MTVG : Multi-text Video Generation with Text-to-Video Models

Gyeongrok Oh[1], Jaehwan Jeong[1], Sieun Kim[1], Wonmin Byeon[2], Jinkyu Kim[3],
Sungwoong Kim[1], Hyeokmin Kwon[1], Sangpil Kim[1]
[1]Department of Artificial Intelligence, Korea University
[2]NVIDIA Research, NVIDIA Corporation
[3]Department of Computer Science and Engineering, Korea University

Figure 1. MTVG produces impressive output that corresponds to the given prompts that consists of chronologically continuous events.

Abstract

Recently, video generation has attracted massive attention and yielded noticeable outcomes. Concerning the characteristics of video, multi-text conditioning incorporating sequential events is necessary for next-step video generation. In this work, we propose a novel multi-text video generation (MTVG) by directly utilizing a pre-trained diffusion-based text-to-video (T2V) generation model without additional fine-tuning. To generate consecutive video segments, visual consistency generated by distinct prompts is necessary with diverse variations, such as motion and content-related transitions. Our proposed MTVG includes Dynamic Noise and Last Frame Aware Inversion which reinitialize the noise latent to preserve visual coherence between videos of different prompts and prevent repetitive motion or contents. Furthermore, we present Structure Guiding Sampling to maintain the global appearance across the frames in a

single video clip, where we leverage iterative latent updates across the preceding frame. Additionally, our Prompt Generator allows for arbitrary format of text conditions consisting of diverse events. As a result, our extensive experiments, including diverse transitions of descriptions, demonstrate that our proposed methods show superior generated outputs in terms of semantically coherent and temporally seamless video. Video examples are available in our project page: https://kuai-lab.github.io/mtvg-page.

1. Introduction

Deep generative models in the computer vision community have a significant attraction due to their unprecedented performance. Especially, text-to-image generation (T2I) [5, 11, 21, 30–32] has successfully produced high-quality images with complex textual descriptions. The driving force behind this success is the advance in hardware accelerators and abundant datasets. While an image captures a single moment, a video delivers a more comprehensive understanding of visual content along the temporal dimension. A naive approach to enable this success in text-to-video generation (T2V) is to scale up the text-video datasets and capacity of the network architecture. However, due to the inherent complexity of modeling temporal dimensions containing motion dynamics and scene transitions, the scarcity of resources and the lack of text-video datasets remain challenging points hindering the progress of T2V.

To handle these aforementioned challenges, recent studies [1, 12, 19, 33] leverage the T2I pre-trained models (e.g., *Stable Diffusion* [31] and *Cogview2* [5]). For data and cost-efficient training, these approaches extend the pre-trained spatial-only layer into the spatial-temporal layer for the video domain and follow the fine-tuning process with video data while transferring the knowledge from the image generation priors. Alternative approaches [18, 20, 22] directly utilize the pre-trained T2I model without additional training. They concentrate on maintaining the global video content across the independently generated frames by redesigning the attention module in a zero-shot manner.

Although the previous efficient T2V approaches have shown promising results, they only focused on a single text-based video generation. Some studies [8, 37] make an effort to embrace 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 [40] leverages the pre-trained T2V network. However, the overlapped denoised process on the consecutive distinct prompts induces visual degradation, and significant inconsistency between the background and objects makes background and object inconsistency. Moreover, traditional long-term T2V methods [1, 8, 19, 33, 49] bridge the gap between each keyframe, followed by generating the keyframes according to a single description. This hierarchical process makes it hard to incorporate the multiple-time variant prompts since it comprehensively generates the global video content in the beginning.

In essence, not only videos are just the sequence of frames, but also the sequence of events. Videos in the wild contain complicated information that changes movement, background, object, and camera motion as time goes on. However, existing methods generate a video with a single prompt that disregards temporal information and restrictively expresses the semantics of the entire story. Therefore, we emphasize that multi-prompt conditioned video generation is necessary to take the video generation one step further.

Our method, multi-text video generation (MTVG), is delicately designed as a multi-prompt video generator without any training or fine-tuning and enforces the temporal coherence between independently generated video clips. MTVG adopts the sequence of prompts to induce a temporal shift in semantics. By building upon the publicly released video generation model, we effectively utilize the pre-trained single-prompt T2V generative model[1] to generate complex scenario videos. Specifically, to preserve the visual coherence between time variable prompts in generation while producing realistic and diverse motion, we introduce two novel techniques: last frame-aware latent initialization and structure-guided sampling. First, the last frame-aware latent 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 extensive experiments, we show that our proposed methods synthesize realistic videos given scenarios that include three representative types of change: object motion, background change, and complex change. Moreover, we examine the effectiveness and legitimacy of proposed modules in ablation studies. To summarize, our main contributions are as follows:

- Our work constructs a novel pipeline to generate video conditioned on multiple prompts without requiring any training or fine-tuning.
- We present two distinct methods that enforce consis-

[1] We used [12] in our experiments, but it can be any T2V pre-trained model based on the diffusion model.

tency (temporal and semantic) between individual video segments and each frame within a single video clip. Furthermore, our proposed method possesses the flexibility to adapt new prompts and semantic transitions.
- Experimental results demonstrate that our proposed pipeline reflects the semantic given multi-prompt while maintaining visually coherent contents against existing zero-shot video generation methods.

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) [2, 34, 36, 39, 41, 48] is a well-known algorithm to generate diverse video from a noise vector utilizing a generator and discriminator. Another approach is auto-regressive transformers [8, 19, 42–44, 47] that leverage discrete representation to depict the motion dynamics. Recently, diffusion-based methods [1, 7, 9, 12, 16, 17, 33, 38, 50] 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 [8, 19, 24] leveraging transformer architecture show remarkable outcomes in long-term video generation. However, they require to use massive training costs and tremendous datasets. Furthermore, although TATS [8] and Phenaki [37] 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 nature of the auto-regressive property. Several works [1, 9, 33, 38, 49] based on the diffusion model leverage temporal interpolation networks and masked strategies for generating smoother videos. VidRD [10] directly utilizes the previous initial latent code to expand the video.

However, most previous works focused on video generation from a single prompt or an event. In this work, we tackle the multi-prompt video generation task that consists of consecutive events across the long-term video. Animate-A-Story [13] utilizes an abundant real-world video corresponding to each story for natural motion. Gen-L-Video [40], 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 [27] and INFUSION [23] 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 [22] 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 [20] and DirecT2V [18] 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.

3. Method

We propose a novel pipeline, MTVG, to generate a temporally and semantically coherent video conditioned on multiple 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 [15] 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:

$$x_t = \sqrt{\bar{\alpha}_t}x + \sqrt{1-\bar{\alpha}_t}\epsilon, \tag{1}$$

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:

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

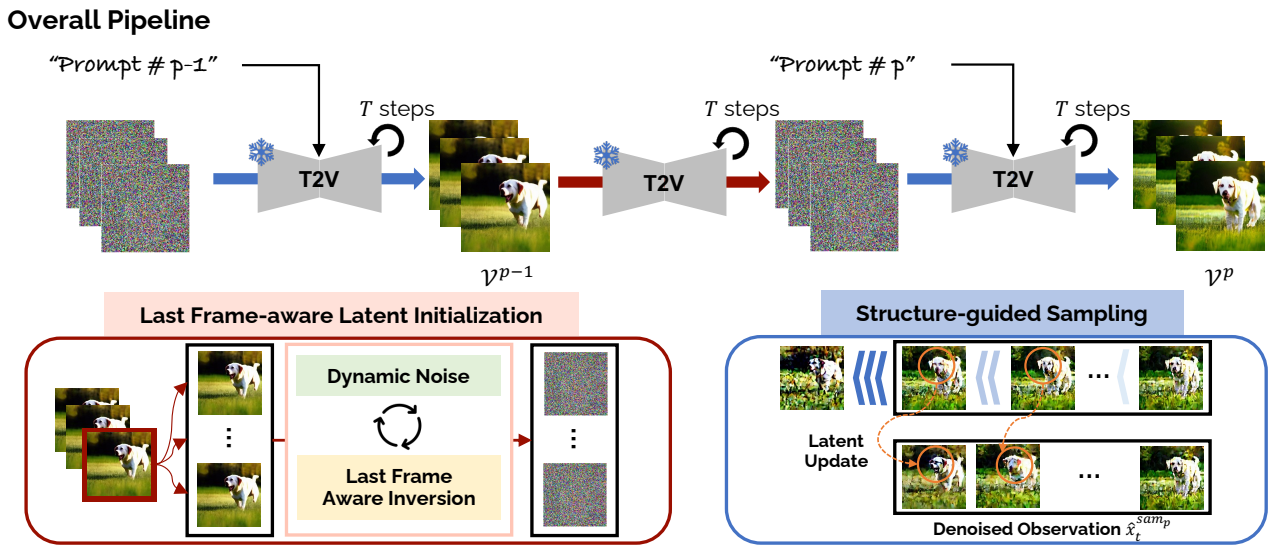


Figure 2. **Overview of MTVG** MTVG synthesizes the consecutive video clips corresponding to distinct prompts. The overall pipeline comprises two major components: last frame-aware latent initialization and structure-guided sampling. First, in the last frame-aware latent initialization, the pre-trained 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.

DDIM DDIM [35] 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:

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

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. MTVG pipeline

We outline our proposed MTVG 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 [12], 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 $\mathrm{V} = \{\mathcal{V}^p\}_{p=0}^{\mathcal{P}-1}$, 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 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 [6, 25, 46], 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 pre-

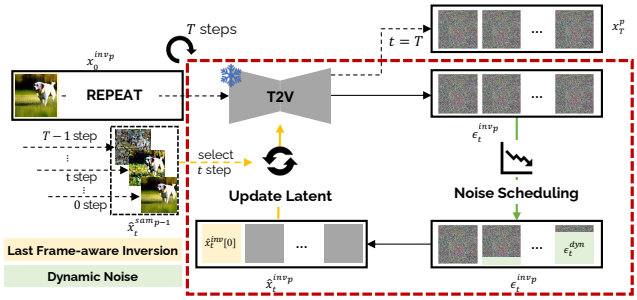


Figure 3. **Last Frame-aware Latent Initialization** Initial latent code is crucial for maintaining global geometric structure. In the last frame-aware latent initialization process, we apply two techniques performing different roles: (i) *dynamic noise* tailors flexibility differentially across each 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.

vious and next prompts and maintains an ability to adapt to the changes. However, the existing approaches [10, 40] 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 [9]. Essentially, the noise vector $\epsilon_t^{dyn} \sim \mathcal{N}(0, \frac{1}{1+\kappa^2}\mathrm{I})$ is added to 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; $\kappa \to 0$ 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:

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

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

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

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

Algorithm 1: Last Frame-aware Latent Initialization

Input: latent code of the 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 $\mathbf{T2V}(\cdot)$

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
```

$x_0^{inv_p} \leftarrow \mathrm{REPEAT}(x_0^{p-1}[-1])$
for t *in* $0, ..., T-1$ **do**
  $\epsilon_t^{inv_p} \leftarrow \mathbf{T2V}(x_t^{inv_p}, t)$
  // Dynamic Noise
  for n *in* $0, ..., N-1$ **do**
    $\kappa_n \leftarrow \mathcal{F}(n)$
    $\epsilon_t^{dyn} \sim \mathcal{N}(0, \frac{1}{1+\kappa_n^2}\mathrm{I})$
    $\epsilon_t^{inv_p}[n] \leftarrow \frac{\kappa_n}{\sqrt{1+\kappa_n^2}}\epsilon_t^{inv_p}[n] + \epsilon_t^{dyn}$
  end
  $\hat{x}_t^{inv_p} \leftarrow (x_t^{inv_p} - \sqrt{1-\bar{\alpha}_t}\epsilon_t^{inv_p})/\sqrt{\bar{\alpha}_t}$
  // Last Frame-aware Inversion
  $\mathcal{L}_{\text{LFAI}} = ||\hat{x}_t^{sam_{p-1}}[-1] - \hat{x}_t^{inv_p}[0]||_2^2$
  $\hat{x}_t^{inv_p} \leftarrow \hat{x}_t^{inv_p} - \delta_{\text{LFAI}}\nabla_{\hat{x}_t}\mathcal{L}_{\text{LFAI}}$
  $x_{t+1}^{inv_p} \leftarrow \sqrt{\bar{\alpha}_{t+1}}\hat{x}_t^{inv_p} + \sqrt{1-\bar{\alpha}_{t+1}}\epsilon_t^{inv_p}$
end
$x_T^p = x_T^{inv_p}$

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:

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

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.

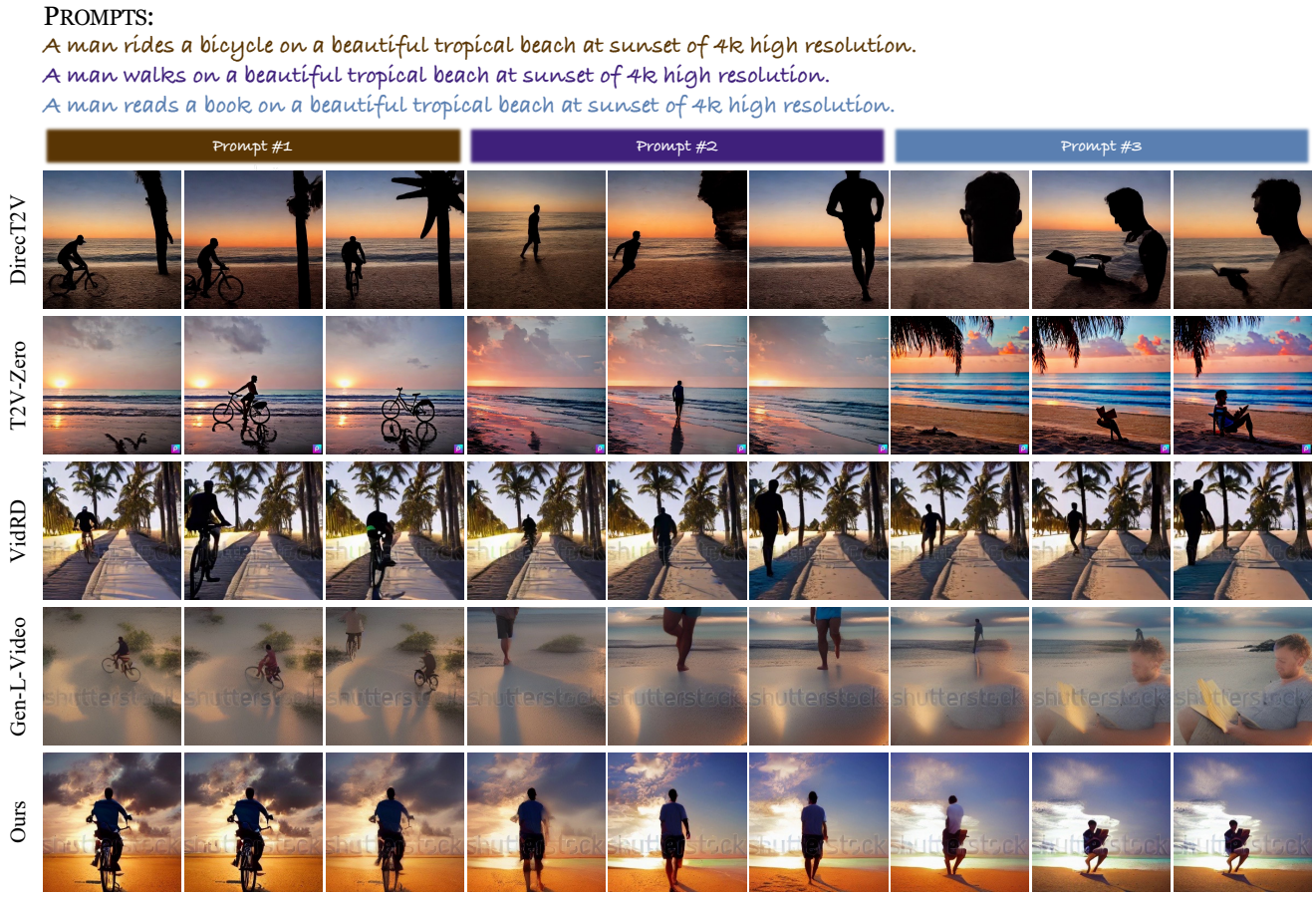


Figure 4. **Qualitative Results** Generation results on given prompts by our method and baseline models. T2V-Zero and DirecT2V build upon the T2I pre-trained model. In contrast, VidRD and Gen-L-Video leverage the same foundation model utilized in our experiments.

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:

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

where $n \in \{1, ..., \text{N}\}$. 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:

$$\hat{x}_t^{sam_p} \leftarrow \hat{x}_t^{sam_p} - \delta_{\text{SGS}} \nabla_{\hat{x}_t} \mathcal{L}_{\text{SGS}}, \tag{8}$$

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 [8, 12, 19] 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 (T2V) model [12] 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 [26], which is a Large Language Model (LLM), 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.

4.2. Qualitative Results

We provide qualitative comparisons along with other recent multi-prompts video generation methods [10, 40], including zero-shot video generation methods [18, 22] 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 [22] and DirecT2V [18], 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 [10] 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 [40] 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 [14, 29] that represents the alignment between given prompts and outputs, and CLIP-Image score [3, 7, 27, 45] 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 [18], 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 [12]. The CLIP-Text score of VidRD [10] 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 [40] 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 similarity decrease between consecutive frames.

Human Evaluation We recruited 100 participants through Amazon Mechanical Turk (AMT) to evaluate five models: T2V-Zero [22], DirecT2V [18], Gen-L-Video [40], VidRD [10], 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. 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

Table 1. **Quantitative Results** 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.

	Automatic Metric		Human Evaluation			
Method	CLIP-Text ↑	CLIP-Image ↑	Temporal ↑	Semantic ↑	Realism ↑	Preference ↑
T2V-Zero [22]	**0.322**	0.808	3.61	3.59	3.45	3.47
DirecT2V [18]	0.301	0.898	2.96	3.04	3.01	3.30
Gen-L-Video [40]	0.308	0.953	3.35	3.38	3.37	3.05
VidRD [10]	0.287	0.951	3.40	3.43	3.56	3.14
Ours	0.309	**0.957**	**3.82**	**3.71**	**3.68**	**3.68**

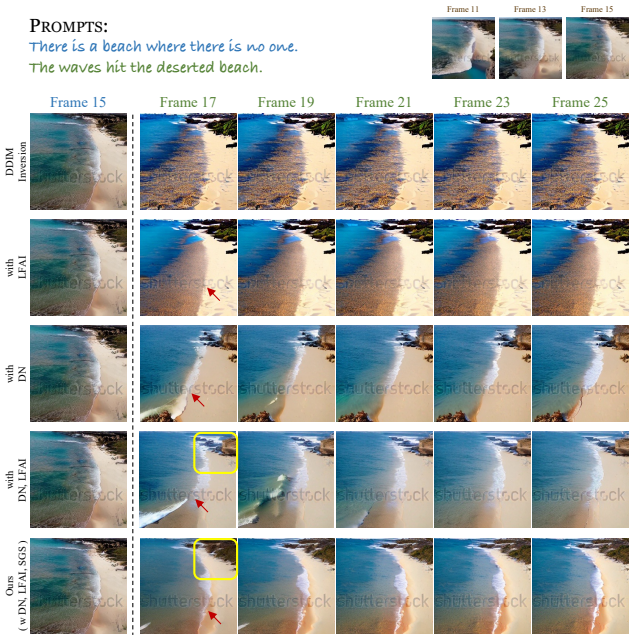


Figure 5. Generated video clips *with* and *without* proposed our modules. The red arrow and yellow box highlight the visual changes between distinct video clips.

shown in Fig. 5. We utilize the basic inversion strategy [35] 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.

Figure 6. Example of our results (bottom) conditioned on multiple prompts and given image (top).

4.5. Applications

Image and Multi-text-based Video Generation Our proposed MTVG 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 our MTVG pipeline. Fig. 6 demonstrates that the generated video successfully preserves the visual appearance and structure of the object in the reference image and shows temporal coherence along the given prompts.

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. 7, the visual examples indicate the entire frame ensures temporal consistency while reflecting the overall storyline.

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

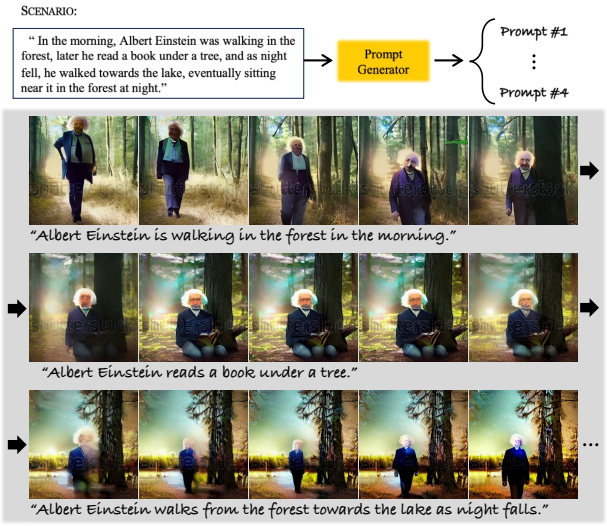


Figure 7. Example that leverages the Large Language Model (LLM). Given the complex scenario, our prompt generator split into each individual prompt using the pre-defined instructions.

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.

References

[1] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023. 2, 3

[2] Tim Brooks, Janne Hellsten, Miika Aittala, Ting-Chun Wang, Timo Aila, Jaakko Lehtinen, Ming-Yu Liu, Alexei Efros, and Tero Karras. Generating long videos of dynamic scenes. *Advances in Neural Information Processing Systems*, 35:31769–31781, 2022. 3

[3] Duygu Ceylan, Chun-Hao P. Huang, and Niloy J. Mitra. Pix2video: Video editing using image diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 23206–23217, 2023. 7, 12

[4] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023. 16, 22, 23

[5] Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *Advances in Neural Information Processing Systems*, 35:16890–16902, 2022. 2

[6] Wenkai Dong, Song Xue, Xiaoyue Duan, and Shumin Han. Prompt tuning inversion for text-driven image editing using diffusion models. *arXiv preprint arXiv:2305.04441*, 2023. 4

[7] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7346–7356, 2023. 3, 7, 12

[8] 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 *European Conference on Computer Vision*, pages 102–118. Springer, 2022. 2, 3, 7

[9] Songwei Ge, Seungjun Nah, Guilin Liu, Tyler Poon, Andrew Tao, Bryan Catanzaro, David Jacobs, Jia-Bin Huang, Ming-Yu Liu, and Yogesh Balaji. Preserve your own correlation: A noise prior for video diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22930–22941, 2023. 3, 5

[10] Jiaxi Gu, Shicong Wang, Haoyu Zhao, Tianyi Lu, Xing Zhang, Zuxuan Wu, Songcen Xu, Wei Zhang, Yu-Gang Jiang, and Hang Xu. Reuse and diffuse: Iterative denoising for text-to-video generation. *arXiv preprint arXiv:2309.03549*, 2023. 3, 5, 7, 8, 12, 16, 17, 18

[11] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10696–10706, 2022. 2

[12] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity video generation with arbitrary lengths. *arXiv preprint arXiv:2211.13221*, 2022. 2, 3, 4, 7, 12, 19, 20, 21

[13] 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. 3

[14] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 7, 12

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

[16] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 3

[17] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models, 2022. 3

[18] Susung Hong, Junyoung Seo, Sunghwan Hong, Heeseong Shin, and Seungryong Kim. Large language models are frame-level directors for zero-shot text-to-video generation. *arXiv preprint arXiv:2305.14330*, 2023. 2, 3, 7, 8, 12, 16, 17, 18

[19] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022. 2, 3, 7

[20] Hanzhuo Huang, Yufan Feng, and ChengShi LanXu JingyiYu SibeiYang. Free-bloom: Zero-shot text-to-video generator with llm director and ldm animator. *arXiv preprint arXiv:2309.14494*, 3, 2023. 2, 3, 12, 14

[21] Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up gans for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10124–10134, 2023. 2

[22] 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. *arXiv preprint arXiv:2303.13439*, 2023. 2, 3, 7, 8, 12, 14, 16, 17, 18

[23] Anant Khandelwal. 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*, pages 3017–3026, 2023. 3

[24] Jian Liang, Chenfei Wu, Xiaowei Hu, Zhe Gan, Jianfeng Wang, Lijuan Wang, Zicheng Liu, Yuejian Fang, and Nan Duan. Nuwa-infinity: Autoregressive over autoregressive generation for infinite visual synthesis. *Advances in Neural Information Processing Systems*, 35:15420–15432, 2022. 3

[25] Thao Nguyen, Yuheng Li, Utkarsh Ojha, and Yong Jae Lee. Visual instruction inversion: Image editing via visual prompting. *arXiv preprint arXiv:2307.14331*, 2023. 4

[26] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744, 2022. 7

[27] Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. Fatezero: Fusing attentions for zero-shot text-based video editing. *arXiv:2303.09535*, 2023. 3, 7, 12

[28] 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. 14

[29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 7, 12

[30] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 2

[31] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 12

[32] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 2

[33] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 2, 3

[34] Ivan Skorokhodov, Sergey Tulyakov, and Mohamed Elhoseiny. 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*, pages 3626–3636, 2022. 3

[35] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 4, 8

[36] 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. *arXiv preprint arXiv:2104.15069*, 2021. 3

[37] 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. *arXiv preprint arXiv:2210.02399*, 2022. 2, 3

[38] Vikram Voleti, Alexia Jolicoeur-Martineau, and Chris Pal. Mcvd-masked conditional video diffusion for prediction, generation, and interpolation. *Advances in Neural Information Processing Systems*, 35:23371–23385, 2022. 3

[39] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. *Advances in neural information processing systems*, 29, 2016. 3

[40] 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, 3, 5, 7, 8, 12, 16, 17, 18

[41] Yaohui Wang, Piotr Bilinski, Francois Bremond, and Antitza Dantcheva. G3an: Disentangling appearance and motion for video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5264–5273, 2020. 3

[42] Dirk Weissenborn, Oscar Täckström, and Jakob Uszkoreit. Scaling autoregressive video models. *arXiv preprint arXiv:1906.02634*, 2019. 3

[43] 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.

[44] 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 *European conference on computer vision*, pages 720–736. Springer, 2022. 3

[45] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. 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*, pages 7623–7633, 2023. 7, 12

[46] Ximing Xing, Chuang Wang, Haitao Zhou, Zhihao Hu, Chongxuan Li, Dong Xu, and Qian Yu. Inversion-by-inversion: Exemplar-based sketch-to-photo synthesis via stochastic differential equations without training. *arXiv preprint arXiv:2308.07665*, 2023. 4

[47] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*, 2021. 3

[48] Sihyun Yu, Jihoon Tack, Sangwoo Mo, Hyunsu Kim, Junho Kim, Jung-Woo Ha, and Jinwoo Shin. Generating videos with dynamics-aware implicit generative adversarial networks. *arXiv preprint arXiv:2202.10571*, 2022. 3

[49] 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. 2, 3

[50] 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. 3

MTVG : Multi-text Video Generation with Text-to-Video Models

Supplementary Material

Overview

This supplementary material introduces experiment details, details of *prompt generator*, more ablation studies, test set, and further qualitative results.

- Section A provides more experiment details about the baseline, metrics, and human evaluation.
- Section B presents the usage of *prompt generator*, including full descriptions used to generate individual prompts from the scenario containing the sequence of events.
- Section C supplements our main paper with additional ablation studies to validate our design choices. We further present the human evaluation results of our proposed module related to Sec. 4.3 in the main paper. In addition, we analyze the effectiveness of individual module (*dynamic noise* and *last frame-aware inversion*).
- Section D introduces the test set for qualitative results.
- Section E provides more qualitative results in diverse domains. We provide more comparison results between state-of-the-art models and qualitative results. Furthermore, we also present more generated videos conditioning on image and multi-text and examples generated by the prompt generator.

A. Experiment Details

Baselines To demonstrate the effectiveness of our proposed MTVG, we compare the outcomes with several existing baselines. We select a baseline that enables synthesizing the videos with multiple prompts without any training or fine-tuning. DirecT2V [18] and Text2Video-Zero (T2V-Zero) [22] leverage Stable Diffusion [31] trained on only text-image pairs. These models utilize the frame-level descriptions to create individual frames constituting the video content. Furthermore, two text-to-video-based methods, Gen-L-Video [40] and VidRD [10], are used to compare with ours. We use LVDM [12] as the foundation model for our experiment to be a fair comparison.

Metrics We report CLIP Similarity ("**CLIP-Text**") [14, 29] and temporal consistency ("**CLIP-Image**") [3, 7, 27, 45] to evaluate our proposed MTVG. CLIP-Text is a commonly employed metric to measure the correlation between two different modalities, image and text. We compute the cosine similarity over all frames corresponding to each prompt to present how well the outcomes reflect the meaning of given conditions. Additionally, CLIP-Image is used to measure the correlation of frame. We compute the cosine similarity between two consecutive frames and take the average value over all frames.

Table 2. **Human evaluation** 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 ↑	Semantics ↑	Realism ↑	Preference ↑
-	-	-	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**

Human evaluation We conduct a human evaluation study to measure four properties of outcomes: temporal consistency, semantic alignment, realism, and preference. Specifically, we request all participants assign a score on a scale 1 (low quality) to 5 (high quality) for the following set of four questions. First,"*How smoothly the content of videos changes in response to the given prompts.*" indicates how each video clip is smoothly connected between the distinct prompts (Temporal Consistency). Second,"*How well does the video correspond with the prompts.*" evaluates how well the generated video reflects a given sequence of prompts (Semantic Alignment). Third, "*How natural and real does this video look, considering the consistency of the background and the objects.*" evaluates the realism of the generated video concerning the background and object consistency (Realism). Finally, "*Considering the three questions above, please rank the overall video quality.*" leads to participants ranking their preference over the generated video based on comprehensive perspective (Preference).

B. Details on Prompt Generator

In this section, we provide additional information of *prompt generator* that is described in our main paper (see Sec. 3.5). Our *prompt generator* can naturally split a single scenario containing multiple events into distinct prompts with a prescribed textural format.

Instruction for *prompt generator* The key point of the *prompt generator* is that each prompt has a single event while maintaining the comprehensive content of the scenario. Inspired by the Free-Bloom [20] and DirecT2V [18], we devise adequate instruction following the five concrete rules (see Fig. 8).

I would like you to play the role of the long sentence separator that breaks down long sentences into **{ The Number of Prompt }** concise short sentences, focusing on **{ The Number of Prompt }** main action verbs.
A long sentence is **{ Scenario }** and there are **{ The Number of Prompt }** short sentences.
First, consider the context of a given long sentence and then change the pronoun to a specific noun given in the long sentence with an indefinite article.
Second, you break down modified long sentences into **{ The Number of Prompt }** concise short sentences, focusing on **{ The Number of Prompt }** main action verbs by following the five rules.

There are some rules as follow:

First, Each short sentence will contain only one action verb and action verb's tenses and forms in each short sentence should match those mentioned in the long sentence.
Second, Each short sentence must be self-contained, following the order of subject, verb, and background.
Third, Each short sentence should contain all background information related to the main verb of a short sentence.
Fourth, Each Short sentence should not include any verbs other than the **{ The Number of Prompt }** main verbs.
Fifth, Each short sentence maintains the present tense, present progressive tense, and present participle as expressed in the long sentence.

Scenario : { user input }, The Number of Prompt : { user input }

Figure 8. This instruction follows the five guidelines to create individual prompts based on a given scenario and the number of prompts by the user.

C. Ablation Study

User Study on Proposed Methods We report the human evaluation results on the effectiveness of the proposed modules. We engage an additional 45 participants in Amazon Mechanical Turk (AMT). All participants look 5 different videos generated by various combinations of our proposed modules and rate the temporal consistency, semantic consistency, realism, and preference over 30 different scenarios. As clearly indicated in Tab. 2, our proposed methods enhance the video quality in terms of the above four criteria.

Analysis on Dynamic Noise (DN) Fig. 9 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 $kappa$ is set to a high value. Thus, we set the scheduling function $\mathcal{F}$ to have the monotonically decreasing form.

Analysis on Last Frame-aware Inversion (LFAI) 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. 10, 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.

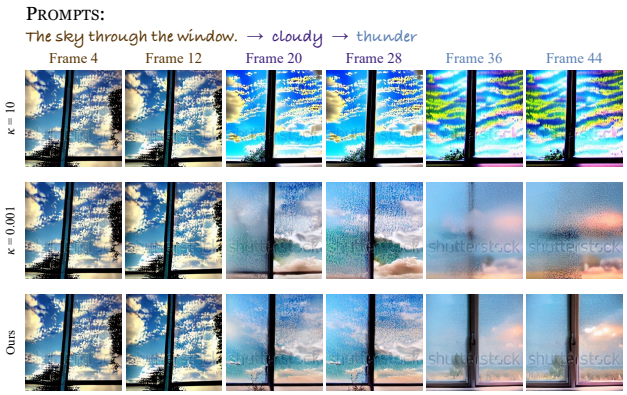


Figure 9. Ablation study for validating the noise schedule. Note that the first and second-row leverage constant value over the entire frames.

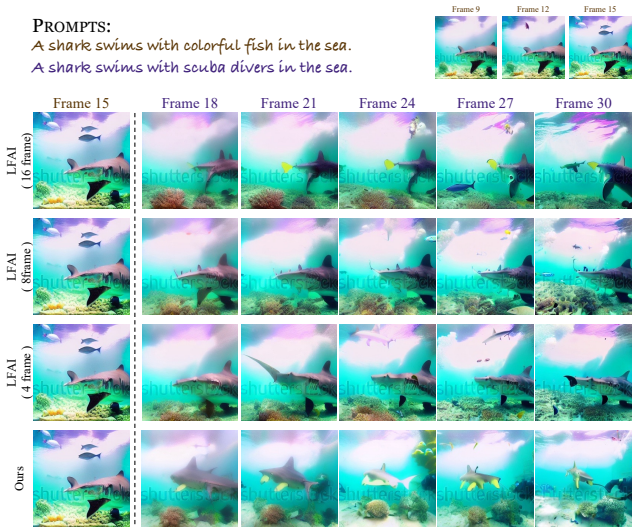


Figure 10. Effectiveness of adjusting the number of influenced frames. The last frame-aware inversion only guides the first frame of the current prompt to maintain the geometric structure between each video clip. Attending only one frame demonstrates better diverse movement in the latter of the frames.

D. Test set

Since there are no evaluation datasets for multi-text-based video generation, we construct a test set by referring to generative model literature communities. Some prompts are derived from the existing works [20, 22, 28]. To evaluate the quality of the generated videos, we design complex scenarios consisting of multiple prompts. Each scenario is divided into three categories: object movements, background transitions, and complex changes. Scenarios consist of two, three, or four prompts while containing diverse objects and backgrounds in different domains. The test set is listed as follows:

Background Transition

- *Scenario* 1.

  1. The teddy bear goes under water in San Francisco.
  2. The teddy bear keeps swimming under the water with colorful fishes.
  3. A panda bear is swimming under water.

- *Scenario* 2.

  1. An astronaut in a white uniform is snowboarding in the snowy hill.
  2. An astronaut in a white uniform is surfing in the sea.
  3. An astronaut in a white uniform is surfing in the desert.

- *Scenario* 3.

  1. A white butterfly sits on a purple flower.
  2. The color of the purple flower where the white butterfly sits turns red.
  3. A white butterfly is sitting on a red flower.

- *Scenario* 4.

  1. The caterpillar is on the leaves.
  2. The caterpillar eats the leaves.
  3. The caterpillar ate all the leaves.

- *Scenario* 5.

  1. The teddy bear is swimming under the sea.
  2. The teddy bear is playing with colorful fishes while swimming under the sea.
  3. The teddy bear is resting quietly among the coral reefs under the sea.
  4. Suddenly a shark appeared next to the teddy bear under the sea.

- *Scenario* 6.

  1. A man runs the starry night road in Van Gogh style.
  2. A man runs the starry night road in Monet style.
  3. A man runs the starry night road in Picasso style.
  4. A man runs the starry night road in Da Vinci style.

- *Scenario* 7.

  1. The whole beautiful night view of the city is shown.
  2. Heavy rain flood the city with beautiful night scenery and flood.
  3. The day dawns over the flooded city.

- *Scenario* 8.

  1. Cherry blossoms bloom around the Japanese-style castle.
  2. Leaves fall around the Japanese-style castle.
  3. Snow falls around the Japanese-style castle.
  4. Snow builds up in trees around the Japanese-style castle.

- *Scenario* 9.

  1. The dog is standing on Times Square Street.
  2. The dog is standing on the Japanese street.
  3. The dog is standing on the China town.
  4. The dog is standing on the street in Korea.

- *Scenario* 10.

  1. In spring, a white butterfly sit on a flower.
  2. In summer, a white butterfly sit on flower.

3. In autumn, a white butterfly sit on flower.
4. In winter, a white butterfly sit on flower.

Object Motion

- *Scenario* 11.

---

1. Two men play tennis in the green gym.
2. Two men playing tennis swing a racket in the green gym.
3. A tennis ball passes between two men playing tennis in the green gym.

- *Scenario* 12.

---

1. A man sits in front of a standing microphone on Times Square Street and plays the guitar.
2. The man sits on the street in Times Square and sings on the guitar.
3. The man sits on Times Square Street and keeps playing the guitar.

- *Scenario* 13.

---

1. A shark swims with colorful fish in the sea.
2. A shark swims with scuba divers in the sea.
3. A shark dances with scuba divers in the sea.

- *Scenario* 14.

---

1. A candle is brightly lit in the dark room.
2. Smoke rises from an unlit candle in the dark room.
3. There is an unlit candle in a dark room.

- *Scenario* 15.

---

1. There is a beach where there is no one.
2. The waves hit the deserted beach.
3. There is a beach that has been swept away by waves.

- *Scenario* 16.

---

1. A dog runs in the snowy mountains.
2. A dog barks on snowy mountain.
3. A dog stands on snowy mountain.
4. A dog lies down on the snowy mountain.

- *Scenario* 17.

---

1. A man runs on a beautiful tropical beach at sunset of 4k high resolution.
2. A man rides a bicycle on a beautiful tropical beach at sunset of 4k high resolution.
3. A man walks on a beautiful tropical beach at sunset of 4k high resolution.
4. A man reads a book on a beautiful tropical beach at sunset of 4k high resolution.

- *Scenario* 18.

---

1. A sheep is standing in a field full of grass.
2. A sheep graze in a field full of grass.
3. A sheep is running in a field full of grass.
4. A sheep is lying in a field full of grass.

- *Scenario* 19.

---

1. A golden retriever has a picnic on a beautiful tropical beach at sunset.
2. A golden retriever is running towards a beautiful tropical beach at sunset.
3. A golden retriever sits next to a bonfire on a beautiful tropical beach at sunset.
4. A golden retriever is looking at the starry sky on a beautiful tropical beach.

- *Scenario* 20.

---

1. A Red Riding Hood girl walks in the woods.
2. A Red Riding Hood girl sells matches in the forest.
3. A Red Riding Hood girl falls asleep in the forest.
4. A Red Riding Hood girl walks towards the lake from the forest.

Complex changes

- *Scenario* 21.

---

1. Side view of an astronaut is walking through a puddle on mars.
2. The astronaut watches fireworks.

- *Scenario* 22.

---

1. The astronaut gets on the spacecraft.
2. The spacecraft goes from Earth to Mars.
3. The spacecraft lands on Mars.

- *Scenario* 23.

---

1. The volcano erupts in the clear weather.
2. Smoke comes from the crater of the volcano, which has ended its eruption in the clear weather.
3. The weather around the volcano turns cloudy.

- *Scenario* 24.

---

1. There is a Mickey Mouse dancing through the spring forest.
2. There is a Mickey Mouse walking through the autumn forest.

3. There is a Mickey Mouse running through the winter forest.

- *Scenario* 25.

---

1. A panda is playing guitar on Times Square.
2. The panda is singing on Times Square.
3. The panda starts dancing.
4. People in Times Square clap for the panda.

- *Scenario* 26.

---

1. A teddy bear walks on the streets of Times Square .
2. The teddy bear enters restaurants.
3. The teddy bear eats pizza.
4. The teddy bear drinks water.

- *Scenario* 27.

---

1. The cartoon-style bear appears in a comic book.
2. The cartoon-style bears in comic books jump out into the real world.
3. The bear in the real world dances.
4. The bear in the real world sits.

- *Scenario* 28.

---

1. A chihuahua in astronaut suit floating in space, cinematic lighting, glow effect.
2. A chihuahua in astronaut suit dancing in space, cinematic lighting, glow effect.
3. A chihuahua in astronaut suit swimming under the water, clean, brilliant effect.
4. A chihuahua in astronaut suit swimming under the water with colorful fishes, clean, brilliant effect.

- *Scenario* 29.

---

1. A waterfall flows in the mountains under a clear sky.
2. A waterfall flows in the fall mountains under a clear sky.
3. A waterfall flows in the winter mountains under a clear sky.
4. A waterfall frozen on a mountain during a snowstorm.

- *Scenario* 30.

---

1. The boulevards are quiet in the clear sky.
2. The boulevards are quiet in the night sky.
3. The boulevards are crowded in the night sky.
4. The boulevards are crowded under the firework sky.

E. Qualitative Results

In this section, we provide more qualitative results of our methods in the multi-text video generation setting. Specifically, Fig. 11 and Fig. 12 represent the qualitative comparison with the state-of-the-art methods [10, 18, 22, 40]. In Fig. 13 $\sim$ Fig. 17, we showcase the multi-text video generation results over the diverse domain. Nota that, Fig 16 and Fig 17 leverage the different foundation model[2] and generate 16 frames per each prompt with 576$\times$1024 resolution. Furthermore, we visualize the generated videos conditioning on image and multi-text (see Fig. 18 and Fig. 19). Finally, we present additional results generated by the *prompt generator* in Fig. 20 and Fig. 21.

[2] VideoCrafter1 [4] is used as the foundation model in this experiment.

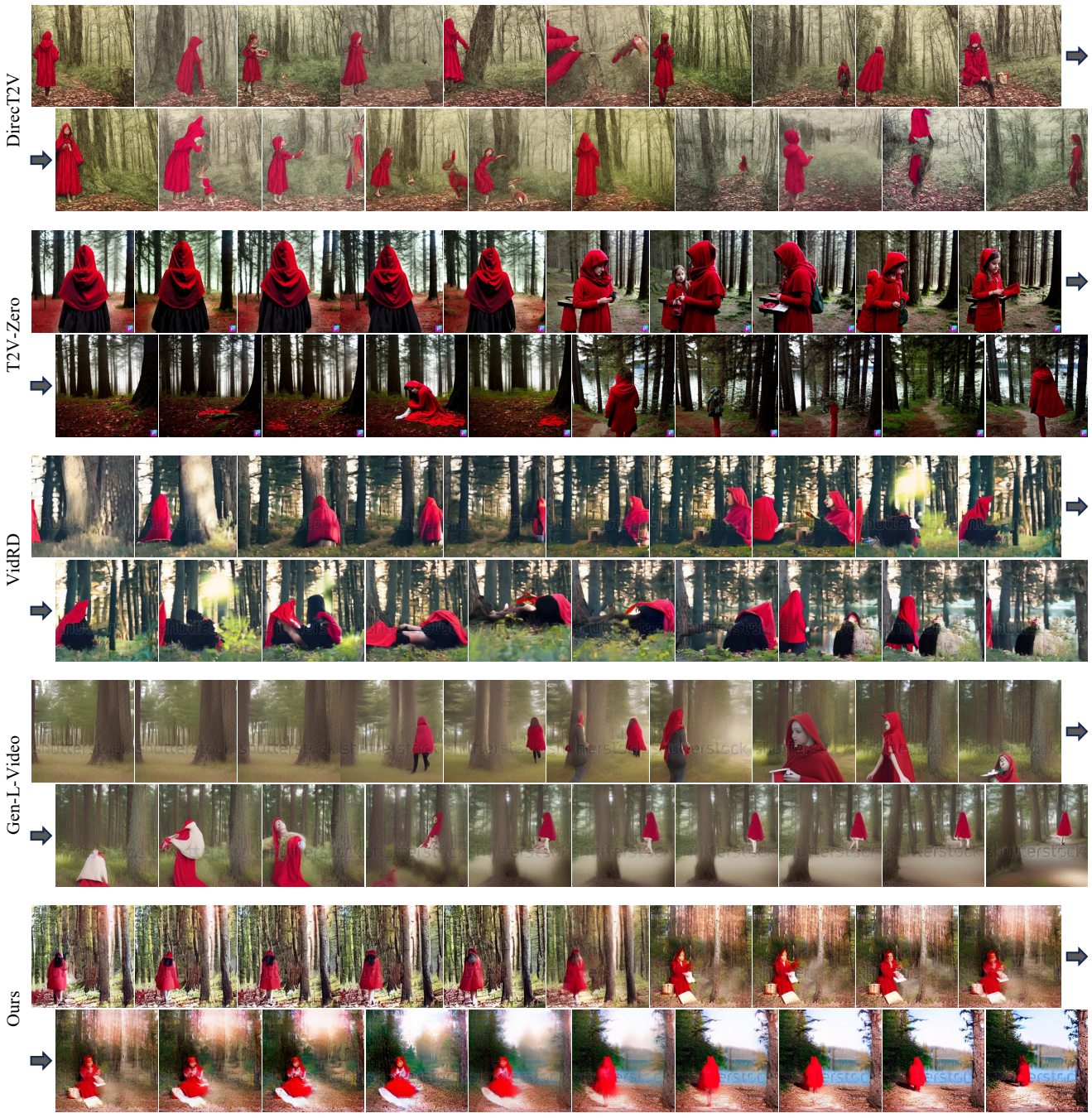


Figure 11. Qualitative comparisons with DirecT2V [18], T2V-Zero [22], VidRD [10], and Gen-L-Video [40]

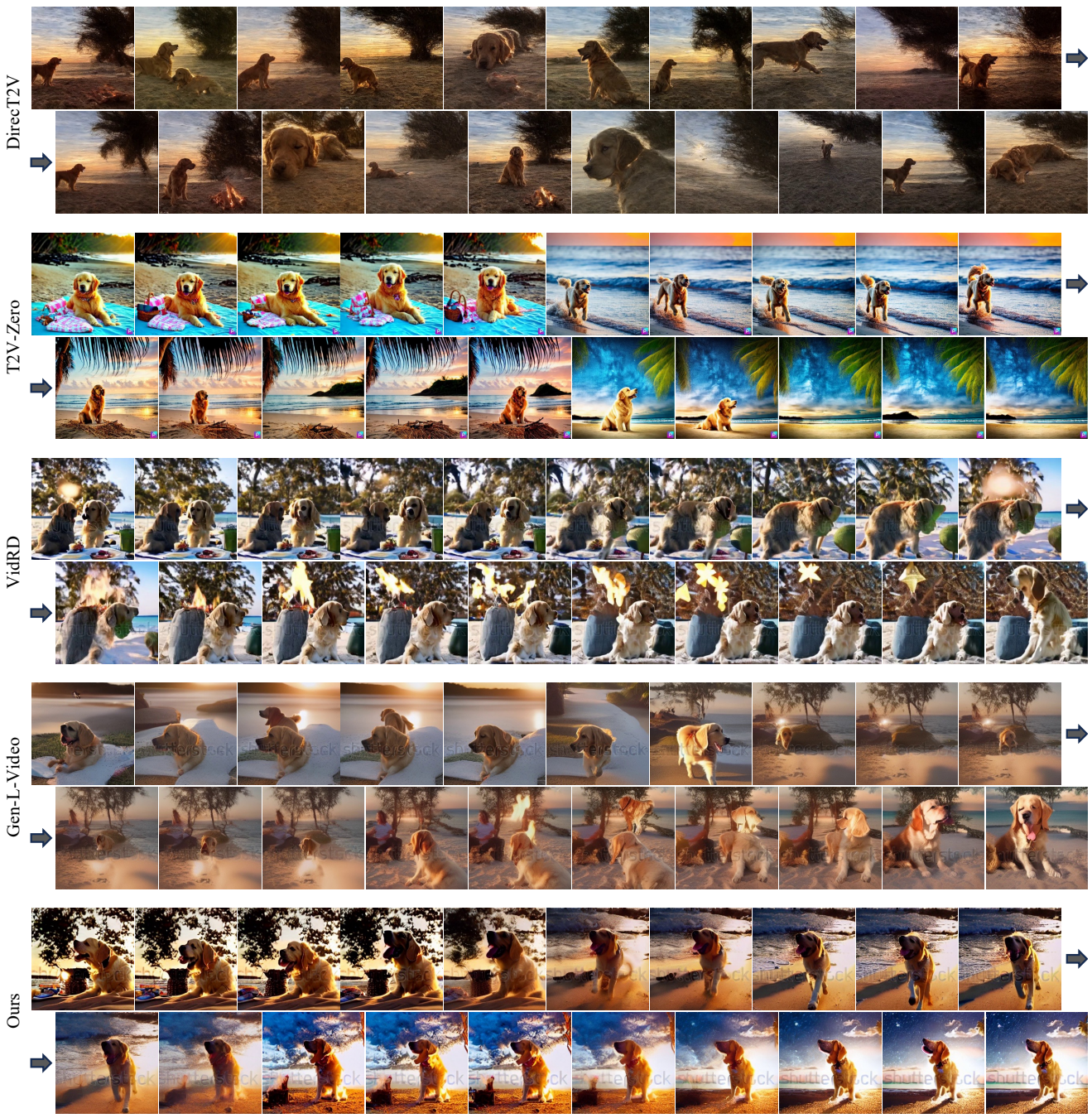


Figure 12. Qualitative comparisons with DirecT2V [18], T2V-Zero [22], VidRD [10], and Gen-L-Video [40]

PROMPTS:

Santa Claus goes snowboarding on a snowy mountain.
Santa Claus rides his sleigh through the snow in the mountain.
Santa Claus walks through the forest to a frozen lake.
Santa Claus has fun skating on the ice.

Figure 13. Qualitative result conditioning on multi-text with LVDM [12].

PROMPTS:
A white dog is running in the beautiful meadow.
A white dog is standing in the beautiful meadow.
A white dog is yawning loudly in the beautiful meadow.
A white dog lies on the ground in the beautiful meadow.

Figure 14. Qualitative result conditioning on multi-text with LVDM [12].

PROMPTS:
A waterfall flows in the mountains under a clear sky.
A waterfall flows in the fall mountains under a clear sky.
A waterfall flows in the winter mountains under a clear sky.
A waterfall frozen on a mountain during a snowstorm.

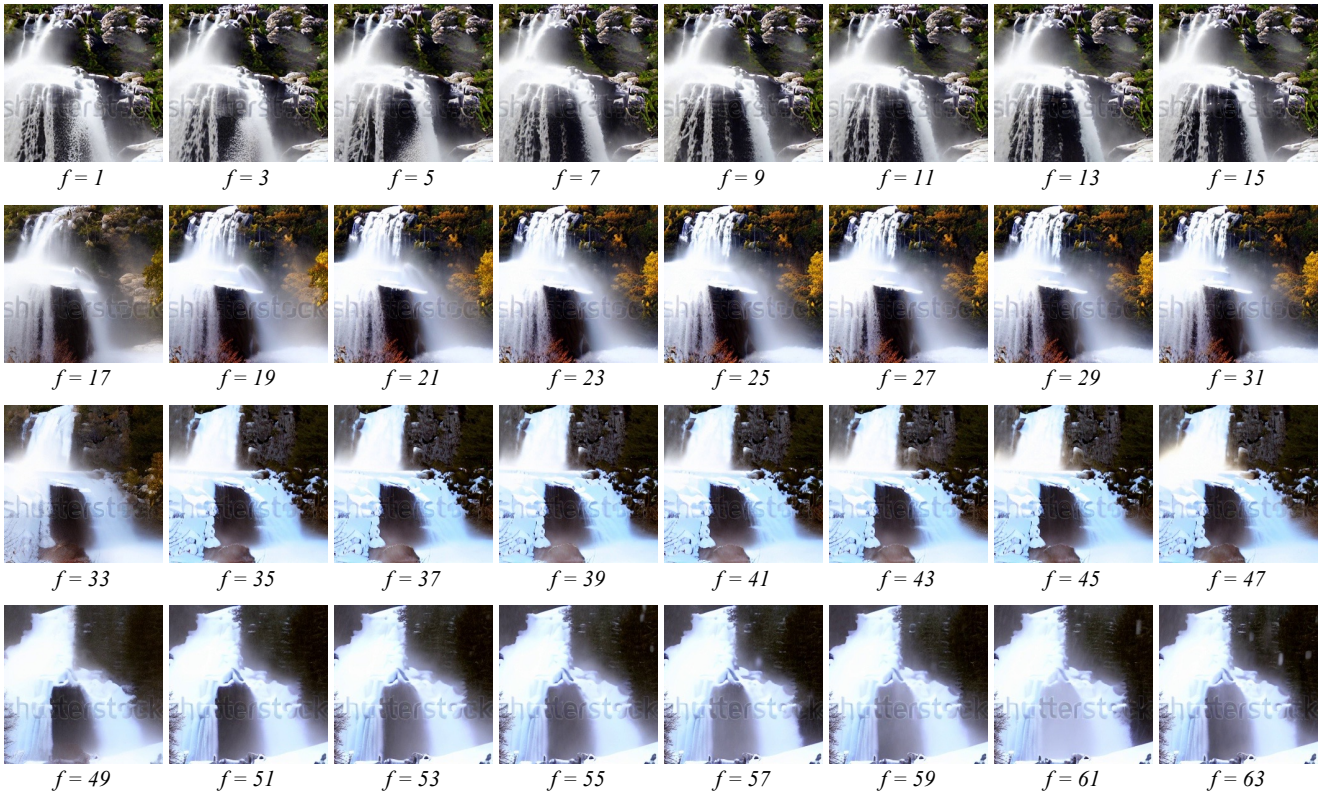


Figure 15. Qualitative result conditioning on multi-text with LVDM [12].

PROMPTS:
An astronaut in a white uniform is snowboarding in the snowy hill.
An astronaut in a white uniform is surfing in the sea.
An astronaut in a white uniform is surfing in the desert.

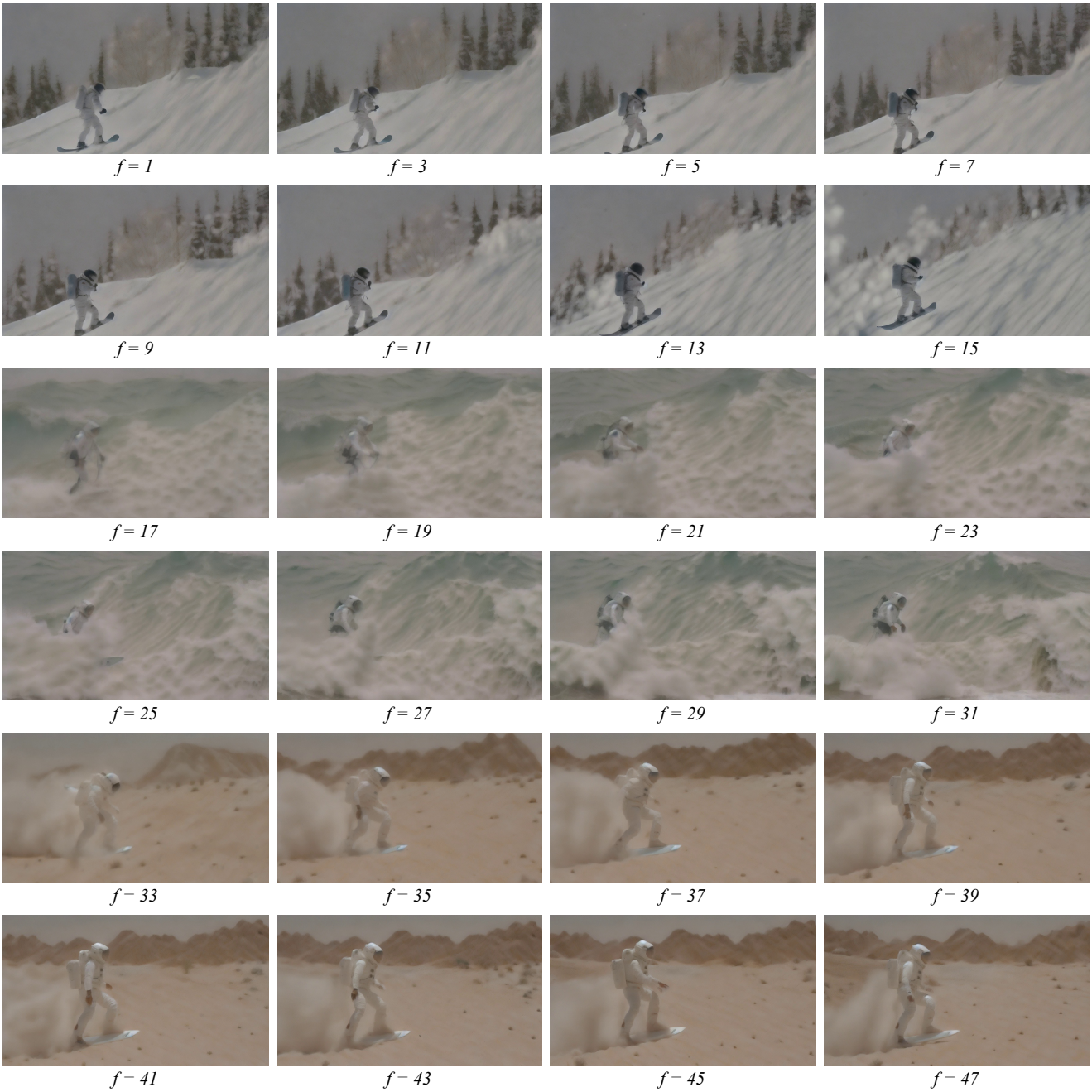
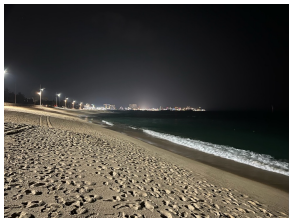


Figure 16. Qualitative result conditioning on multi-text with VideoCrafter1 [4].

PROMPTS:
A white dog is running in the beautiful meadow.
A white dog is standing in the beautiful meadow.
A white dog is yawning loudly in the beautiful meadow.
A white dog lies on the ground in the beautiful meadow.

Figure 17. Qualitative result conditioning on multi-text with VideoCrafter1 [4].



PROMPTS:

People walks on the beach at night.

There are sand castles on the beach under the fireworks at night.

Very few people remain on the beach at night and they gradually fade away.

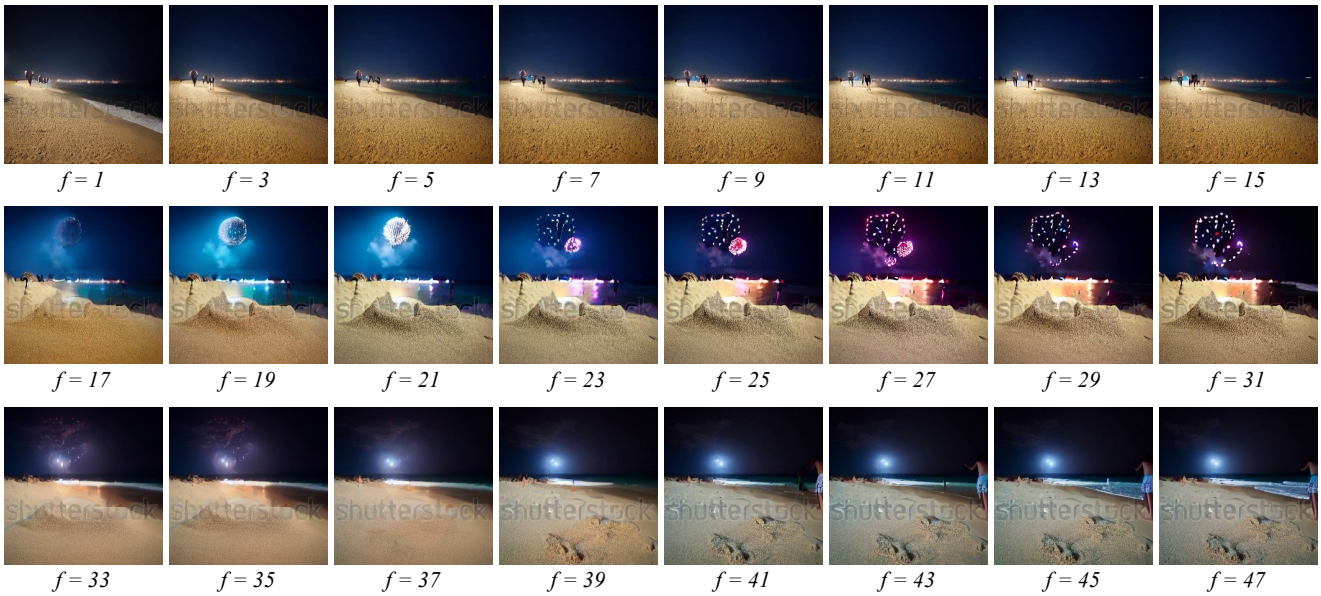


Figure 18. Example of generated video conditioning on image and multi-text.

PROMPTS:

The volcano erupts in the clear weather.

Smoke comes from the crater of the volcano, which has ended its eruption in the clear weather.

The weather around the volcano turns cloudy.

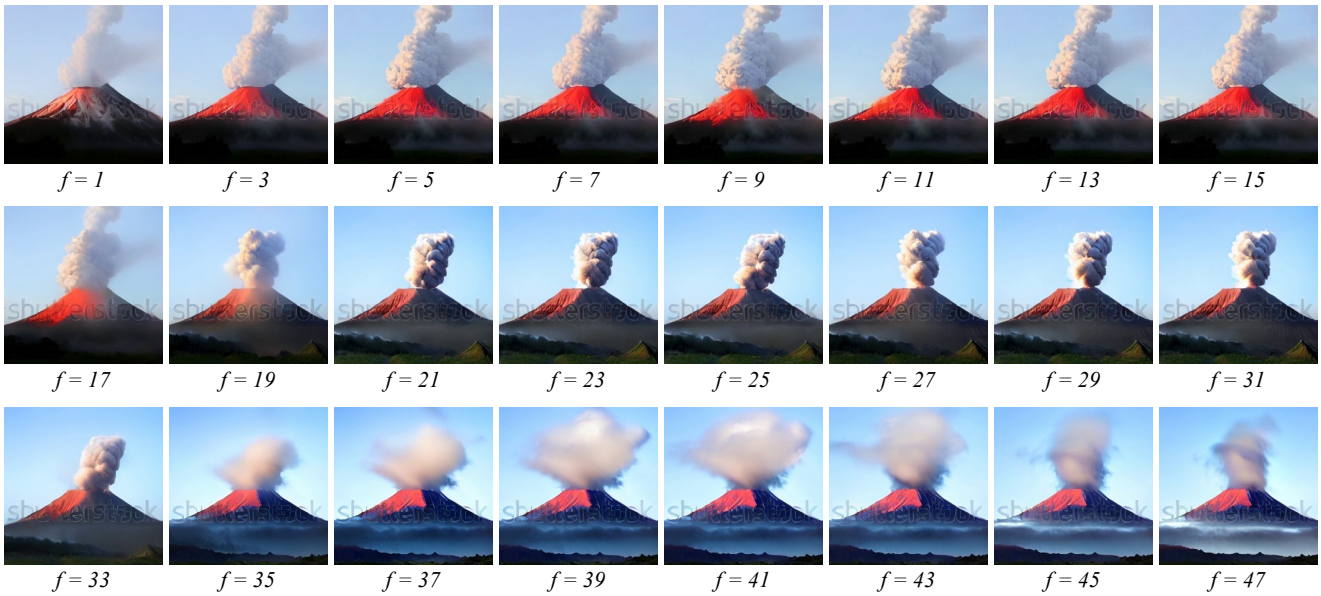


Figure 19. Example of generated video conditioning on image and multi-text.

SCENARIO: → Prompt Generator → PROMPTS:

Elmo is eating cookies in the square, sitting in the snowy square and smiling, and standing next to the snowman in the snowy square.

P1. Elmo is eating cookies in the snowy square.
P2. Elmo is sitting in the snowy square and smiling.
P3. Elmo is standing next to the snowman in the snowy square.

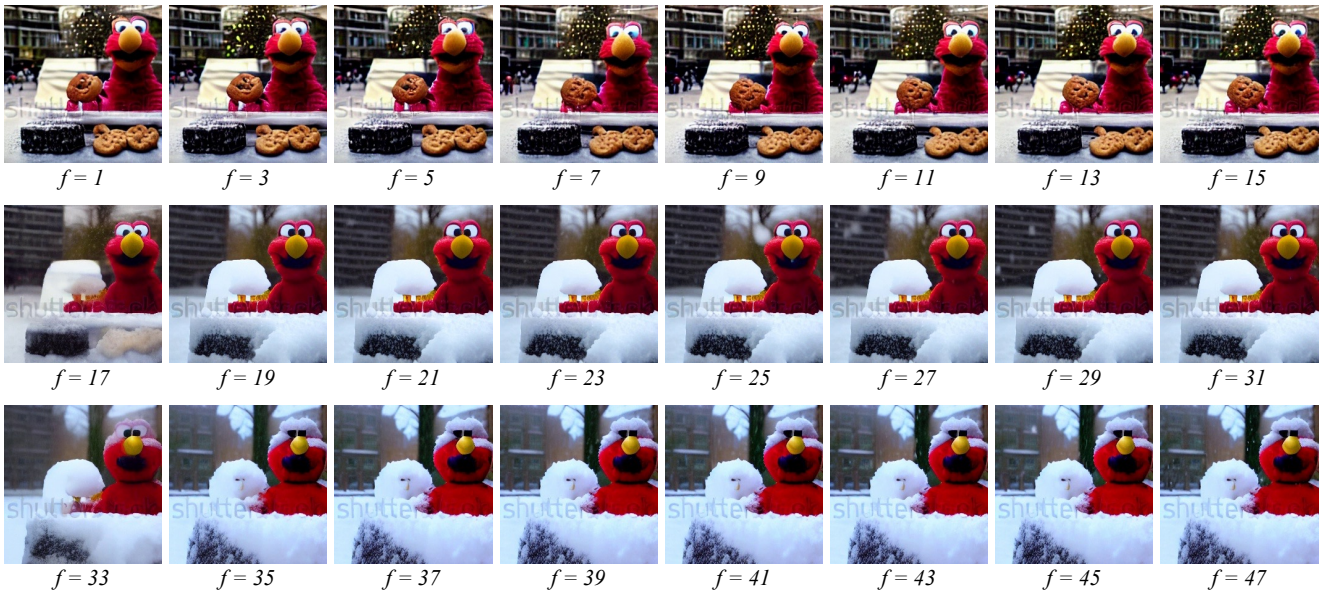


Figure 20. Examples of multi-text video generation utilizing the *prompt generator*.

SCENARIO: → Prompt Generator → PROMPTS:

Embarking on a motorcycle journey, the man weaves through a traffic jam, rides in the desert, walks under the night sky, and gazes at the aurora.

P1. A man embarks on a motorcycle journey.
P2. A man runs through a traffic jam on a busy road.
P3. A man rides a motorcycle in the desert.
P4. A man walks in the desert at night.
P5. A man looks at the sky with aurora in the desert.

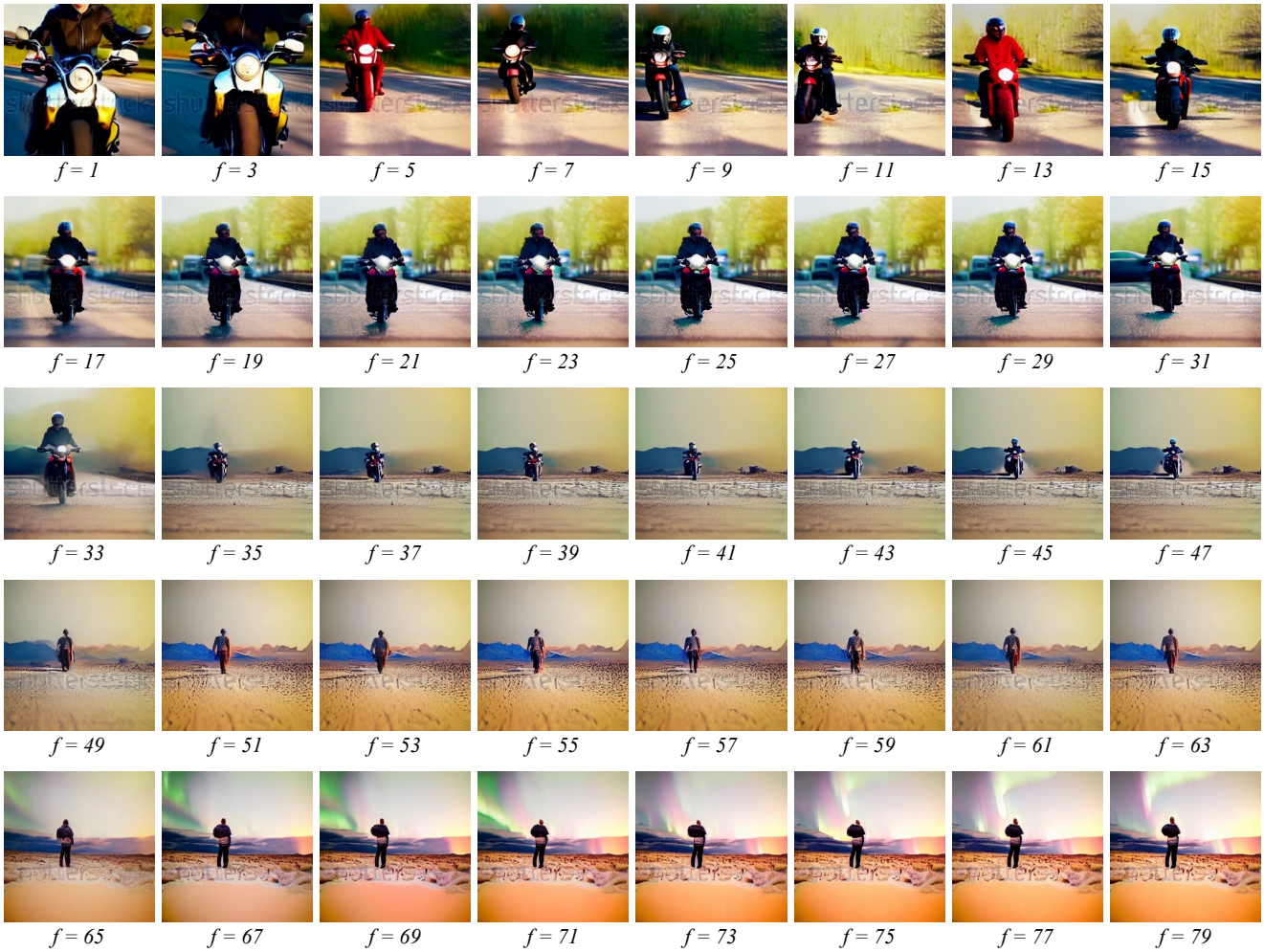


Figure 21. Examples of multi-text video generation utilizing the *prompt generator*.