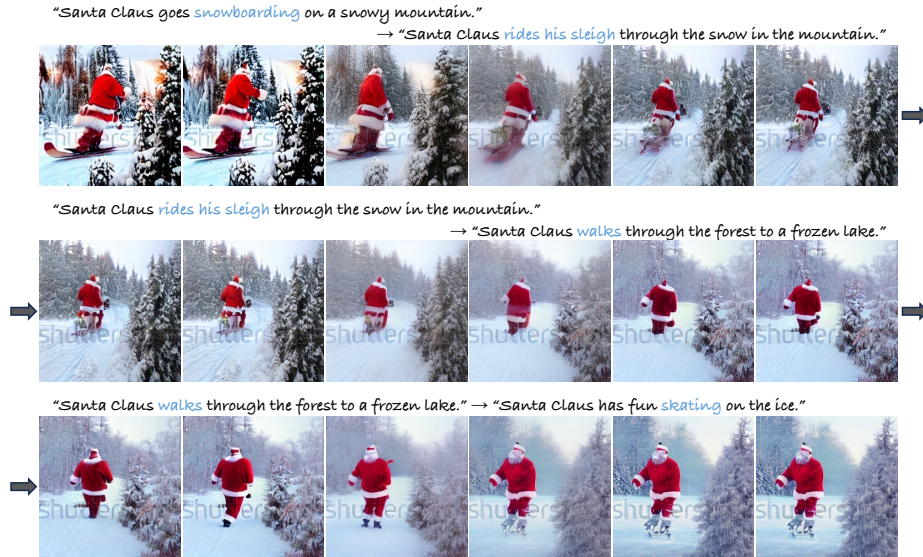


Fig. 1: An example of multi-event video generation. MEVG produces impressive output that corresponds to the given prompts and consists of chronologically continuous events.

In essence, videos in the wild consist of consecutive events with dynamic movements, backgrounds, objects, and viewpoint changes over time. However, existing approaches mainly generate a video with a single prompt that disregards semantic transitions from event to event and restrictively expresses the entire story when the story consists of multiple events. Multi-event-based video generation requires three significant criteria: 1) smooth transition between each video clip, 2) semantic alignment between the prompt from the user and the generated video, and 3) ensuring diversity of the content and motion in the video.

Recently, some studies [10, 38] have made progress in embracing the multiple descriptions that contain chronological sequence events. Despite this breakthrough, substantial training efforts are necessary using extensive text-video datasets because they present the additional networks to make multi-prompt video generation. The concurrent work [41] leverages the pre-trained text-to-video generation model (T2V). However, the overlapped denoised process on the consecutive distinct prompts induces visual degradation and significant inconsistency between the background and objects. Moreover, traditional long-term T2V methods [2, 10, 21, 34, 50] hierarchically generate a video by bridging the gap between each keyframe, followed by generating keyframes given a single description. This hierarchical video generation process makes it challenging to incorporate the multiple-time variant prompts since it comprehensively generates the global video content in the beginning.

Our method, multi-event video generation method (MEVG), is delicately designed to generate a video clip consisting of multiple events without any

video data nor fine-tuning process. MEVG successively enforces the temporal coherence between independently generated video clips given multiple event descriptions. By building upon the publicly released diffusion-based video generation model, we effectively utilize the pre-trained single-prompt T2V generative model¹ to generate complex scenario videos. Specifically, to preserve the visual coherence between time variable prompts in video generation while producing realistic and diverse motion, we introduce two novel techniques: a last frame-aware latent vector initialization method and a structure-guided sampling strategy for the diffusion-based generative model. First, the last frame-aware latent vector initialization stage includes (i) *dynamic noise*, which diversifies motion across frames, and (ii) *last frame-aware inversion*, which guides to generate consistent contents between prompts. *Structure-guided sampling* further improves the visual consistency by progressively updating the latent code during the sampling process. In addition, to incorporate sequentially structured prompts, we leverage the Large Language Model (LLM) as a prompt generator. Since a single story has sequentially incorporated events within one sentence, LLM separates a complex story into multiple prompts, each having only one event.

From our extensive experiments and user studies, we demonstrate that our proposed methods generate realistic videos that include three representative types of change: object motion, background, and complex content changes. Moreover, we examine the effectiveness and legitimacy of each proposed method by conducting ablation studies. To summarize, our main contributions are as follows:

- Our proposed diffusion-based video generation method generates a video consisting of multiple events without requiring any training or additional video data.
- We present a last-frame aware initialization method and dynamic noise adjustment strategy for the latent vector that enhances temporal and semantic consistency between individual videos where each video shows a different event.
- We present a novel prompt generator that transforms course text inputs into optimal text instructions for a text-to-video generative model, ensuring coherence of semantic transitions in the generated video.
- We show that our proposed video generation method outperforms the other zero-shot video generation methods in reflecting multiple events while maintaining visually coherent content for video generation.

2 Related Work

Text-to-Video Generation. Text-to-video (T2V) generation has shown remarkable progress. Three primary methodologies are utilized in the field of computer vision. A Generative Adversarial Network (GAN) [3, 35, 37, 40, 42, 49] is

¹ We used [15]. From our experiments, our proposed method can be applied to any kind of diffusion-based T2V pre-trained model.

a well-known algorithm to generate diverse video from a noise vector utilizing a generator and discriminator. Another approach is auto-regressive transformers [10, 21, 43–45, 48] that leverage discrete representation to depict the motion dynamics. Recently, diffusion-based methods [1, 2, 5, 9, 11, 15, 17, 19, 34, 39, 51] have shown significant progress in learning data distribution while iteratively removing noise from the initial gaussian noise.

Long-term video generation has recently been a popular topic in the computer vision community. Auto-regressive approaches [10, 21, 26] leveraging transformer architecture show plausible results in long-term video generation. However, they require massive training costs and datasets. Furthermore, although TATS [10] and Phenaki [38] can generate videos driven by a sequence of prompts, accumulated errors over time cause drastic changes in video content and visual quality degradation due to the auto-regressive property. Several works [2, 11, 34, 39, 50] based on the diffusion model leverage temporal interpolation networks and masked strategies for generating smoother videos. VidRD [12] directly utilizes the previous initial latent code to expand the video.

Note that most previous works focused on video generation from a single prompt or an event. However, in this work, we tackle the multi-event video generation task, which consists of consecutive events in a long-term video. Animate-A-Story [14] utilizes abundant real-world video corresponding to each story for natural motion. Moreover, SEINE [6] focuses on the transition from short to long video generation models by utilizing millions of datasets. Gen-L-Video [41], another approach, uses overlapped frames between two successive prompts. Although this strategy makes the outcomes more realistic, undesirable contents occur due to the overlapping denoising process.

Zero-shot approach. FateZero [29] and INFUSION [25] edit video by leveraging the pre-trained image diffusion models while ensuring temporal consistency. These methods utilize attention maps and spatial features to preserve the structure and temporal coherence over the frame. In the generation field, Text2Video-Zero [24] synthesizes the video to keep the global structure across the sequence of frames without video data while encoding the motion dynamics to provide diverse movement. To encode the movement, latent codes enclose the motion dynamics with direction parameters. Free-bloom [22] and DirecT2V [20] are distinct approaches employing the text-to-image models while sharing a similar conceptual framework. Both utilize the Large Language Model (LLM) in order to maintain the semantic information for each generated frame. DirecT2V leverages self-attention to preserve the appearance of the video; in addition, Free-bloom leverages joint distribution to sample the initial code for consistent frames.

Inspired by these approaches, we leverage a pre-trained text-to-video (T2V) generation model to extend a short, monotonous video into an exciting video containing variable events. To pursue the naturalness of results, challenges exist in maintaining visual coherence and guaranteeing diversity. Therefore, we adjust the latent code near the preceding video and grant dynamic changes through gradual perturbation.

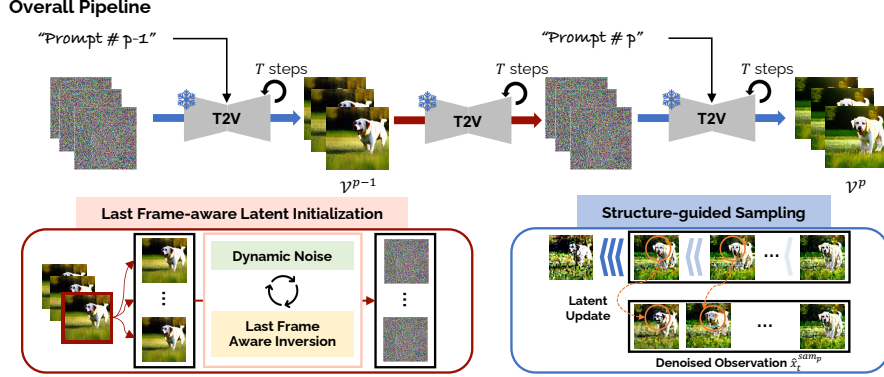


Fig. 2: MEVG synthesizes the consecutive video clips corresponding to distinct prompts. The overall pipeline comprises two major components: last frame-aware latent initialization and structure-guided sampling. First, in the last frame-aware latent initialization, the pre-trained text-to-video generation model adopts the repeated frame as an input to invert into the initial latent code with two novel techniques: *dynamic noise* and *last frame-aware inversion*. Second, *structure-guided sampling* enforces continuity within a video clip by updating the latent code.

3 Method

We propose a novel pipeline to generate a temporally and semantically coherent video conditioned on multiple event-based prompts. Specifically, our goal is to generate multiple video clips without disturbing the natural flow and recurrent video pattern across the semantic transitions in the given prompts. In this section, we first provide the introductory diffusion model that is the basis for our research and an overview of our proposed pipeline (see Sec. 3.1 and Sec. 3.2). Next, we present the technical details of our two main components: (i) *last frame-aware latent initialization*, (ii) *structure-guided sampling*, in Sec. 3.3 and Sec. 3.4. Finally, we introduce the prompt generator, harnessing the powerful ability of Large Language Model (LLM) to handle a complex story containing multiple meaningful events (Sec. 3.5).

3.1 Preliminaries

DDPM. The diffusion probabilistic model [18] has two components: a forward diffusion process and a backward diffusion process. In the forward process, data distribution transforms into noise distribution by adding noise iteratively. For every timestep t , noise $\epsilon \sim N(0, I)$ diffract the original data x utilizing the variance schedule β_t as follows:

$$x_t = \sqrt{\alpha_t}x + \sqrt{1 - \alpha_t}\epsilon, \quad (1)$$

where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{i=0}^t \alpha_i$. In training time, the diffusion model predicts noise during every step to reconstruct the original data distribution. This reverse process $q(x_{t-1}|x_t)$ is parameterized as follows:

$$p_\theta(x_{t-1}|x_t) := \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)). \quad (2)$$

DDIM. DDIM [36] is a variant of DDPM that leverages the non-Markovian manner instead of the Markov chain. DDIM sampling strategy makes the diffusion process to be deterministic, which can be written as follows:

$$\begin{aligned} x_{t-1} = & \sqrt{\bar{\alpha}_{t-1}} \left(\underbrace{\frac{x_t - \sqrt{1 - \bar{\alpha}_t} \epsilon_\theta(x_t, t)}{\sqrt{\bar{\alpha}_t}}}_{\hat{x}_t} \right) \\ & + \sqrt{1 - \bar{\alpha}_{t-1} - \sigma_t^2} \underbrace{\epsilon_\theta(x_t, t)}_{\epsilon_t} - \sigma_t n_t, \end{aligned} \quad (3)$$

where $\hat{x}_t$ indicates the denoised observation of x_0 at each diffusion step t , ϵ_t denotes the predicted noise at each time step t and σ controls whether the model is stochastic or deterministic. In this paper, we use the modified DDIM Inversion to strengthen the naturalness of the video, maintaining overall visual consistency despite the semantic changes along the temporal axis in the given prompts.

3.2 MEVG pipeline

We outline our proposed MEVG that utilizes the former video clip to generate subsequent video clips considering the given prompts (see Fig. 2). Our method is built upon the latent video diffusion model [15], which leverages the low-dimensional latent space $x \in \mathbb{R}^{F \times c \times h \times w}$, where c , h , and w denote latent space dimension and F indicates the total number of frames. The video output $\mathcal{V} \in \mathbb{R}^{F \times 3 \times H \times W}$ are obtained by passing the latent code x through the decoder $\mathcal{D}$, where $H \times W$ is the resolution of the frame.

To get the final result $\mathbf{V} = \{\mathcal{V}^p\}_{p=0}^{\mathcal{P}-1}$, where $\mathcal{P}$ denotes the number of given prompts, we first sample the video conditioned on the first prompt. An initial video clip is created by Gaussian distribution $x_T \sim \mathcal{N}(0, I)$ as well as a static image as conditional guidance to capture the essential visual

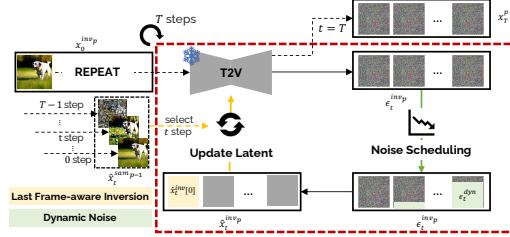


Fig. 3: Last Frame-aware Latent Initialization Initial latent code is crucial for maintaining global geometric structure. We apply two techniques performing different roles: (i) *dynamic noise* tailors flexibility differentially across each frame, and (ii) *last frame-aware inversion* restricts the model to minimize the divergence of the entire frames from the content of the preceding video clip.

content corresponding to user intention. After generating the initial video clip, we extend a preceding video in accordance with the semantic context of the subsequent prompt, driven by two major elements: *last frame-aware latent initialization* and *structure-guided sampling*. *Last frame-aware latent initialization* is the initialization process of the noise latent that helps to preserve the spatial information while generating more diverse contents. *Structure-guided sampling* then enforces motion consistency between frames during the backward diffusion process.

3.3 Last Frame-aware Latent Initialization

The inversion technique, which reconstructs the initial latent code from the visual input (e.g., image or video), is used in real-world applications [8, 27, 47], where accurate spatial layout or visual content reconstruction is important. Video generation should have visual coherence across the entire sequence of frames while adapting the movement of objects and the background transition. To achieve this goal, we aim to find the optimal latent code that helps preserve global coherence between the generated videos for the previous and next prompts and maintains an ability to adapt to the changes. However, the existing approaches [12, 41] present repetitive video patterns (e.g., similar camera movement and object position) and awkward scene transition caused by overlapped content in a single frame.

To solve these challenges, we reuse the generated video (essentially the last frame) from the previous prompt to generate the frames for the new prompt. Basically, the last frame of the previously generated video is copied over the entire sequence as an initial conditioning input. We then propose dynamic noise to enforce the diversity of the generated video. This process preserves the overall visual contents, such as an object and background across the video, and also improves the generation diversity.

Dynamic Noise. To generate diverse motion of the object and smooth transition of the background given prompts, we add the video noise prior similar to [11]. Essentially, the noise vector $\epsilon_t^{dyn} \sim \mathcal{N}(0, \frac{1}{1+\kappa^2}\mathbf{I})$ is added to

Algorithm 1: Last Frame-aware Latent Initialization

Input: latent code of the previous prompt x_0^{p-1} , denoised observation of the last frame from the previous prompt $\{\hat{x}_t^{sam_{p-1}}[-1]\}_{t=0}^{T-1}$, noise scheduling function $\mathcal{F}(\cdot)$, and pre-trained T2V model **T2V**($\cdot$)

Result: Initial latent code x_T^p of next prompt p

// T = Number of diffusion steps
// N = Number of frames
// p = Index of next prompt

- 1 $x_0^{inv_p} \leftarrow \text{REPEAT}(x_0^{p-1}[-1])$
- 2 **for** t in $0, \dots, T-1$ **do**
- 3 $\epsilon_t^{inv_p} \leftarrow \text{T2V}(x_t^{inv_p}, t)$
- // **Dynamic Noise**
- 4 **for** n in $0, \dots, N-1$ **do**
- 5 $\kappa_n \leftarrow \mathcal{F}(n)$
- 6 $\epsilon_t^{dyn} \sim \mathcal{N}(0, \frac{1}{1+\kappa_n^2}\mathbf{I})$
- 7 $\epsilon_t^{inv_p}[n] \leftarrow \frac{\kappa_n}{\sqrt{1+\kappa_n^2}}\epsilon_t^{inv_p}[n] + \epsilon_t^{dyn}$
- 8 **end**
- 9 $\hat{x}_t^{inv_p} \leftarrow (x_t^{inv_p} - \sqrt{1-\alpha_t}\epsilon_t^{inv_p})/\sqrt{\alpha_t}$
- // **Last Frame-aware Inversion**
- 10 $\mathcal{L}_{\text{LFAI}} = \|\hat{x}_t^{sam_{p-1}}[-1] - \hat{x}_t^{inv_p}[0]\|_2^2$
- 11 $\hat{x}_t^{inv_p} \leftarrow \hat{x}_t^{inv_p} - \delta_{\text{LFAI}} \nabla_{\hat{x}_t} \mathcal{L}_{\text{LFAI}}$
- 12 $x_{t+1}^{inv_p} \leftarrow \sqrt{\alpha_{t+1}}\hat{x}_t^{inv_p} + \sqrt{1-\alpha_{t+1}}\epsilon_t^{inv_p}$
- 13 **end**
- 14 $x_T^p = x_T^{inv_p}$

the predicted noise $\epsilon_t^{inv_p}$ during inversion stage for next prompt p . κ regulates the dynamics and variability of the frames within a single video segment; $\kappa \rightarrow 0$ increases video variations.

Since the beginning of the new video should be similar to the preceding video clip, and then more changes occur toward the end of the video, we design a noise scheduling function $\mathcal{F} = \exp(-x)$ that monotonically decreases the κ . κ_n corresponding to the frame index n is determined by:

$$\kappa_n = \mathcal{F}(n), \quad 0 \leq n < N, \quad (4)$$

where N is the total number of frames within one video clip. Finally, the predicted noise $\epsilon_t^{inv_p}$ is obtained as follows:

$$\epsilon_t^{inv_p}[n] = \frac{\kappa_n}{\sqrt{1 + \kappa_n^2}} \epsilon_t^{inv_p}[n] + \epsilon_t^{dyn}, \quad 0 \leq n < N, \quad (5)$$

where $[\cdot]$ denotes the index of frame.

Last Frame-aware Inversion. While *dynamic noise* helps to generate diverse video contents, they cause a temporal inconsistency problem between individual video clips conditioned on consecutive prompts. *Last frame-aware inversion* coerces to maintain a visual correlation between different video clips guided by the denoised observation $\hat{x}_t$. Since the denoised observation $\hat{x}_t$ is the predicted noise-free latent at diffusion step t (see Eq. 3), it contains a sketchy spatial layout and video context. We regularize the initial frame of the current video clip $\hat{x}_t^{inv_p}[0]$ using the denoised observation of the last frame $\hat{x}_t^{sam_{p-1}}[-1]$ from the previous clip. This process ensures the visual consistency between two video clips. We minimize the objective $\mathcal{L}_{\text{LFAI}}$ using L2 loss as follows:

$$\mathcal{L}_{\text{LFAI}} = \|\hat{x}_t^{sam_{p-1}}[-1] - \hat{x}_t^{inv_p}[0]\|_2^2. \quad (6)$$

It basically aligns the denoised observations between the sampling process for the previous prompt and the inversion process for the next prompt at each diffusion step t . After all, we update the $\hat{x}_t^{inv_p}$, the denoised observation during the inversion procedure, along the direction that minimizes the $\mathcal{L}_{\text{LFAI}}$ and δ_{LFAI} controls the guidance strength.

Consequently, through this procedure, we maintain the flexibility allowed by the *dynamic noise* and regularize the overall visual content by employing the denoised observation $\hat{x}_t$. We present the procedure of *last frame-aware latent initialization* in Alg. 1 to facilitate understanding, and the overall procedure are illustrated in Fig. 3.

3.4 Structure-guided Sampling

The video clip is generated for the next prompt using the initial latent x_T^p produced at the previous step. Although the video clip for the new prompt should preserve the appearance of the previous video clip by using the last frame-aware initial latent, undesirable variation in scene texture and object placement often

occurs due to the stochastic nature of the sampling process. To improve the visual consistency within a video clip, we progressively update the predicted original $\hat{x}_t^{sam_p}$ of the current video clip in the sampling process. Specifically, we formulate the objective as follows:

$$\mathcal{L}_{SGS} = \|\hat{x}_t^{sam_p}[1:n] - \hat{x}_t^{sam_p}[:n-1]\|_2^2, \quad (7)$$

where $n \in \{1, \dots, N\}$. Note that for the first frame ($n = 0$), we compute $\mathcal{L}_{SGS}$ using the denoised observation of the last frame from the previous prompt $\{\hat{x}_t^{sam_{p-1}}[-1]\}_{t=0}^{T-1}$. Finally, we update $\hat{x}_t^{sam_p}$ as follows:

$$\hat{x}_t^{sam_p} \leftarrow \hat{x}_t^{sam_p} - \delta_{SGS} \nabla_{\hat{x}_t} \mathcal{L}_{SGS}, \quad (8)$$

where δ_{SGS} is responsible for guidance scale. Eq. 7 and Eq. 8 are iteratively conducted frame-by-frame at each diffusion step. Guidance on the denoised observation leads to a similar global geometric structure between frames within a single video clip.

3.5 Prompt Generator

In real-world scenarios, multiple sequential events can be described in one sentence or paragraph. For instance, “*The dog runs across the wide field, then comes to a halt, yawns softly, and lies down.*”. However, existing long-video generation models [10, 15, 21] are designed to generate only a single event. Therefore, the generated video does not reflect the entire text when the prompt contains multiple events. To address this issue, the Large Language Model (LLM) has been utilized to generate an appropriate input for the pre-trained T2V models. We introduce a prompt generator to segment the comprehensive description into the prescribed textual format. We put the exemplar and guidelines in the supplementary materials.

4 Experiments

4.1 Implementation Details

We directly leverage the released pre-trained text-to-video generation model [15] to generate a multi-text conditioned video. In our experiments, we generate multi-text conditioned video, and each video clip consists of 16 frames with 256×256 resolution. Moreover, we employ ChatGPT [28], which is a Large Language Model, to separate the complex scenarios into individual prompts that comprise the sequence of events. We set each guidance weight δ_{LFAI} and δ_{SGS} to 1000 and 7 in our experiments. All experiments are performed on a single NVIDIA GeForce RTX 3090.

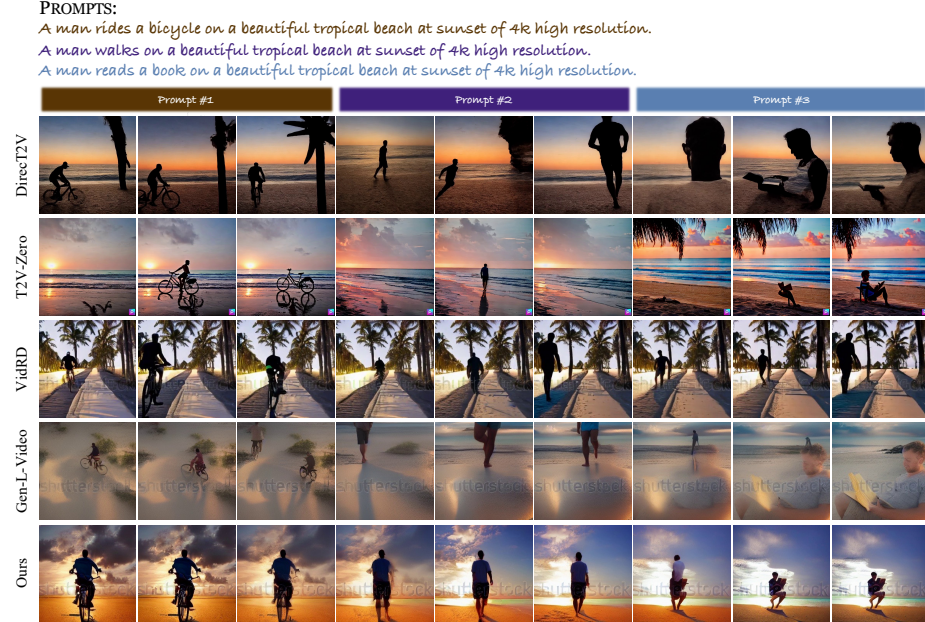


Fig. 4: Generation results on given prompts by our method and baseline models. T2V-Zero and DirecT2V build upon the T2I pre-trained model. In contrast, VidRD and Gen-L-Video leverage the same foundation model utilized in our experiments.

Table 1: Compared with baseline methods in terms of two primary categories: automatic metric and human evaluation. Note that we use **bold** to highlight the best scores, and underline indicates the second-best scores.

Method	Automatic Metric		Human Evaluation			
	CLIP-Text $\uparrow$	CLIP-Image $\uparrow$	Temporal $\uparrow$	Semantic $\uparrow$	Realism $\uparrow$	Preference $\uparrow$
T2V-Zero [24]	0.322	0.808	<u>3.61</u>	<u>3.59</u>	3.45	<u>3.47</u>
DiracT2V [20]	0.301	0.898	2.96	3.04	3.01	3.30
Gen-L-Video [41]	0.308	<u>0.953</u>	3.35	3.38	3.37	3.05
VidRD [12]	0.287	0.951	3.40	3.43	<u>3.56</u>	3.14
Ours	<u>0.309</u>	0.957	3.82	3.71	3.68	3.68

4.2 Qualitative Results

We provide qualitative comparisons along with other recent multi-prompts video generation methods [12,41], including zero-shot video generation methods [20,24] which leveraging frame-level descriptions. In the supplementary material and the project page, we provide additional videos for frame-level and video-level comparison to show more qualitative examples. As shown in Fig. 4, our proposed method achieves better video quality, especially in two points: naturalness and temporal coherence. First, compared with T2V-Zero [24] and DirecT2V [20], we

observe that only leveraging the image-based model fails to generate reasonable video flow in terms of naturalness; thus, utilizing the video-based approaches is necessary. Second, generated videos by our method show a strong visual relation between each video segment without the recurrent video pattern. To be more specific, visual examples of VidRD [12] exhibit recurrent video patterns between two distinct video clips; e.g., the movement pattern of a man within each video segment mirrors the previous one. Additionally, although Gen-L-Video [41] generates more diverse movement, they can not preserve the structure coherence of objects, and the background is not stable across the entire frame. On the other hand, our method not only smoothly bridges the gap between two individual video clips but also maintains the overall temporal coherence of the content.

4.3 Quantitative Results

Automatic Metrics. We report the CLIP-Text score [16, 30] that represents the alignment between given prompts and outputs, and CLIP-Image score [4, 9, 29, 46] that shows the similarity between two consecutive frames. We measure the metrics over the 30 scenarios, each consisting of multiple prompts. For a fair evaluation, we randomly sampled 20 videos per scenario. As shown in Tab. 1, our method generally outperforms the other state-of-the-art methods. Among the baseline, the generated video of Text2Video-Zero (T2V-Zero) aligns well with the semantics of given prompts as this approach yields a single frame corresponding to a prompt of the same video segments.

However, they fail to generate temporally coherent video. In contrast, DirecT2V [20], which utilizes the frame-specific descriptions sharing high-level stories, shows a higher CLIP-Image score, whereas we observe a decrease in performance for the CLIP-Text score. Comparison with text-to-video-based approaches exhibits a relatively low difference in all metrics due to the common foundation model [15]. The CLIP-Text score of VidRD [12] is notably lower than the baselines. The visual content is substantially maintained without a significant performance drop in CLIP-Image score since the VidRD directly utilizes the latent code of the previous video clip with minimal deviations from the sampling step. Gen-L-Video [41] performs well in capturing the meaning of the prompts. However, the global content variations during the sampling process caused by overlapping prompts lead to a decrease in similarity between consecutive frames.

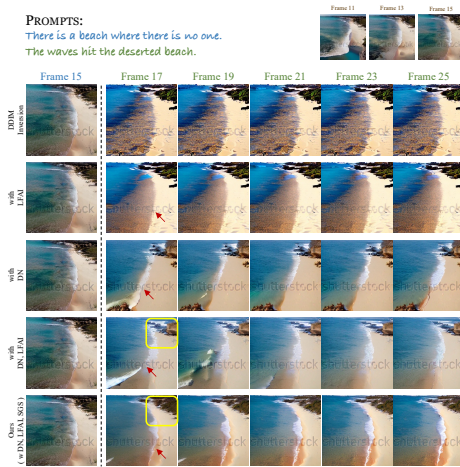


Fig. 5: Generated video clips *with* and *without* proposed our modules. The red arrow and yellow box highlight the visual changes between distinct video clips.

Human Evaluation. We recruited 100 participants through Amazon Mechanical Turk (AMT) to evaluate five models: T2V-Zero [24], DirecT2V [20], Gen-L-Video [41], VidRD [12], and our method. We employ a Likert scale ranging from 1 (low quality) to 5 (high quality). Participants score each method considering temporal consistency, semantic alignment, realism, and preference over 30 videos generated by different scenarios. As clearly indicated in Tab. 1, generated videos from our method significantly outperform other state-of-the-art approaches in all four criteria, regardless of individual frame quality. In particular, based on human evaluation results, we observe that preserving the identity of the object and background is crucial for human preference. When compared with text-to-video-based methods, temporal inconsistency between each video clip caused by the semantic transition of given prompts results in lower human evaluation scores in spite of the same foundation model as ours.

4.4 Ablation Studies

Effectiveness of Proposed Methods. We qualitatively show the effectiveness of *last frame-aware inversion* (LFAI), *dynamic noise* (DN), and *structure-guided sampling* (SGS), as shown in Fig. 5. We utilize the basic inversion strategy [36] as a base model. Although basic DDIM inversion somewhat preserves the overall structure of visual content between each video clip, the detailed texture and background show severe changes. DN model relaxes this problem but shows the disconnection between two video clips since the DN module gives flexibility to the model by using the i.i.d noise in the inversion procedure. Combining two modules, LFAI and DN, generates natural-looking videos since LFAI module preserves the structure of the content in the previous frame. However, we find that the stochastic characteristics of the sampling process introduce slight fluctuations in the videos, and iterative update of latent code (SGS) during the sampling process is beneficial in enhancing the realism of video content, as shown in the fifth row at Fig. 5. Moreover, as reflected in Tab. 2, we provide an additional human evaluation to validate the impact of our proposed module by human judges.

Analysis on Dynamic Noise.

Fig. 6 indicates that κ controls the flexibility of the frame sequence. We modify the noise scheduling function $\mathcal{F}$ into the static value to validate the effectiveness of our method. When κ applies to the entire frames as a smaller value, we figure out that the latter frames can not preserve the geometric structure. However, frozen video is observed when κ is set to a high value. As a result, we achieve the smooth transition and flexibility

Table 2: We report user study results on ablation studies using four different criteria: Temporal, Semantics, Realism, and Preference. Note that we use **bold** to highlight the best scores, and underline indicates the second-best scores.

Method			Human Evaluation			
LFAI	DN	SGS	Temporal $\uparrow$	Semantics $\uparrow$	Realism $\uparrow$	Preference $\uparrow$
-	-	-	3.42	3.40	3.45	2.89
✓	-	-	3.48	3.40	3.50	<u>2.93</u>
-	✓	-	3.53	3.46	<u>3.63</u>	2.51
✓	✓	-	<u>3.61</u>	3.58	3.58	2.78
✓	✓	✓	3.70	<u>3.47</u>	3.69	3.27



Fig. 6: Ablation study for validating the noise schedule. Note that the first and second-row leverage constant value over the entire frames.

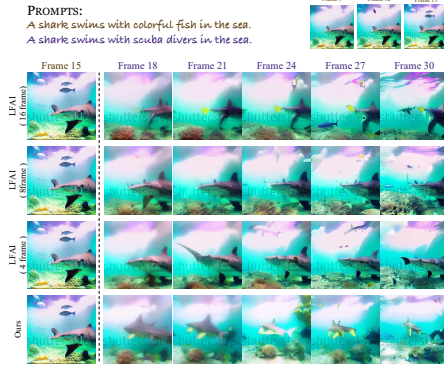


Fig. 7: Effectiveness of adjusting the number of influenced frames.

between consecutive video clips by adopting the scheduling function $\mathcal{F}$, which decrease steadily.

Analysis on Last Frame-aware Inversion. We introduce the *last frame-aware inversion* to prevent the visual inconsistency in terms of object spatial location and scene texture driven by the Dynamic Noise. In the LFAI process, we guide the first frame to adjust the initial latent code that is correlated to the geometric structure of the previous video clip. Here, we explore the influence of the number of frames that offer guidance by the last frame of the previous video clip. As shown in Fig. 7, we observe that only the first frame is sufficient to maintain the visual structure at the beginning of the frames. On the contrary, the increase in the number of affected frames makes the stationary movement in objects; e.g., the shark only moves on the right side. Conversely, the restriction of the affected frame as only a single one gives increased flexibility to the subsequent frames. This flexibility enhances their ability to effectively convey the meaning of the subsequent prompts and generate a diverse range of movement.

4.5 Applications

Video Generation with Large Language Model (LLM). In real-world scenarios, more intricate descriptions are generally used, which have the time-variant events in a single narrative. *Prompt generator* (see Sec. 3.5) to separate into the individual prompts for handling the consecutive events. As shown in Fig. 8 (left), the visual examples indicate the entire frame ensures temporal consistency while reflecting the overall storyline.

Image and Multi-event-based Video Generation. Our proposed MEVG is capable of generating video with a given image and multi-text, multi-text-image-to-video generation (MTI2V). For generating video, we first encode the seeding image using the encoder into the latent vector and duplicate it as the number of frames. Then, we follow MEVG pipeline. Fig. 8 (right) demonstrates that the

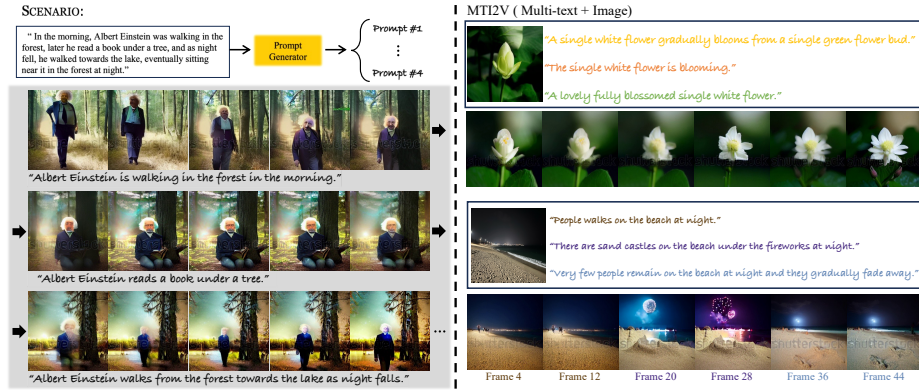


Fig. 8: (Left) Example that leverages the Large Language Model (LLM). Given the complex scenario, our prompt generator split into each individual prompt using the pre-defined instructions. **(Right)** Example of our results conditioned on multiple prompts and given image.

generated video successfully preserves the visual appearance and structure of the object in the reference image and shows temporal coherence along the given prompts.

5 Conclusion

We introduced a novel method that generates multi-text-based videos by taking temporally consecutive descriptions. Specifically, we propose two techniques, *last frame-aware latent initialization* and *structure-guided sampling*, to preserve the visual and temporal consistency in the generated video. Our proposed method can generate much more natural and temporally coherent videos than the other state-of-the-art methods with qualitative and quantitative results. Our pipeline also handles a single story containing time-variant events by utilizing the Large Language Model (LLM). In addition, our proposed method can generate videos conditioned on both the multi-prompts and a reference image and can be used in various applications.

Limitation and Future Work Although our proposed method yields promising outcomes in preserving visual consistency and generation diversity over the distinct prompts, there exists potential for future works as listed: 1) Our method requires a certain text format as our model inherits the characteristics of a pre-trained single-prompt video generator and 2) The absence of benchmark datasets for multi-text video generation makes it hard to conduct a quantitative evaluation. Video generation with diverse input conditions and curating multi-prompts video datasets are promising future directions.

Acknowledgements

This research was supported by Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism in 2024((International Collaborative Research and Global Talent Development for the Development of Copyright Management and Protection Technologies for Generative AI, RS-2024-00345025, 25%),(Research on neural watermark technology for copyright protection of generative AI 3D content, RS-2024-00348469, 25%),(Development of sketch-based semantic 3D modeling technology for creating user-centric Metaverse content spaces for indoor spaces, RS-2023-00227409, 10%))), Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT)(RS-2019-II190079, 10%), Artificial intelligence industrial convergence cluster development project funded by the Ministry of Science and ICT(MSIT, Korea)& Gwangju Metropolitan City(contribution rate: 15%), and IITP under the Leading Generative AI Human Resources Development(IITP-2024-RS-2024-00397085, 15%) grant funded by the Korea government(MSIT).

References

1. Bar-Tal, O., Chefer, H., Tov, O., Herrmann, C., Paiss, R., Zada, S., Ephrat, A., Hur, J., Li, Y., Michaeli, T., et al.: Lumiere: A space-time diffusion model for video generation. arXiv preprint arXiv:2401.12945 (2024)
2. Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S.W., Fidler, S., Kreis, K.: Align your latents: High-resolution video synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22563–22575 (2023)
3. Brooks, T., Hellsten, J., Aittala, M., Wang, T.C., Aila, T., Lehtinen, J., Liu, M.Y., Efros, A., Karras, T.: Generating long videos of dynamic scenes. *Advances in Neural Information Processing Systems* **35**, 31769–31781 (2022)
4. Ceylan, D., Huang, C.H.P., Mitra, N.J.: Pix2video: Video editing using image diffusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 23206–23217 (October 2023)
5. Chen, H., Zhang, Y., Cun, X., Xia, M., Wang, X., Weng, C., Shan, Y.: Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7310–7320 (2024)
6. Chen, X., Wang, Y., Zhang, L., Zhuang, S., Ma, X., Yu, J., Wang, Y., Lin, D., Qiao, Y., Liu, Z.: Seine: Short-to-long video diffusion model for generative transition and prediction. In: The Twelfth International Conference on Learning Representations (2023)
7. Ding, M., Zheng, W., Hong, W., Tang, J.: Cogview2: Faster and better text-to-image generation via hierarchical transformers. *Advances in Neural Information Processing Systems* **35**, 16890–16902 (2022)
8. Dong, W., Xue, S., Duan, X., Han, S.: Prompt tuning inversion for text-driven image editing using diffusion models. arXiv preprint arXiv:2305.04441 (2023)
9. Esser, P., Chiu, J., Atighehchian, P., Granskog, J., Germanidis, A.: Structure and content-guided video synthesis with diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7346–7356 (2023)

10. Ge, S., Hayes, T., Yang, H., Yin, X., Pang, G., Jacobs, D., Huang, J.B., Parikh, D.: Long video generation with time-agnostic vqgan and time-sensitive transformer. In: European Conference on Computer Vision. pp. 102–118. Springer (2022)
11. Ge, S., Nah, S., Liu, G., Poon, T., Tao, A., Catanzaro, B., Jacobs, D., Huang, J.B., Liu, M.Y., Balaji, Y.: Preserve your own correlation: A noise prior for video diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22930–22941 (2023)
12. Gu, J., Wang, S., Zhao, H., Lu, T., Zhang, X., Wu, Z., Xu, S., Zhang, W., Jiang, Y.G., Xu, H.: Reuse and diffuse: Iterative denoising for text-to-video generation. arXiv preprint arXiv:2309.03549 (2023)
13. Gu, S., Chen, D., Bao, J., Wen, F., Zhang, B., Chen, D., Yuan, L., Guo, B.: Vector quantized diffusion model for text-to-image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10696–10706 (2022)
14. He, Y., Xia, M., Chen, H., Cun, X., Gong, Y., Xing, J., Zhang, Y., Wang, X., Weng, C., Shan, Y., et al.: Animate-a-story: Storytelling with retrieval-augmented video generation. arXiv preprint arXiv:2307.06940 (2023)
15. He, Y., Yang, T., Zhang, Y., Shan, Y., Chen, Q.: Latent video diffusion models for high-fidelity video generation with arbitrary lengths. arXiv preprint arXiv:2211.13221 (2022)
16. Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y.: Clipscore: A reference-free evaluation metric for image captioning. arXiv preprint arXiv:2104.08718 (2021)
17. Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D.P., Poole, B., Norouzi, M., Fleet, D.J., et al.: Imagen video: High definition video generation with diffusion models. arXiv preprint arXiv:2210.02303 (2022)
18. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. *Advances in neural information processing systems* **33**, 6840–6851 (2020)
19. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models (2022)
20. Hong, S., Seo, J., Hong, S., Shin, H., Kim, S.: Large language models are frame-level directors for zero-shot text-to-video generation. arXiv preprint arXiv:2305.14330 (2023)
21. Hong, W., Ding, M., Zheng, W., Liu, X., Tang, J.: Cogvideo: Large-scale pretraining for text-to-video generation via transformers. arXiv preprint arXiv:2205.15868 (2022)
22. Huang, H., Feng, Y., SibiYang, C.L.J.: Free-bloom: Zero-shot text-to-video generator with llm director and ldm animator. arXiv preprint arXiv:2309.14494 **3** (2023)
23. Kang, M., Zhu, J.Y., Zhang, R., Park, J., Shechtman, E., Paris, S., Park, T.: Scaling up gans for text-to-image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10124–10134 (2023)
24. Khachatryan, L., Movsisyan, A., Tadevosyan, V., Henschel, R., Wang, Z., Navasardyan, S., Shi, H.: Text2video-zero: Text-to-image diffusion models are zero-shot video generators. arXiv preprint arXiv:2303.13439 (2023)
25. Khandelwal, A.: Infusion: Inject and attention fusion for multi concept zero-shot text-based video editing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3017–3026 (2023)
26. Liang, J., Wu, C., Hu, X., Gan, Z., Wang, J., Wang, L., Liu, Z., Fang, Y., Duan, N.: Nuwa-infinity: Autoregressive over autoregressive generation for infinite visual synthesis. *Advances in Neural Information Processing Systems* **35**, 15420–15432 (2022)

27. Nguyen, T., Li, Y., Ojha, U., Lee, Y.J.: Visual instruction inversion: Image editing via visual prompting. arXiv preprint arXiv:2307.14331 (2023)
28. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* **35**, 27730–27744 (2022)
29. Qi, C., Cun, X., Zhang, Y., Lei, C., Wang, X., Shan, Y., Chen, Q.: Fatezero: Fusing attentions for zero-shot text-based video editing. arXiv:2303.09535 (2023)
30. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: *International conference on machine learning*. pp. 8748–8763. PMLR (2021)
31. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 (2022)
32. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 10684–10695 (2022)
33. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems* **35**, 36479–36494 (2022)
34. Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., et al.: Make-a-video: Text-to-video generation without text-video data. arXiv preprint arXiv:2209.14792 (2022)
35. Skorokhodov, I., Tulyakov, S., Elhoseiny, M.: Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 3626–3636 (2022)
36. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
37. Tian, Y., Ren, J., Chai, M., Olszewski, K., Peng, X., Metaxas, D.N., Tulyakov, S.: A good image generator is what you need for high-resolution video synthesis. arXiv preprint arXiv:2104.15069 (2021)
38. Villegas, R., Babaeizadeh, M., Kindermans, P.J., Moraldo, H., Zhang, H., Saffar, M.T., Castro, S., Kunze, J., Erhan, D.: Phenaki: Variable length video generation from open domain textual description. arXiv preprint arXiv:2210.02399 (2022)
39. Voleti, V., Jolicoeur-Martineau, A., Pal, C.: Mcvd-masked conditional video diffusion for prediction, generation, and interpolation. *Advances in Neural Information Processing Systems* **35**, 23371–23385 (2022)
40. Vondrick, C., Pirsiaavash, H., Torralba, A.: Generating videos with scene dynamics. *Advances in neural information processing systems* **29** (2016)
41. Wang, F.Y., Chen, W., Song, G., Ye, H.J., Liu, Y., Li, H.: Gen-l-video: Multi-text to long video generation via temporal co-denoising. arXiv preprint arXiv:2305.18264 (2023)
42. Wang, Y., Bilinski, P., Bremond, F., Dantcheva, A.: G3an: Disentangling appearance and motion for video generation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 5264–5273 (2020)
43. Weissenborn, D., Täckström, O., Uszkoreit, J.: Scaling autoregressive video models. arXiv preprint arXiv:1906.02634 (2019)

44. Wu, C., Huang, L., Zhang, Q., Li, B., Ji, L., Yang, F., Sapiro, G., Duan, N.: Godiva: Generating open-domain videos from natural descriptions. arXiv preprint arXiv:2104.14806 (2021)
45. Wu, C., Liang, J., Ji, L., Yang, F., Fang, Y., Jiang, D., Duan, N.: Nüwa: Visual synthesis pre-training for neural visual world creation. In: European conference on computer vision. pp. 720–736. Springer (2022)
46. Wu, J.Z., Ge, Y., Wang, X., Lei, S.W., Gu, Y., Shi, Y., Hsu, W., Shan, Y., Qie, X., Shou, M.Z.: Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7623–7633 (2023)
47. Xing, X., Wang, C., Zhou, H., Hu, Z., Li, C., Xu, D., Yu, Q.: Inversion-by-inversion: Exemplar-based sketch-to-photo synthesis via stochastic differential equations without training. arXiv preprint arXiv:2308.07665 (2023)
48. Yan, W., Zhang, Y., Abbeel, P., Srinivas, A.: Videogpt: Video generation using vq-vae and transformers. arXiv preprint arXiv:2104.10157 (2021)
49. Yu, S., Tack, J., Mo, S., Kim, H., Kim, J., Ha, J.W., Shin, J.: Generating videos with dynamics-aware implicit generative adversarial networks. arXiv preprint arXiv:2202.10571 (2022)
50. Zhang, D.J., Wu, J.Z., Liu, J.W., Zhao, R., Ran, L., Gu, Y., Gao, D., Shou, M.Z.: Show-1: Marrying pixel and latent diffusion models for text-to-video generation. arXiv preprint arXiv:2309.15818 (2023)
51. Zhou, D., Wang, W., Yan, H., Lv, W., Zhu, Y., Feng, J.: Magicvideo: Efficient video generation with latent diffusion models. arXiv preprint arXiv:2211.11018 (2022)