Large Language Models are Frame-level Directors for Zero-shot Text-to-Video Generation

Susung Hong **Junyoung Seo** **Sunghwan Hong** **Heeseong Shin** **Seungryong Kim**

Korea University, Seoul, Korea

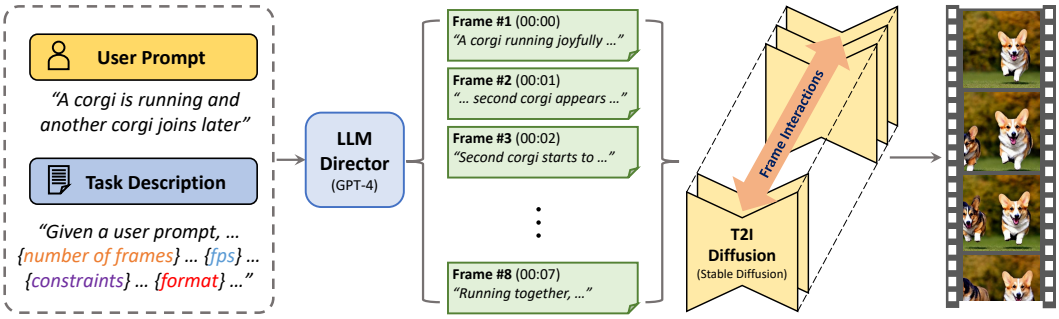

Figure 1: **Overall pipeline of the proposed DirecT2V framework.** Our framework consists of two parts: directing an abstract user prompt with an LLM director (GPT-4 [34]) and video generation with a modified T2I diffusion (Stable Diffusion [41]).

Abstract

In the paradigm of AI-generated content (AIGC), there has been increasing attention in extending pre-trained text-to-image (T2I) models to text-to-video (T2V) generation. Despite their effectiveness, these frameworks face challenges in maintaining consistent narratives and handling rapid shifts in scene composition or object placement from a single user prompt. This paper introduces a new framework, dubbed DirecT2V, which leverages instruction-tuned large language models (LLMs) to generate frame-by-frame descriptions from a single abstract user prompt. DirecT2V utilizes LLM directors to divide user inputs into separate prompts for each frame, enabling the inclusion of time-varying content and facilitating consistent video generation. To maintain temporal consistency and prevent object collapse, we propose a novel value mapping method and dual-softmax filtering. Extensive experimental results validate the effectiveness of the DirecT2V framework in producing visually coherent and consistent videos from abstract user prompts, addressing the challenges of zero-shot video generation. The code and demo shall be available at `https://github.com/KU-CVLAB/DirecT2V`.

1 Introduction

Within the paradigm of AI-generated content (AIGC), there has been increasing attention in expanding the capabilities of pre-trained text-to-image (T2I) models to text-to-video (T2V) generation [23, 48, 21, 48, 2, 61]. One notable advancement in this area is the Text2Video-Zero (T2V-Z) framework, which introduced a fine-tuning-free approach utilizing a pre-trained text-to-image diffusion model [41, 43] for generating videos from text descriptions [23]. Additionally, several other studies [23, 59, 38] have

Preprint. Under review.

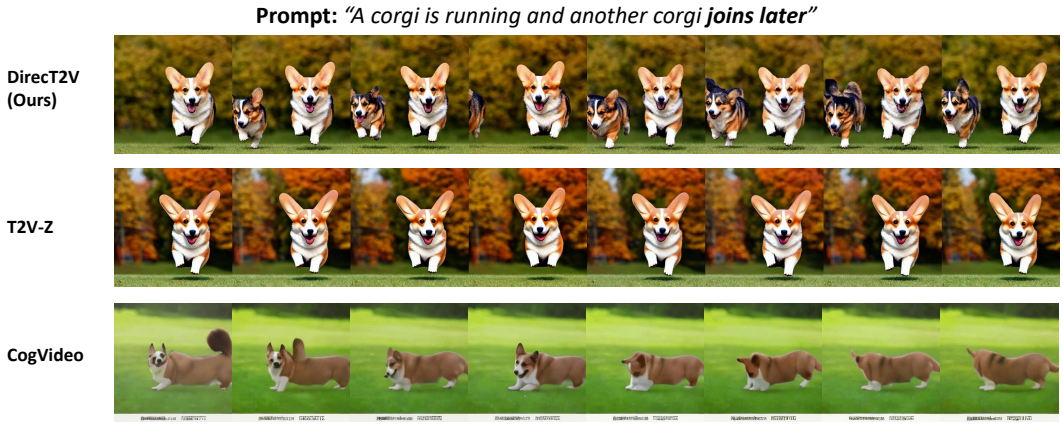

Figure 2: **Teaser.** DirecT2V, using LLMs as frame-level directors, enables zero-shot narrative text-to-video generation, while current zero-shot (Text2Video-Zero [23];T2V-Z) or tuned (CogVideo [21]) baselines do not contain high-level stories, *e.g.*, joining of the second corgi.

focused on enhancing temporal consistency in existing text-to-image diffusion models by redesigning the self-attention module, enabling video generation without the need for further training. These methods have successfully reduced the requirement for expensive fine-tuning processes, saving both time and resources while ensuring accessibility.

Although the zero-shot T2V generation frameworks [23, 59, 38] have shown effectiveness, they are not without challenges. One significant challenge is the conditioning of a single user prompt for all frames, which may struggle to maintain consistent narratives and varying contexts over time. The model's limited ability to comprehend the temporal dynamics of complex actions from a single prompt that only provides abstract information can result in videos that overlook important motions, actions, or events [23]. Moreover, accurately representing motions poses difficulties as the framework lacks effective mechanisms to handle rapid or substantial shifts in scene composition or object placement.

This paper proposes a solution to address the aforementioned challenges by leveraging instruction-tuned large language models (LLMs) [36, 57], such as GPT-4 [34] and PaLM2 [11], for generating frame-by-frame descriptions in video creation from a single abstract user prompt. The method, called DirecT2V, enables zero-shot video creation by utilizing carefully designed task prompts tailored for instruction-tuned LLMs. By leveraging the language understanding capabilities of LLMs [4, 7, 44], DirecT2V complements the abstract user prompts with specific descriptions to facilitate zero-shot frame generation. This is achieved by incorporating LLM directors to divide user inputs into separate prompts for each frame, implicitly separating static and dynamic elements within the user prompts. This separation allows for the integration of these components into text-to-image models, enabling the inclusion of time-varying content, which was previously unattainable, and facilitating consistent video generation.

Although the complemented prompt may help, creating temporally cohesive and visually captivating videos from a text-to-image model remains extremely challenging due to the stochasticity of diffusion models [15, 50, 22]. Therefore, in addition to high-level narrative constraints, we allow frame interactions to happen for achieving temporal coherence and flexibility between frames. In specific, we propose to adaptively adjust self-attention values in T2I diffusion models based on the timestep, which we call value mapping, and apply dual-softmax for obtaining the confidence map within the self-attention layers to eliminate unreliable mapping between frames.

To validate our framework, we present extensive experimental results that demonstrate the effectiveness of the proposed methods in addressing the challenges of zero-shot video generation from abstract user prompts. Empirical results validate the effectiveness of the DirecT2V framework in producing visually coherent and consistent videos from abstract user prompts.

2 Related Work

Incorporating large language models. Large language models (LLMs), such as GPT-3 [4], PaLM [7], and BLOOM [44], have been shown to be effective in a wide range of tasks, *e.g.*, decision making [26], program synthesis [1], and prompt engineering [62]. Notably, their zero-shot capabilities have demonstrated strong generalization power that almost resembles the linguistic ability of humans. By transferring such knowledge, numerous methods [43, 25, 24, 4] have excelled at even tasks involving different modalities, *i.e.*, audio, text, and images. Specifically, a recently introduced technique called instruction-finetuning, achievable via supervision or RLHF [51, 8], enabled accurate manipulation of LLMs that aligns with human intent. Another line of works, including [3, 12], have proposed to combine pre-trained language models with diffusion-based generative models, aiming to generate prompts that produce more reliable results. Among them, InstructPix2Pix [3] carefully combined two large-scale models, GPT-3 [4] and Stable Diffusion [41], and enabled generating controllable images from a reliable set of prompts obtained from fine-tuned LLMs. This approach effectively combines LLMs' prompt handling capabilties and stable diffusion's powerful text-based image generation ability. In this paper, we take a step forward and show that not only the aforementioned capabilities of large-scale models can be inherited, but also extend to video generation task, where temporal consistency and modeling of actions are posed as additional challenges.

Text-to-video generation. In the stream of research on AI-generated content (AIGC), text-to-video generation has been receiving considerable attention as a forefront research area, exploring various methodologies to generate videos from textual inputs. Among them, some methods employ autoregressive transformers or diffusion processes [58, 55, 14, 16]. NÜWA [58] introduces a 3D transformer [54]-based encoder-decoder framework and aims to tackle various tasks, including text-to-video generation, while Phenaki [55] presents a bidirectional masked transformer for a video creation from arbitrary-length text prompt sequences. Similarly, Imagen Video [14] leverages diffusion models for cascading pipeline [16] and introduces a freamework to spatial and temporal super-resolution.

Notably, a recent trend is that owing to remarkable generation ability of large-scale text-to-image models, numerous methods attempted to transfer their knowledge and even extend to other tasks, including text-to-video generation. CogVideo [21] builds upon CogView2 [10], a text-to-image model, and employs a multi-frame-rate hierarchical training strategy, encouraging text and video alignment. Make-a-video [48] tackles T2V task more efficiently by using the synthetic data for self-supervision. Another line of works [18, 61] that exploit LDM [41] enables high-resolution video generation by introducing temporal tuning technique that efficiently fine-tunes the parameters. Taking a step forward, Text2Video-Zero (T2V-Z) [23] introduces tuning-free zero-shot video generation without requiring intensive training or large-scale video datasets.

Although aforementioned works may synthesize temporally consistent and high fidelity videos, it is notable that a single user prompt is responsible for the actions in all the frames in a video, making the output videos lacking story. As illustrated in Fig. 5, the time-dependent dynamics are often disregarded and only limitedly expressed, *i.e.*, translation of objects. In this work, we obtain user temporal contents-aware prompts from a single user prompt using an instruction-tuned LLM [36, 34, 11, 57], and use them to synthesize videos that successfully capture both static and dynamic components.

3 Method

3.1 Preliminaries

Numerous works in text-to-image (T2I) field, which include GLIDE [32], Dall-E 2 [40], latent diffusion models (LDM) [41] and Imagen [43], have been actively employing diffusion models for their high-fidelity generation. In this section, we first explain details of LDM [41] whose methods are adopted in Stable Diffusion.

LDM [41] is a diffusion model that performs the forward and reverse process within the latent space of an autoencoder denoted as $D(E(\cdot))$, where E and D symbolize the encoder and decoder, respectively. Given an input image $x \in \mathbb{R}^{H\times W\times 3}$ and its latent tensor $z_0 := E(x) \in \mathbb{R}^{h\times w\times c}$ where $h < H$ and $w < W$, during the forward process, Gaussian noises are progressively added to the

signal z_0, following

$$q(z_t|z_{t-1}) = \mathcal{N}(z_t; \sqrt{1-\beta_t}z_{t-1}, \beta_t I), \quad t = 1, \ldots, T, \tag{1}$$

where $q(z_t|z_{t-1})$ signifies the conditional density of z_t given z_{t-1}, and β_t for all t's are hyperparameters that defines the noise schedule. The forward process is repeated until the initial signal z_0 is entirely obscured, yielding $z_T \sim \mathcal{N}(0, I)$. The objective of the diffusion models is then to learn the reverse process defined as:

$$p_\theta(z_{t-1}|z_t) = \mathcal{N}(z_{t-1}; \mu_\theta(z_t, t), \Sigma_\theta(z_t, t)), \quad t = T, \ldots, 1, \tag{2}$$

which enables the discovering of a valid signal z_0 from the standard Gaussian noise z_T. To sample from $p_\theta(z_{t-1}|z_t)$, LDM [41] instead predicts the reparametrization $\epsilon_\theta(z_t, t)$. To achieve text-conditioned image sampling, the text embedding [39] of a user prompt ω is conditioned along the intermediate noised images within the cross-attention layers [54, 35, 13, 43, 41] and classifier-free guidance is leveraged for better alignment to user prompts. Moreover, to enhance the quality of the output images, self-attention is performed at different resolutions [15, 9, 20, 23] as an additional technique.

3.2 Motivation

Text-to-video generation is a challenging task, particularly in a zero-shot setting where video-specific priors like temporal consistency and motion realism cannot be learned due to the absence of video data during training. Moreover, the generated video should align with the provided text descriptions, capturing the narrative essence of the scene. Recently, T2V-Z [23] has tackled this task by lifting text-to-image (T2I) models to text-to-video (T2V) generation. In specific, motions are introduced by translating the latent tensor in x-/y-axis in the early process of generation. Then, temporal consistency is enforced by manipulating the self-attention mechanism after a predefined timestep T' in the reverse process.

Despite its noticeable performance, T2V-Z [23] is confined to having simple motions based on linear movements and stochastic perturbation from the diffusion process. Not only this is distant from motion realism, it also hinders the alignment between the video and the text description. For example, given the occasional presence or absence of objects within frames (*e.g.*, *"joining"* or *"exiting"*) and the representation of changing object states (*e.g.*, *"jump"* or *"explode"*). Therefore, when generating videos that align with user prompts, there is often a need to handle dynamic contents within the prompt, subject to narrative (plot) and temporal consistency.

To overcome these challenges, we find that pre-trained large language models (LLMs) can be exploited to transform abstract user prompts into frame-by-frame image descriptions, thereby addressing the multifaceted challenges of temporal consistency and visual quality. In the following, we itemize each components which can be captured by LLM:

- Actions can be identified and decomposed from the user prompts, which can be used to create visual representations of these actions across different frames.
- Object descriptions can be detected, which can be later described across the frames.
- Contextual information can be captured to create the background and context for the video.
- Camera angles and movements can be understood and inferred, which can provide a more dynamic and visually engaging video.
- Plot or storyline can be extracted, empowering the generation model to follow a coherent and logical sequence of events.

We show that each element is well expressed and it is illustrated in Fig. 3. In Sec. 3.3, we detail how the proposed method satisfies the aforementioned items.

In addition, another aspect that LDM [41] does not consider, mainly due to its stochasticity, but we aim to encompass is temporal consistency. Without consideration of such constraint, which is of prime importance in video generation tasks, the discontinuity between adjacent frames may occur. However, recent studies have shown that by sharing the same self-attention module within the U-Net of diffusion models, object rigidity can be achieved [23, 59, 38]. However, this approach heavily relies on the prior attention maps for the attention computation at arbitrary time-step, which enforces

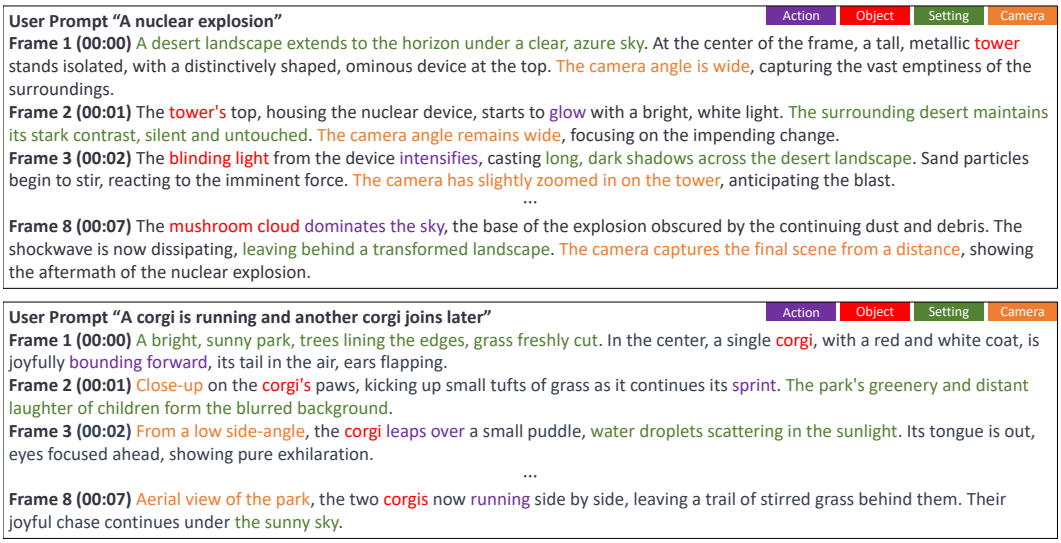

Figure 3: **Examples of frame-level directing with LLM.** Given an abstract user prompt, our LLM director outputs frame-wise prompts that complement the initial prompts with insufficient information. For complete instructions, see the appendix.

unchanging contexts, as exemplified in Fig. 5. To remedy this, we explore methods for interacting frames within the video in order to enhance the flexibility and overall quality of generated videos and also to preserve temporal consistency.

3.3 Frame-wise directing with LLM

In order to effectively leverage instruction-tuned LLMs [36, 34, 11, 57] for video generation, we claim that it is crucial to take into account the narrative consistency, in other words, the *storyline* reflected to the video. To achieve this goal, we propose a dynamic prompting strategy, in order to grant controllability over the desired attributes of the video without hurting the narratives. For this, we provide LLMs with a user prompt indicating the narrative for the overall scene with a task description to ensure the continuity of the narrative and controllability of the various attributes of the video, such as the number of frames and frames per second (FPS).

As shown in Fig. 2, given a prompt *"A corgi is running, and another corgi joins later,"* we expect the frame-level prompts to describe a single corgi in the earlier frames and to have two corgis in the latter frames. This is achieved by leveraging the LLM for complementing the user prompt by accounting for different items we mentioned in Sec. 3.2. We show the resultant generated prompts in Fig. 3.

3.4 Rotational value mapping

Given frame-level dynamic prompts that account for the story within the video, the remaining challenge for lifting T2I models for T2V is generating frames satisfying the temporal consistency. This requires adjacent frames to have similar time-invariant components, such as object appearances and the background, while still allowing temporal variations, such as movements, to happen. To address this challenge in a zero-shot manner, T2V-Z [23] propagates the key and value projection of the first frame of the video across every other frames. However, this constrains the context and content of the overall video to resemble the first frame, without the ability to distinguish time-variant/invariant components. Not only this limits the flexibility of the video, but also dissatisfies the narrative consistency as dynamic motions and transitions can be introduced to the scene through time-variant components.

To overcome these, we introduce Value Mapping (VM), a method that injects temporal consistency, while enabling the use of diverse contents, such as objects and textures, across the video frames. Different from prior works [23, 38], this method adjusts the value of self-attention in relation to the timestep, effectively preventing the objects visual collapsing and ensuring temporal consistency.

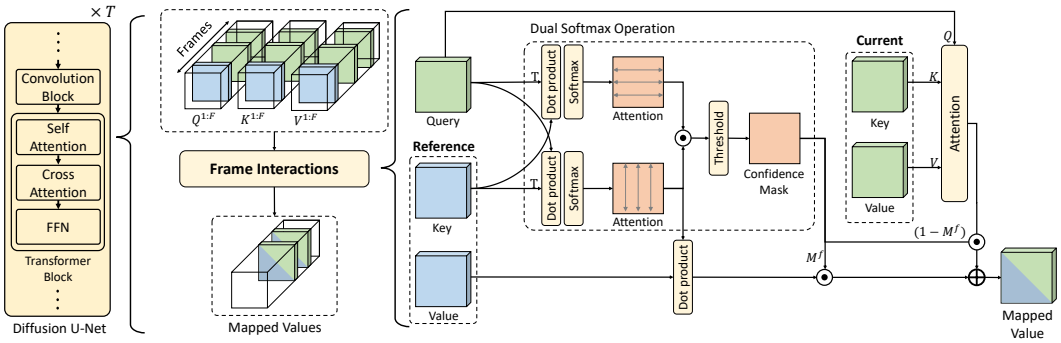

Figure 4: **The proposed frame interaction mechanism.** Within the self-attention layers in a diffusion U-Net [42], value mapping and dual-softmax are introduced to ensure temporal consistency and reduce unreliable matching between frames, respectively.

Given the formal definition of self-attention layer within a diffusion U-Net [15, 42]:

$$\text{Attention}(Q^f, K^f, V^f) = \text{Softmax}\left(\frac{Q^f (K^f)^{\mathrm{T}}}{\sqrt{d}}\right) V^f \quad \text{for} \quad \forall f, \tag{3}$$

where the notation Q^f, K^f, and V^f is the query, key, and value of the f-th frame, respectively, and we denote $X^{1:F} = [X^1, \ldots, X^F]$. In VM, we modify Eq. 3 to following:

$$\text{VM}(Q^f, K^{1:F}, V^{1:F}) = \text{Softmax}\left(\frac{Q^f (K^{r(t')})^{\mathrm{T}}}{\sqrt{d}}\right) V^{r(t')} \quad \text{for} \quad \forall f, \tag{4}$$

where $r : \{1, \ldots, T'\} \to \{1, \ldots, F\}$, and t' is the number of function evaluations from when the VM starts in the diffusion reverse process.

To decide the reference frame $r(t')$ for value mapping, we can simply consider randomly selecting $r(t')$ every timestep for allowing bidirectional flow of value mapping. However, we find this stochastic approach to often result in degenerate results due to the finite number of frames and timesteps. In this regard, we introduce Rotational Value Mapping (RVM), sequentially applying the value of self-attention based on the timestep by rotating over the frames periodically. Specifically, we define the reference frame as $r(t') = \text{Mod}(\lfloor t'/m \rfloor, F) + 1$, where m is a hyperparameter representing the period of timesteps. By setting m to a sufficiently large value, RVM becomes equivalent to the cross-frame attention in the original T2V-Z model [23].

3.5 Reducing unreliable matching via dual softmax filtering

Although VM can diversify the contextual information within the generated video, on its own, VM may face difficulties when accounting for rapid movements or drastic changes. In Eq. 4, VM enforces a mapping from $V^{r(t')}$ to V^f, where the query-key map $Q^f (K^{r(t')})^{\mathrm{T}}$ can be viewed as a correspondence map [52] that establishes a matching between source frame $r(t')$ and target frame f. However, when drastic changes happen between frames, there may exist cases where a reliable matching cannot be established, as an object may not co-occur between the frames. This restricts the target frame from incorporating attributes that are absent in the source frame, thereby preventing desired variations.

To address the issue of unreliable matching, we propose a means for mapping values when a reliable correspondence is established [30, 6, 53, 47]. This allows us to account for unreliable matching by propagating the original value of the target frame V^f instead of mapping from the reference frame, $V^{r(t')}$. To derive confidence values, we follow the dual softmax method [6] and then apply a threshold to these confidence values using a specified quantile. Starting from Eq. 4, the dual softmax [6], denoted as C_{d}, is defined as follows:

$$C_{\text{dual}} = \text{Softmax}(Q^t (K^{r(t')})^{\mathrm{T}}) \odot \text{Softmax}(K^{r(t')} (Q^t)^{\mathrm{T}}), \tag{5}$$

where $\odot$ represents the Hadamard product. This is followed by thresholding and masking to map only the reliable values:

$$\text{VM}'(Q^f, K^{1:F}, V^{1:F}) = (1 - M^f) \odot \text{Attention}(Q^f, K^f, V^f) + M^f \odot \text{VM}(Q^f, K^{1:F}, V^{1:F}) \tag{6}$$

for all f's, where $M^f = \mathbb{1}(C_{\text{dual}} > \phi)$, and ϕ is a pre-defined quantile of C_{dual}. This method allows only confident inter-frame matching, reflecting desired variances while preventing distortion throughout the video sequence.

4 Experiments

4.1 Implementation details

In this work, we employ GPT-4 [34] as our instruction-tuned LLM and T2V-Z [23], utilizing a single NVIDIA 3090 RTX GPU for efficient video sampling. For the generation process, we employ the PNDM scheduler [29], which is a member of the deterministic diffusion samplers family [22, 49, 29]. We configure the scheduler parameters with $T = 100$ and $T' = 96$. Furthermore, we adopt the classifier-free guidance [17], using a scale of 12.0. Both T2V-Z and our method employ motion dynamics; however, for a fair comparison, we refrain from using it in the main paper by setting its intensity to zero. We further visualize results with motion dynamics in the appendix. The code will be made publicly available.

4.2 Zero-shot video generation results

In Fig. 5, we showcase the zero-shot video generation capabilities of DirecT2V. Our framework generates per-frame prompts based on a user's description of a general scene, and the prompts are used to depict dynamic actions and time-varying content in the produced videos.

From the results, given a prompt *"A corgi is running and another corgi joins later,"* DirecT2V successfully portrays the second corgi appearing in the intermediate frames. In contrast, the second corgi is either always present or entirely absent for T2V-Z. As shown for the other prompt, *"A rainbow forming after a rain shower,"* demonstrates our method's ability to synthesize videos with narrative consistency, whereas T2V-Z with just the user prompt shows the rainbow exists for every frame.

Notably, the generated video from CogVideo [21] also lacks some components of the given user prompts, even though the model is fine-tuned using text-video pairs. For example, the addition of another corgi or a change in weather is not incorporated into the generated videos. These results corroborates the effectiveness of our frame-wise prompting approach.

4.3 Controlling video attributes with LLMs

In this section, we show that by providing specific instructions to LLMs, we can control the attributes of the video, such as the number of frames and frames per second (FPS). For lifting FPS, we provide instructions to the LLM to divide the prompts, and use it to generate videos similar to the method described above. This approach effectively handles situations where the number of frames exceeds the batch size. For controlling attributes other than FPS, we provide further results in the appendix.

To divide the prompt, the original frame-wise director extracts F frames at a specified frame rate R. We then provide a prompt that says, *"Now, at a frame rate of* $\{2 \times R\}$ *fps, divide each frame in the previous result into two separate image descriptions. This should eventually result in* $\{2 \times F\}$ *frames."* We repeat this process with the new $F := 2F$.

For the caching process, if the number of frames exceeds the batch size, we divide it into sections with a size of floored half of the batch size. Initially, we perform rotational value mapping for the frames, the number of which is the floored half of the batch size. Subsequently, we utilize the cached attention from the first process to perform rotational value mapping for the full batch size. The results are displayed in Fig. 6.

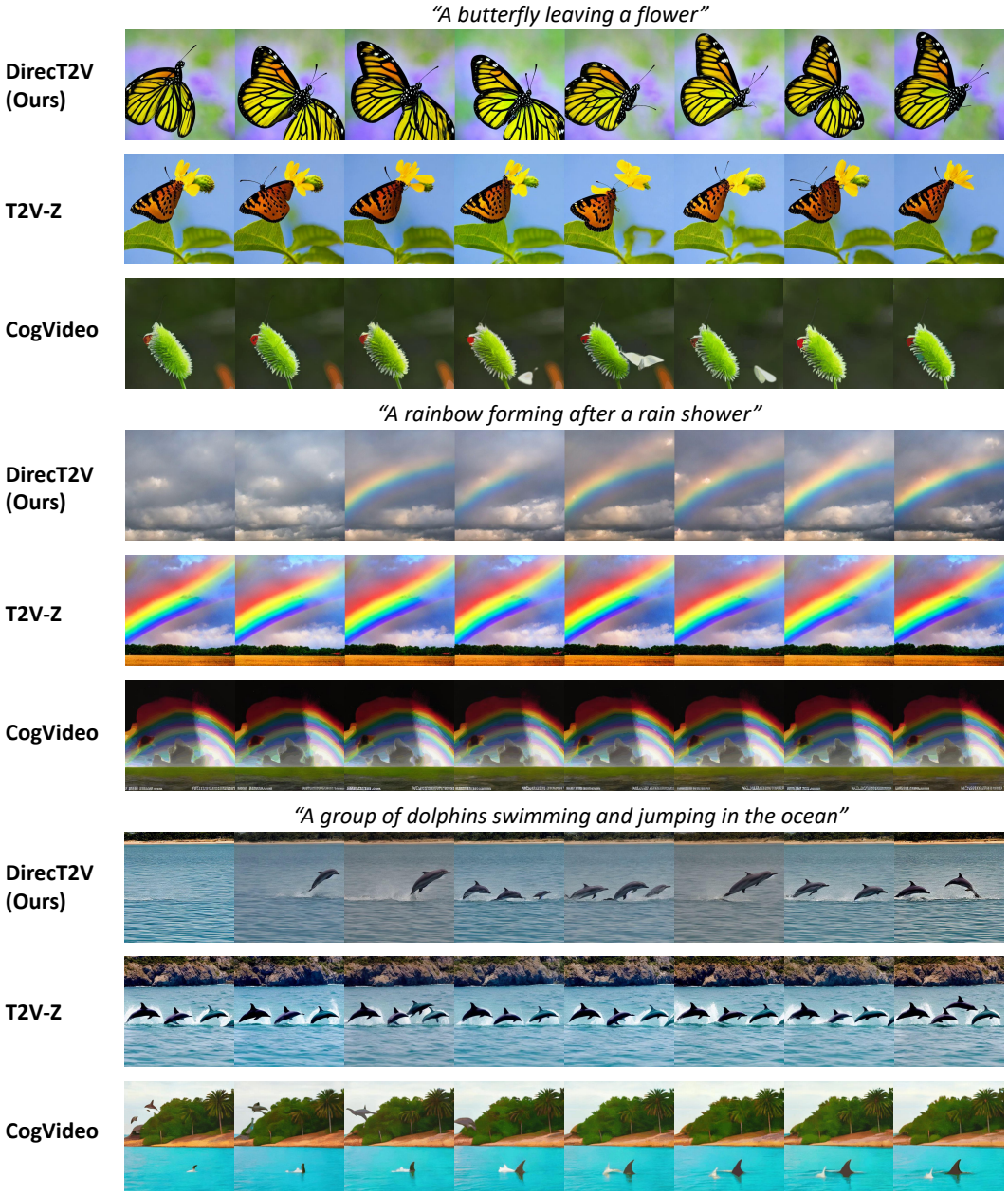

Figure 5: **Zero-shot video generation results.** Given an abstract user prompt, we compare DirecT2V with T2V-Z [23] and CogVideo [21]. Note that CogVideo is trained with video-text dataset, instead of zero-shot generation.

Figure 6: **Lifted frame rate.** By iteratively dividing frame-wise prompts, we can generate a video with an arbitrary frame rate.

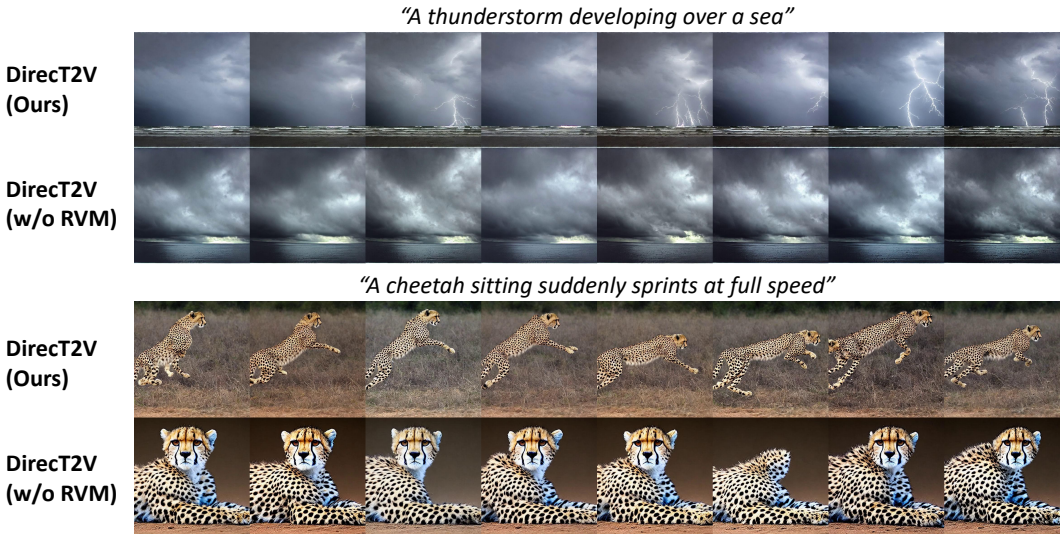

Figure 7: **Ablation study on RVM.** With frame-wise directing, RVM is essential for achieving narrative consistency, since the ablated counterpart (without RVM) does not reflect the frame-wise prompt well.

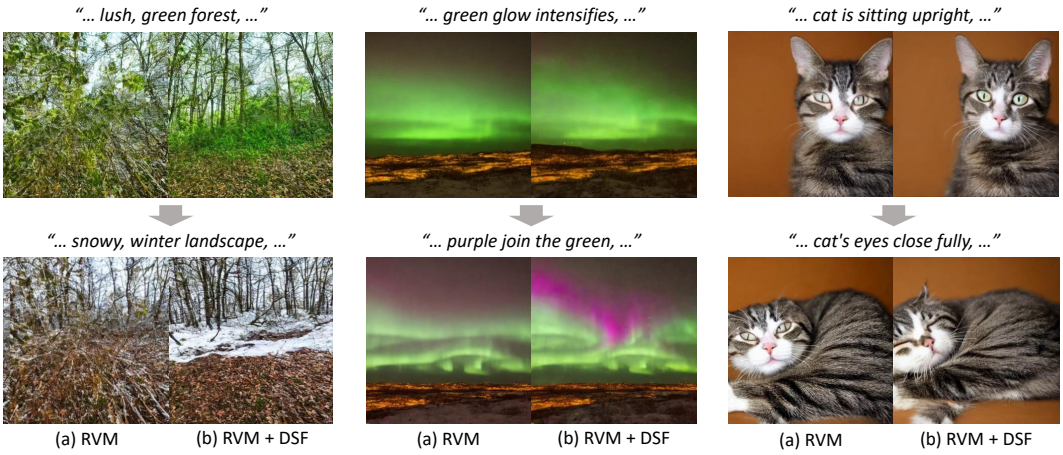

Figure 8: **Results of dual-softmax filtering.** For each prompts, results on the left are only with RVM, while the right are with RVM and Dual-Softmax Filtering (DSF).

4.4 Ablation study

Rotational value mapping. In Fig. 7, we display the ablation study for our Rotational Value Mapping (RVM) approach. This experiment showcases the generated video outcomes alongside the results obtained without employing rotational selection, *i.e.*, utilizing only the keys and values of the initial frame, as seen in [23].

When given a prompt like *"A thunderstorm developing over a sea,"* DirecT2V effectively depicts the thunderstorm's subsequent progression, as described by the GPT-4-generated prompt for that specific frame. On the other hand, the ablated version, with a fixed value in the second row, fails to develop a thunderstorm using the given fixed value. For another prompt, *"A cheetah sitting suddenly sprints at full speed,"* our technique illustrates its capacity to create videos that maintain narrative coherence. In contrast, DirecT2V without RVM leads to a persistent first-frame prompt. These findings emphasize RVM's ability to generate videos featuring dynamic actions and preserved time-varying content.

Dual-softmax filtering. In the experiment depicted in Fig. 8, we present the outcomes obtained when applying dual-softmax filtering, which successfully addresses the negative impact of inaccurate matches. For this study, we select two separate frames from the generated video and exhibit them along with the prompt proportion. It is worth noting that in the third column, the absence of dual-softmax filtering results in inevitable inaccurate matching due to value mapping, leading to the cat's eyes from a different frame being mapped onto the closed eyes.

5 Conclusion

In this paper, we have presented a novel approach for zero-shot video creation from textual prompts, tackling the intricate challenges of maintaining temporal consistency and visual quality in the generated videos. By employing GPT-4, a state-of-the-art instruction-tuned language model, we demonstrated its capability to generate detailed and temporally consistent image descriptions, which can be effectively integrated into text-to-image models. We introduced two key innovations for frame interactions, rotational value mapping and dual softmax filtering, which significantly enhance the flexibility and overall quality of the generated videos. These techniques help maintain object consistency, prevent distortion, and ensure temporal coherence throughout the video. Through extensive experiments, we showcased the effectiveness of our approach in generating visually compelling and temporally consistent videos from textual prompts. Our method's versatility and adaptability make it suitable for extension to other related tasks, broadening its potential impact across various applications and research domains. In conclusion, our work represents a significant advancement in the field of zero-shot video generation, setting a new benchmark and paving the way for future research and applications.

References

[1] Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021.

[2] 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. *CVPR*, 2023.

[3] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. 2023.

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

[5] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. *arXiv preprint arXiv:2303.13873*, 2023.

[6] Xing Cheng, Hezheng Lin, Xiangyu Wu, Fan Yang, and Dong Shen. Improving video-text retrieval by multi-stream corpus alignment and dual softmax loss. *arXiv preprint arXiv:2109.04290*, 2021.

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

[8] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

[9] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *NeurIPS*, 34:8780–8794, 2021.

[10] Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *arXiv preprint arXiv:2204.14217*, 2022.

[11] Google. Palm 2 technical report, 2023.

[12] Yaru Hao, Zewen Chi, Li Dong, and Furu Wei. Optimizing prompts for text-to-image generation. *arXiv preprint arXiv:2212.09611*, 2022.

[13] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. 2022.

[14] 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.

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 33:6840–6851, 2020.

[16] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *Journal of Machine Learning Research*, 23(47):1–33, 2022.

[17] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

[18] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv preprint arXiv:2204.03458*, 2022.

[19] Susung Hong, Donghoon Ahn, and Seungryong Kim. Debiasing scores and prompts of 2d diffusion for robust text-to-3d generation. *arXiv preprint arXiv:2303.15413*, 2023.

[20] Susung Hong, Gyuseong Lee, Wooseok Jang, and Seungryong Kim. Improving sample quality of diffusion models using self-attention guidance. *arXiv preprint arXiv:2210.00939*, 2022.

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

[22] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *NeurIPS*, 2022.

[23] 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.

[24] Yuma Koizumi, Yasunori Ohishi, Daisuke Niizumi, Daiki Takeuchi, and Masahiro Yasuda. Audio captioning using pre-trained large-scale language model guided by audio-based similar caption retrieval. *arXiv preprint arXiv:2012.07331*, 2020.

[25] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022.

[26] Shuang Li, Xavier Puig, Chris Paxton, Yilun Du, Clinton Wang, Linxi Fan, Tao Chen, De-An Huang, Ekin Akyürek, Anima Anandkumar, et al. Pre-trained language models for interactive decision-making. *Advances in Neural Information Processing Systems*, 35:31199–31212, 2022.

[27] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. *arXiv preprint arXiv:2301.07093*, 2023.

[28] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. *arXiv preprint arXiv:2211.10440*, 2022.

[29] Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. Pseudo numerical methods for diffusion models on manifolds. *ICLR*, 2022.

[30] Oisin Mac Aodha, Ahmad Humayun, Marc Pollefeys, and Gabriel J Brostow. Learning a confidence measure for optical flow. *IEEE transactions on pattern analysis and machine intelligence*, 35(5):1107–1120, 2012.

[31] Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, and Daniel Cohen-Or. Latent-nerf for shape-guided generation of 3d shapes and textures. *arXiv preprint arXiv:2211.07600*, 2022.

[32] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

[33] OpenAI. Introducing chatgpt, 2022.

[34] OpenAI. Gpt-4 technical report, 2023.

[35] Hadas Orgad, Bahjat Kawar, and Yonatan Belinkov. Editing implicit assumptions in text-to-image diffusion models. *arXiv preprint arXiv:2303.08084*, 2023.

[36] 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.

[37] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.

[38] 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 preprint arXiv:2303.09535*, 2023.

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

[40] 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.

[41] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022.

[42] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[43] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.

[44] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022.

[45] Hoigi Seo, Hayeon Kim, Gwanghyun Kim, and Se Young Chun. Ditto-nerf: Diffusion-based iterative text to omni-directional 3d model. *arXiv preprint arXiv:2304.02827*, 2023.

[46] Junyoung Seo, Wooseok Jang, Min-Seop Kwak, Jaehoon Ko, Hyeonsu Kim, Junho Kim, Jin-Hwa Kim, Jiyoung Lee, and Seungryong Kim. Let 2d diffusion model know 3d-consistency for robust text-to-3d generation. *arXiv preprint arXiv:2303.07937*, 2023.

[47] Junyoung Seo, Gyuseong Lee, Seokju Cho, Jiyoung Lee, and Seungryong Kim. Midms: Matching interleaved diffusion models for exemplar-based image translation. *arXiv preprint arXiv:2209.11047*, 2022.

[48] 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.

[49] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021.

[50] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2020.

[51] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020.

[52] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020.

[53] Prune Truong, Martin Danelljan, Luc Van Gool, and Radu Timofte. Learning accurate dense correspondences and when to trust them. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5714–5724, 2021.

[54] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017.

[55] 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.

[56] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. *arXiv preprint arXiv:2212.00774*, 2022.

[57] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

[58] 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 *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVI*, pages 720–736. Springer, 2022.

[59] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Weixian Lei, Yuchao Gu, 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. *arXiv preprint arXiv:2212.11565*, 2022.

[60] Jiale Xu, Xintao Wang, Weihao Cheng, Yan-Pei Cao, Ying Shan, Xiaohu Qie, and Shenghua Gao. Dream3d: Zero-shot text-to-3d synthesis using 3d shape prior and text-to-image diffusion models. *arXiv preprint arXiv:2212.14704*, 2022.

[61] 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.

[62] Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. Large language models are human-level prompt engineers. *arXiv preprint arXiv:2211.01910*, 2022.

Appendix

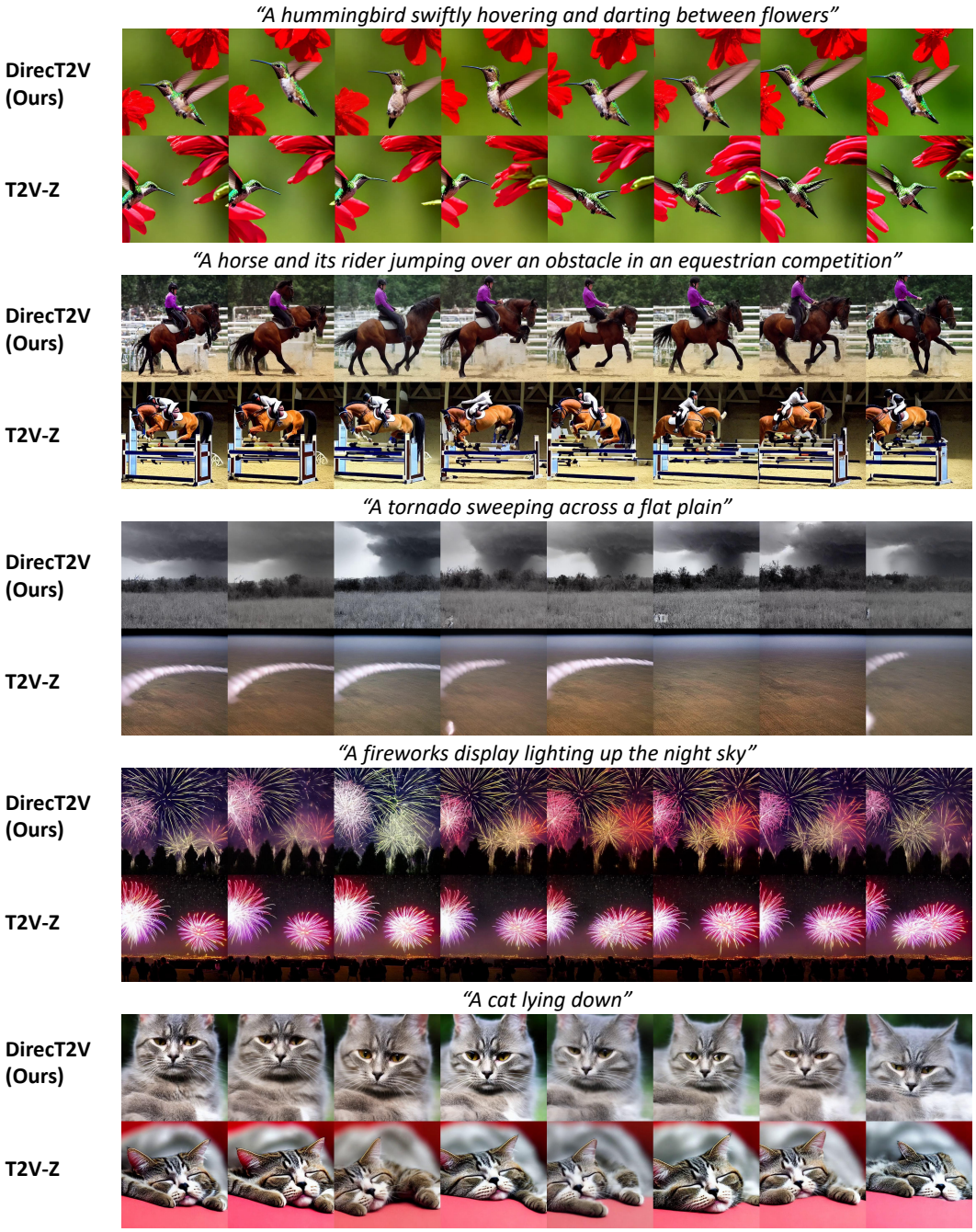

Figure 9: **Zero-shot video generation results with motion dynamics [23].** We compare our method with T2V-Z [23].

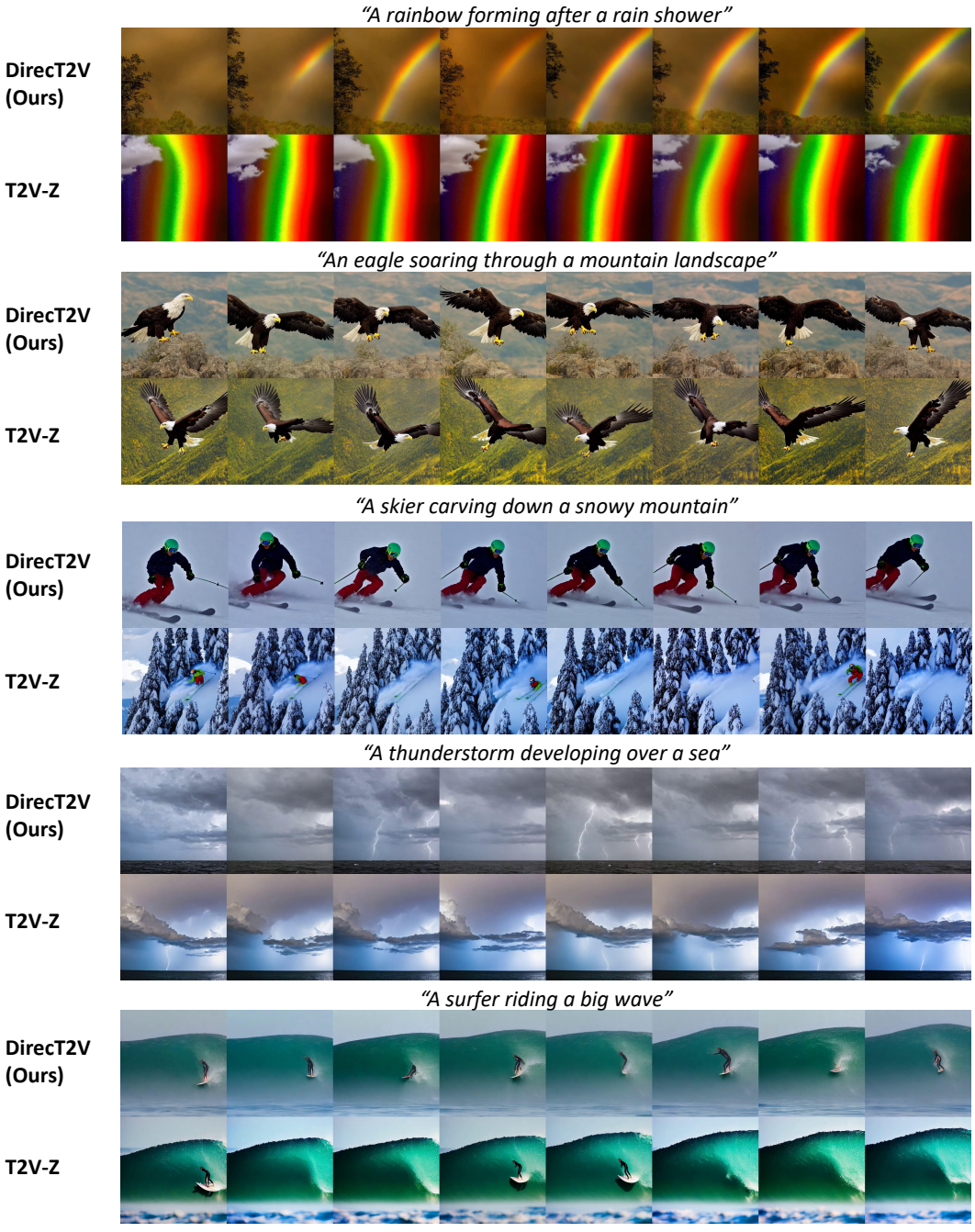

Figure 10: **Zero-shot video generation results with motion dynamics [23].** We compare our method with T2V-Z [23].

A Zero-shot video generation results with motion dynamics

Our framework can naturally be extended to provide motion dynamics [23], a feature that enables the capturing of explicit camera movement, *i.e.*, translation. Given accurately predicted motion dynamics, our approach would not only encapsulate the context of the narrative but also dynamically move with it. The results are provided in Fig. 9 and Fig. 10, and the video is available on our homepage.

B Experimental setting

We present examples of the full prompts used to produce the results shown in Fig. 2 of the main paper. These examples are depicted in Fig. 11. Additionally, in Fig. 12, we provide examples of the frame-wise prompts utilized in creating videos.

Complete instruction for Fig 2, including task description Action Object Setting Camera

Given a user prompt, envision a video scene and create eight distinct brief image descriptions at a frame rate of 1 fps. Ensure that each description is self-contained, encompassing all objects and their properties within the scene. The difference between two adjacent descriptions must be small, considering the small interval. Use the following format: ["", "", ..., ""].

Before you write each description, you must follow these instructions. These are of primary importance:

1. Identify and decompose specific actions or events included in the user prompts
2. Detect and describe objects mentioned in the user prompts
3. Describe environmental settings, which include contexts and backgrounds
4. Infer the desired camera angles and movements

Prompt: {user prompt}

Figure 11: **The complete instruction for Fig. 3, including the task description.**

A corgi is running and another corgi joins later	Frame #1 (00:00)	A corgi is running on a grassy field, its ears flopping as it moves.
	Frame #2 (00:01)	The corgi continues running, a second corgi starts to appear in the background.
	Frame #3 (00:02)	The second corgi starts to run, playfully chasing the first corgi.
	Frame #4 (00:03)	The first corgi maintains its pace, the second corgi getting closer.
	Frame #5 (00:04)	Both corgis are running side by side, their short legs moving quickly.
	Frame #6 (00:05)	The second corgi starts to take the lead, the first corgi following closely.
	Frame #7 (00:06)	Both corgis continue running, their tails wagging happily as they race.
	Frame #8 (00:07)	The first corgi begins to catch up, the two corgis running neck and neck.
A rainbow forming after a rain shower	Frame #1 (00:00)	Raindrops fall against a backdrop of cloudy sky, the smell of fresh rain in the air.
	Frame #2 (00:01)	The rain starts to lighten, the clouds beginning to break, allowing a few beams of sunlight through.
	Frame #3 (00:02)	A faint arc of color starts to form in the sky, the first hint of a rainbow.
	Frame #4 (00:03)	The colors become more vibrant as the sunlight strengthens, the rain shower coming to an end.
	Frame #5 (00:04)	A full arc of a rainbow becomes visible against the partly cloudy sky, each band of color distinct.
	Frame #6 (00:05)	The rainbow's colors become even more vibrant, reflected in the puddles left by the rain shower.
	Frame #7 (00:06)	The rainbow stands bright and complete against the clearing sky, a beautiful result of the passing storm.
	Frame #8 (00:07)	The scene ends with the full rainbow stretching across the sky, a symbol of calm after the storm.
A group of dolphins swimming and jumping in the ocean	Frame #1 (00:00)	A calm ocean surface, ripples indicating movement beneath the water.
	Frame #2 (00:01)	A dolphin's dorsal fin emerges from the water, followed by a sleek, silver body.
	Frame #3 (00:02)	The first dolphin leaps out of the water, arcing through the air.
	Frame #4 (00:03)	As the first dolphin re-enters the water, two more dolphins break the surface.
	Frame #5 (00:04)	The group of dolphins swim in unison, their bodies gliding effortlessly through the waves.
	Frame #6 (00:05)	Another dolphin leaps, twisting in the air before diving back into the sea.
	Frame #7 (00:06)	The dolphins continue their aquatic ballet, their jumps punctuating the oceanic rhythm.
	Frame #8 (00:07)	As the dolphins submerge, the ocean surface returns to its serene state, waiting for the next performance.

Figure 12: **Examples of frame-level prompts directed by LLMs.**

C More ablation studies and analyses

Attribute controls. In the main paper, we discuss the effects of regulation of frame rate control. In this section, we also provide our results for attribute control. These results are presented in Fig. 13. In this experiment, we demonstrate the capability to manipulate the number of frames effectively, as well as to remove or add the camera setting of the video scene from a simple prompt.

Number of frames (8 → 16)
1. A wide-angle shot of a verdant park, bustling with people and their pets. A small, enthusiastic Corgi enters the frame from the right, ears flopping and tail wagging.
2. The Corgi running at full speed across the grass, its small legs moving quickly. The camera pans to follow its motion.
3. The camera zooms in on the Corgi's determined face, its tongue hanging out due to the exertion, and eyes focused ahead.
4. The camera pulls back to a medium shot, showing the Corgi darting past a group of park-goers who stop to watch.
5. The Corgi slows down momentarily to sniff at a tree. Leaves rustle in the wind and people walk by in the background.
6. Suddenly, the Corgi perks up its ears, looking towards the left side of the frame. The camera follows its gaze.
7. Cut to a medium shot of another Corgi at the edge of the frame, standing a few meters away. This Corgi looks curious and eager.
8. The second Corgi starts running towards the first. Its fur is a little darker and it has a distinctive marking on its face.
9. The camera switches to a low-angle shot, capturing the second Corgi's joyful leap as it begins to sprint.
10. The two Corgis run towards each other, the green of the park providing a vibrant backdrop. The camera follows the second Corgi's movements.
11. A brief close-up of the first Corgi's surprised and delighted expression as it notices the other Corgi.
12. Both Corgis are now in the same frame, their paths meeting. They sniff each other in greeting, tails wagging excitedly.
13. A medium shot showing the two Corgis beginning to run together, bounding across the grass and away from the camera.
14. The camera tracks their movement as they chase each other around a tree, barking playfully.
15. An aerial shot shows the two Corgis circling the tree, their path weaving through the park-goers.
16. The final shot pulls back to a wide angle, capturing the whole park. The two Corgis continue to play, becoming small, joyous specks in the bustling scene.

Camera constraint (with → without)
1. An open grassy field, a bright day. In the center of the frame, a spirited Corgi, with its golden-brown coat, is in mid-stride, its small legs blurring with motion.
2. The Corgi's ears flap in the breeze as it continues its run, a trail of kicked-up grass particles floats behind it. The background reveals a distant wooden fence enclosing the field.
3. A fallen Frisbee appears in the scene, halfway between the running Corgi and the fence. The Corgi's focused eyes indicate its target.
4. Close up on the Corgi as it nimbly dodges a patch of wildflowers, its body slightly tilted to the side, showcasing its agility.
5. The Corgi reaches the Frisbee and chomps down on it, its eyes squinting as it does. Evident joy and accomplishment are on its face.
6. A movement at the edge of the frame; another Corgi, this one with a black and white coat, is shown behind the wooden fence, peeking through a gap.
7. The second Corgi, visibly excited, slips through the fence gap, its eyes fixed on the Frisbee-holding Corgi.
8. The two Corgis, now both in full run, are moving towards each other, the Frisbee held triumphantly by the first. Their tails are wagging, illustrating the joy of their shared activity.

Figure 13: **Video attribute control using LLMs.**

Comparison of GPT-4, GPT-3.5, and Bard. We compare the frame-level directing abilities of GPT-4 [34], GPT-3.5 [33], and Bard [11]. In Fig. 14, we present the conversations with those LLMs. We find that they generally follow the instruction prompt well.

Prompt:
Given a user prompt, envision a video scene and create eight distinct brief image descriptions at a frame rate of 1 fps. Ensure that each description is self-contained, encompassing all objects and their properties within the scene. The difference between two adjacent descriptions must be small, considering the small interval. Use the following format: ["", "", ..., ""].

Prompt: A corgi is running and another corgi joins later

GPT-4:
["A golden corgi stands alert on a green park lawn, a blue tennis ball at its feet.",
"A blur of golden fur as the corgi lunges forward, the blue tennis ball securely in its mouth.",
"The corgi, ball in mouth, streaks across the lawn leaving shallow imprints on the dewy grass.",
"From the corner of the frame, another corgi, this one tri-colored, enters, watching the golden corgi's progress intently.",
"The tri-colored corgi, with a wagging tail, dashes onto the lawn, giving chase to the golden corgi.",
"Both corgis now running parallel, their small legs moving in a blur against the verdant backdrop of the park.",
"Golden and tri-colored fur mix as the second corgi reaches the first, both bounding joyously with the blue ball still in the golden corgi's mouth.",
"Both corgis pause momentarily, panting lightly, before the golden corgi drops the ball, an invitation for the next round of the chase."]

GPT-3.5:
["A grassy field is stretched out under a bright blue sky. A lively corgi with short legs and a fluffy coat runs towards the camera, its ears flapping in the wind.",
"The first corgi continues its sprint, its paws kicking up small tufts of grass as it dashes across the field.",
"The second corgi, identical in appearance to the first, bounds into the frame from the left side, wagging its tail eagerly.",
"The two corgis now run side by side, their short legs propelling them forward in a synchronized motion.",
"Both corgis display excitement and joy as they frolic through the field, their tongues hanging out and happy barks escaping their mouths.",
"A gentle breeze rustles the nearby trees as the corgis continue their playful run, their fluffy coats catching the sunlight.",
"The two corgis effortlessly navigate through the field, their movements fluid and graceful despite their small stature.",
"With the vast expanse of the field behind them, the corgis sprint ahead, their tails wagging in unison, disappearing into the distance."]

Bard
["A corgi is standing in the middle of a field.",
"The corgi starts to run.",
"Another corgi enters the scene from the left.",
"The two corgis run towards each other.",
"The corgis collide and tumble to the ground.",
"The corgis get up and shake themselves off.",
"The two corgis continue to run together.",
"The corgis run out of the scene."]

Figure 14: **Comparison between frame-level prompting of GPT-4 [34], GPT-3.5 [33], and Bard [11].**

Rotational and random value mapping. We also test the random method of value mapping, which can replace our RVM. This means the reference frames are randomly selected contrary to RVM. The results, however, show that the method adopting random mapping obtains results with occasionally distorted objects, possibly due to the instability caused by the small number of timesteps.

Figure 15: **Examples of results from the different value mappings.**

D Limitations and broader impact

Limitations. The performance of the proposed method, DirecT2V, may vary depending on the instruction-tuned large language models (LLMs) [34, 33, 11] (see Fig. 14). As a result, any biases or limitations within these models may adversely affect the quality of the resulting videos. This is because LLMs can produce ambiguous or distracting descriptions, leading to less accurate or coherent video frames. Further research might explore the incorporation of additional constraints to create more vision-friendly frame-by-frame prompts. Moreover, DirecT2V's dependence on pre-trained text-to-image diffusion models introduces another layer of dependency. These models have encountered challenges in accurate counting and positioning [43, 27]. A potential solution to this problem could involve the use of the encoder from an even larger language model [43].

In essence, the development and enhancement of both instruction-tuned LLMs and T2I diffusion models equipped with attention mechanisms present promising landmarks for the future improvement of DirecT2V.

Broader impact. To the best of our knowledge, our research presents DirecT2V as the first framework that explicitly leverages the temporal and narrative knowledge embedded within large language models for high-level visual creation, specifically video creation [23, 48, 21, 48, 14, 55, 58, 61]. This philosophy can be extended to other high-level visual tasks, such as zero-shot text-to-3D [19, 37, 56, 31, 28, 5, 46, 60, 45]. Nevertheless, the ability of DirecT2V to generate realistic videos from textual prompts raises concerns about its potential to contribute to the spread of misinformation and deepfake content. The increased difficulty in distinguishing between authentic and fabricated videos may exacerbate existing concerns about the dissemination of false information.