Large Language Models are Frame-Level Directors for Zero-Shot Text-to-Video Generation

Susung Hong [1] **Junyoung Seo** [1] **Heeseong Shin** [1] **Sunghwan Hong** [1] **Seungryong Kim** [1]

Abstract

In the paradigm of AI-generated content (AIGC), there has been increasing attention to transferring knowledge from 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 shifts in scene composition or object placement from a single abstract user prompt. Exploring the ability of large language models (LLMs) to generate time-dependent, frame-by-frame prompts, this paper introduces a new framework, dubbed DirecT2V. DirecT2V leverages instruction-tuned LLMs as directors, enabling the inclusion of time-varying content and facilitating consistent video generation. To maintain temporal consistency and prevent mapping the value to a different object, we equip a diffusion model with a novel value mapping method and dual-softmax filtering.

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 (Khachatryan et al., 2023; Singer et al., 2022; Hong et al., 2022; Singer et al., 2022; Blattmann et al., 2023; Zhou et al., 2022a). One notable advancement in this area is the Text2Video-Zero (T2V-Z) framework, which introduced a fine-tuning-free approach utilizing a pre-trained T2I diffusion model (Rombach et al., 2022; Saharia et al., 2022) for generating videos from text descriptions (Khachatryan et al., 2023) in a zero-shot manner. Additionally, several other studies (Khachatryan et al., 2023; Wu et al., 2022b; Qi et al., 2023) have 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 these methods have proven effective (Khachatryan et al., 2023; Wu et al., 2022b; Qi et al., 2023), they are not without drawbacks. One significant challenge is the use of a single user prompt to condition all frames, which may struggle to maintain consistent narratives and varying contexts over time. In contrast to images, which can be described by one or a few sentences, videos contain sequences of time-varying actions and contexts, requiring much more descriptive information. This challenge is not fully addressed in such works, as they all assume a single prompt conditions the entire video. Consequently, the limited ability to comprehend the temporal dynamics of complex actions from a single prompt, which provides only abstract information, can result in videos that overlook important aspects such as motions, actions, or events.

To address these limitations, we devise DirecT2V, a novel framework that utilizes instruction-tuned large language models (LLMs) (Ouyang et al., 2022; Wei et al., 2021; OpenAI, 2023; Google, 2023) to generate frame-by-frame descriptions for zero-shot video creation from a single abstract user prompt. LLM directors divide user inputs into separate prompts for each frame, separating static and dynamic elements, enabling the integration of time-varying content into text-to-image models. To achieve temporal coherence and flexibility between frames, we propose methods that adaptively integrate into the self-attention mechanism in T2I diffusion models: value mapping, which selects an arbitrary frame and propagates the self-attention value to others based on diffusion timesteps, and dual-softmax filtering, which obtains and leverages confidence masks from the self-attention layers to eliminate unreliable mappings between frames, thereby addressing the challenges of creating temporally cohesive and visually captivating videos from text-to-image models.

[1]Korea Univeristy, Seoul, Korea. Correspondence to: Seungryong Kim <seungryong_kim@korea.ac.kr>.

First Workshop on Controllable Video Generation at ICML 2024, Vienna, Austria. Copyright 2024 by the author(s).

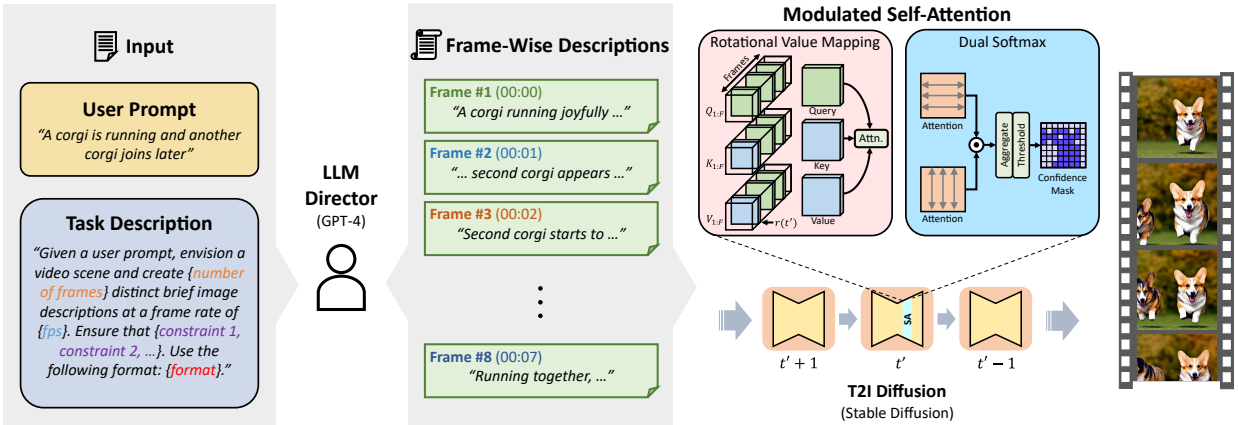

Figure 1. **Overall pipeline of our DirecT2V framework.** DirecT2V consists of two parts: directing an abstract user prompt with an LLM director (*e.g.*, GPT-4 (OpenAI, 2023)) and video generation with a T2I diffusion (*e.g.*, Stable Diffusion (Rombach et al., 2022)) equipped with modulated self-attention.

2. Method

2.1. Frame-wise directing with LLM

In order to effectively leverage instruction-tuned LLMs (Ouyang et al., 2022; OpenAI, 2023; Google, 2023; Wei et al., 2021) 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. 1, 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. **??**. We show the resultant generated prompts in Fig. 7.

2.2. Incorporating various contexts with 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 while harmonizing with descriptive frame-level prompts. 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. Partially addressing this challenge, T2V-Z (Khachatryan et al., 2023) 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 (Khachatryan et al., 2023; Qi et al., 2023), this method adjusts the value of self-attention in relation to the timestep, effectively preventing the objects visual collapsing and ensuring temporal consistency.

Given the formal definition of self-attention layer within a diffusion U-Net (Ho et al., 2020; Ronneberger et al., 2015):

$$\text{Attention}(Q_f, K_f, V_f) = \text{Softmax}\left(\frac{Q_f (K_f)^{\mathrm{T}}}{\sqrt{d}}\right) V_f \quad (1)$$

for all frames f in the set $\{1, \ldots, F\}$, where d is a channel dimension, and the notation Q_f, K_f, and V_f is the query, key, and value of the f-th frame, respectively. In VM, we modify Eq. 1 to following:

$$\text{VM}(Q_f, K_{1:F}, V_{1:F}) = \text{Softmax}\left(\frac{Q_f (K_{r(t')})^{\mathrm{T}}}{\sqrt{d}}\right) V_{r(t')} \quad (2)$$

for all f's, where $r : \{1, \ldots, T'\} \to \{1, \ldots, F\}$. Here, t' is the number of timesteps from the timestep at which we initiate the value mapping 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,

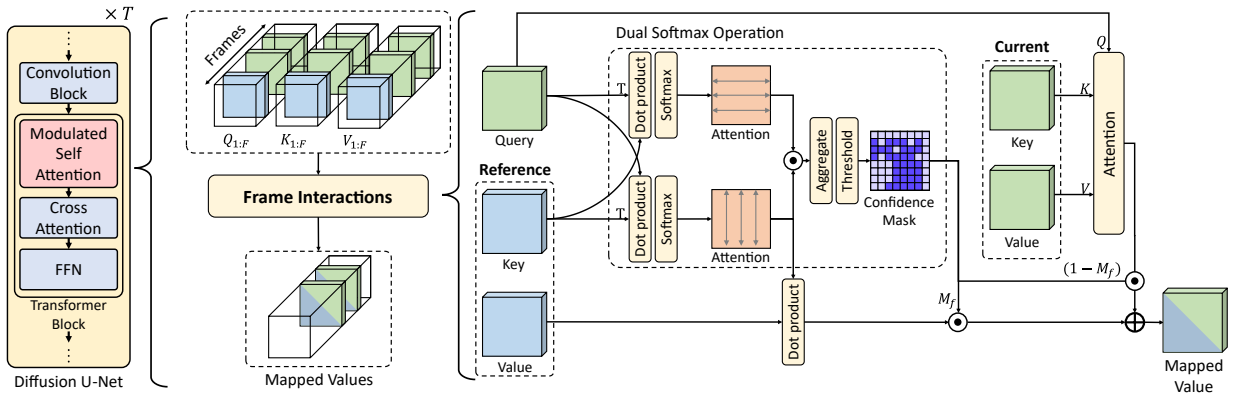

Figure 2. **The proposed frame interaction mechanism.** Within our augmented self-attention layers in a diffusion U-Net (Ronneberger et al., 2015), value mapping and dual-softmax are introduced to ensure temporal consistency and reduce unreliable matching between frames, respectively.

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. On a high level, RVM not only enables the complete, bidirectional flow of value mapping but also makes each frame equally contributable. Specifically, we define the reference frame as $r(t') = \text{Mod}(\lfloor t'/m \rfloor, F) + 1$, where m represents 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 (Khachatryan et al., 2023). Empirically, setting m to 4 stabilizes the outcome while maintaining the capability to incorporate various content specified in the frame-level prompts.

2.3. Reducing unreliable matching via dual softmax filtering

Although VM can diversify the contextual information within the generated video, on its own, it face difficulties when accounting for rapid movements or drastic changes. In Eq. 2, 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 (Teed & Deng, 2020) 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 (Mac Aodha et al., 2012; Cheng et al., 2021; Truong et al., 2021; Seo et al., 2022). 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 leverage dual softmax (Cheng et al., 2021) and apply a threshold to these values using a specified quantile. Starting from Eq. 2, the dual softmax (Cheng et al., 2021), denoted as C_{dual}, is defined as follows:

$$C_{\text{dual}} = \text{Softmax}(Q_t(K_{r(t')})^{\mathrm{T}}) \odot \text{Softmax}(K_{r(t')}(Q_t)^{\mathrm{T}}), \tag{3}$$

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

$$\begin{aligned} &\text{VM}'(Q_f, K_{1:F}, V_{1:F}) \\ &= (1 - M_f) \odot \text{Attention}(Q_f, K_f, V_f) \\ &+ M_f \odot \text{VM}(Q_f, K_{1:F}, V_{1:F}) \end{aligned} \tag{4}$$

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

3. Experiments

3.1. Zero-shot video generation results

Qualitative results. In Fig. 3, 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 with our attention mechanism, the prompts are animated into a video containing dynamic actions and time-varying content.

From the results, given the 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 with the other prompt, *"A rainbow forming after a rain shower,"* our method demonstrates the ability to synthesize videos with narrative consistency,

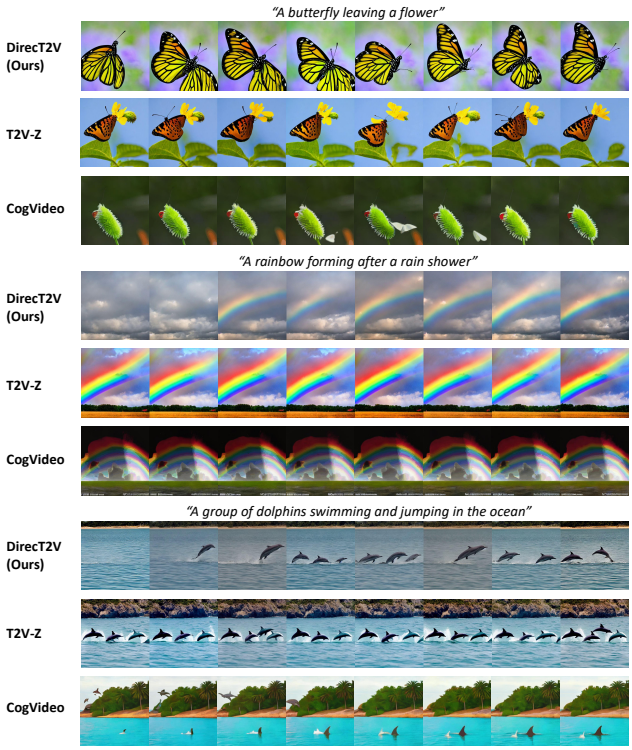

Figure 3. **Zero-shot video generation results.** Given an abstract user prompt, we compare DirecT2V with T2V-Z (Khachatryan et al., 2023) and CogVideo (Hong et al., 2022). Note that CogVideo is trained with video-text dataset, instead of zero-shot generation.

Figure 4. **Lifted frame rate.** By iteratively dividing frame-wise prompts, we can generate a video with an arbitrary frame rate.

whereas T2V-Z, with just the user prompt, shows the rainbow existing in every frame.

Notably, the generated video from CogVideo (Hong et al., 2022) 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 corroborate the effectiveness of our frame-wise prompting approach.

Quantitative results. Although our zero-shot lifting from image diffusion models does not have a ground-truth video reference dataset, several works use the CLIP score as a metric to evaluate faithfulness. Following this, we display in Fig. 5 the CLIP similarity score comparison for each frame. For the comparison, we used the proposed attention mechanism in Text2Video-Zero (T2V-Z) (Khachatryan et al., 2023), Tune-A-Video (TAV) (Wu et al., 2022b), our rotational value mapping, and rotational value mapping with dual softmax filtering. Note that for TAV, we only leveraged

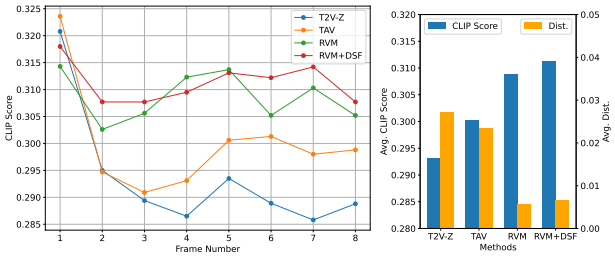

Figure 5. **Lifted frame rate.** By iteratively dividing frame-wise prompts, we can generate a video with an arbitrary frame rate.

the attention mechanism for zero-shot video generation. In terms of average similarity, our attention mechanism clearly outperforms the others. *Avg. Dist.* in this table refers to the average L1 distance between frame 1 and the other frames, and we can observe that both T2V-Z and TAV are only faithful to the first prompt.

3.2. 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. This approach effectively handles situations where the number of frames exceeds the batch size.

Originally, the frame-wise director extracts F frames at a specified frame rate R. We then provide the following prompt: *"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. 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. 4.

4. Conclusion

In this paper, we present 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, we demonstrate its capability to generate detailed and temporally consistent image descriptions, which can be effectively integrated into text-to-image models. We introduce 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.

References

An, J., Zhang, S., Yang, H., Gupta, S., Huang, J.-B., Luo, J., and Yin, X. Latent-shift: Latent diffusion with temporal shift for efficient text-to-video generation. *arXiv preprint arXiv:2304.08477*, 2023.

Austin, J., Odena, A., Nye, M., Bosma, M., Michalewski, H., Dohan, D., Jiang, E., Cai, C., Terry, M., Le, Q., et al. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021.

Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S. W., Fidler, S., and Kreis, K. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22563–22575, 2023.

Brooks, T., Holynski, A., and Efros, A. A. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18392–18402, 2023.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Chen, R., Chen, Y., Jiao, N., and Jia, K. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. *arXiv preprint arXiv:2303.13873*, 2023a.

Chen, W., Wu, J., Xie, P., Wu, H., Li, J., Xia, X., Xiao, X., and Lin, L. Control-a-video: Controllable text-to-video generation with diffusion models. *arXiv preprint arXiv:2305.13840*, 2023b.

Cheng, X., Lin, H., Wu, X., Yang, F., and Shen, D. Improving video-text retrieval by multi-stream corpus alignment and dual softmax loss. *arXiv preprint arXiv:2109.04290*, 2021.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Ding, M., Zheng, W., Hong, W., and Tang, J. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *arXiv preprint arXiv:2204.14217*, 2022.

Esser, P., Chiu, J., Atighehchian, P., Granskog, J., and Germanidis, A. Structure and content-guided video synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7346–7356, 2023.

Google. Palm 2 technical report, 2023.

Guo, Y., Yang, C., Rao, A., Wang, Y., Qiao, Y., Lin, D., and Dai, B. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.

Hao, Y., Chi, Z., Dong, L., and Wei, F. Optimizing prompts for text-to-image generation. *arXiv preprint arXiv:2212.09611*, 2022.

Ho, J. and Salimans, T. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *NeurIPS*, 33:6840–6851, 2020.

Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D. P., Poole, B., Norouzi, M., Fleet, D. J., et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022a.

Ho, J., Saharia, C., Chan, W., Fleet, D. J., Norouzi, M., and Salimans, T. Cascaded diffusion models for high fidelity image generation. *Journal of Machine Learning Research*, 23(47):1–33, 2022b.

Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., and Fleet, D. J. Video diffusion models. *arXiv preprint arXiv:2204.03458*, 2022c.

Hong, S., Ahn, D., and Kim, S. Debiasing scores and prompts of 2d diffusion for robust text-to-3d generation. *arXiv preprint arXiv:2303.15413*, 2023.

Hong, W., Ding, M., Zheng, W., Liu, X., and Tang, J. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022.

Huang, H., Feng, Y., Shi, C., Xu, L., Yu, J., and Yang, S. Free-bloom: Zero-shot text-to-video generator with llm director and ldm animator. *arXiv preprint arXiv:2309.14494*, 2023.

Karras, T., Aittala, M., Aila, T., and Laine, S. Elucidating the design space of diffusion-based generative models. *NeurIPS*, 2022.

Khachatryan, L., Movsisyan, A., Tadevosyan, V., Henschel, R., Wang, Z., Navasardyan, S., and Shi, H. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. *arXiv preprint arXiv:2303.13439*, 2023.

Koizumi, Y., Ohishi, Y., Niizumi, D., Takeuchi, D., and Yasuda, M. Audio captioning using pre-trained large-scale language model guided by audio-based similar caption retrieval. *arXiv preprint arXiv:2012.07331*, 2020.

Li, J., Li, D., Xiong, C., and Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pp. 12888–12900. PMLR, 2022a.

Li, S., Puig, X., Paxton, C., Du, Y., Wang, C., Fan, L., Chen, T., Huang, D.-A., Akyürek, E., Anandkumar, A., et al. Pre-trained language models for interactive decision-making. *Advances in Neural Information Processing Systems*, 35:31199–31212, 2022b.

Li, X., Chu, W., Wu, Y., Yuan, W., Liu, F., Zhang, Q., Li, F., Feng, H., Ding, E., and Wang, J. Videogen: A reference-guided latent diffusion approach for high definition text-to-video generation. *arXiv preprint arXiv:2309.00398*, 2023a.

Li, Y., Liu, H., Wu, Q., Mu, F., Yang, J., Gao, J., Li, C., and Lee, Y. J. Gligen: Open-set grounded text-to-image generation. *arXiv preprint arXiv:2301.07093*, 2023b.

Lin, C.-H., Gao, J., Tang, L., Takikawa, T., Zeng, X., Huang, X., Kreis, K., Fidler, S., Liu, M.-Y., and Lin, T.-Y. Magic3d: High-resolution text-to-3d content creation. *arXiv preprint arXiv:2211.10440*, 2022.

Liu, L., Ren, Y., Lin, Z., and Zhao, Z. Pseudo numerical methods for diffusion models on manifolds. *ICLR*, 2022.

Luo, Z., Chen, D., Zhang, Y., Huang, Y., Wang, L., Shen, Y., Zhao, D., Zhou, J., and Tan, T. Videofusion: Decomposed diffusion models for high-quality video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10209–10218, 2023.

Mac Aodha, O., Humayun, A., Pollefeys, M., and Brostow, G. J. Learning a confidence measure for optical flow. *IEEE transactions on pattern analysis and machine intelligence*, 35(5):1107–1120, 2012.

Metzer, G., Richardson, E., Patashnik, O., Giryes, R., and Cohen-Or, D. Latent-nerf for shape-guided generation of 3d shapes and textures. *arXiv preprint arXiv:2211.07600*, 2022.

Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., and Chen, M. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

OpenAI. Introducing chatgpt, 2022.

OpenAI. Gpt-4 technical report, 2023.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

Poole, B., Jain, A., Barron, J. T., and Mildenhall, B. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.

Qi, C., Cun, X., Zhang, Y., Lei, C., Wang, X., Shan, Y., and Chen, Q. Fatezero: Fusing attentions for zero-shot text-based video editing. *arXiv preprint arXiv:2303.09535*, 2023.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *CVPR*, pp. 10684–10695, 2022.

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S. K. S., Ayan, B. K., Mahdavi, S. S., Lopes, R. G., et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.

Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A. S., Yvon, F., Gallé, M., et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022.

Seo, H., Kim, H., Kim, G., and Chun, S. Y. Ditto-nerf: Diffusion-based iterative text to omni-directional 3d model. *arXiv preprint arXiv:2304.02827*, 2023a.

Seo, J., Lee, G., Cho, S., Lee, J., and Kim, S. Midms: Matching interleaved diffusion models for exemplar-based image translation. *arXiv preprint arXiv:2209.11047*, 2022.

Seo, J., Jang, W., Kwak, M.-S., Ko, J., Kim, H., Kim, J., Kim, J.-H., Lee, J., and Kim, S. Let 2d diffusion model know 3d-consistency for robust text-to-3d generation. *arXiv preprint arXiv:2303.07937*, 2023b.

Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models. In *ICLR*, 2021.

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., Radford, A., Amodei, D., and Christiano, P. F. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33: 3008–3021, 2020.

Teed, Z. and Deng, J. 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*, pp. 402–419. Springer, 2020.

Truong, P., Danelljan, M., Van Gool, L., and Timofte, R. Learning accurate dense correspondences and when to trust them. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5714–5724, 2021.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *NeurIPS*, 30, 2017.

Villegas, R., Babaeizadeh, M., Kindermans, P.-J., Moraldo, H., Zhang, H., Saffar, M. T., Castro, S., Kunze, J., and Erhan, D. Phenaki: Variable length video generation from open domain textual description. *arXiv preprint arXiv:2210.02399*, 2022.

Wang, H., Du, X., Li, J., Yeh, R. A., and Shakhnarovich, G. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. *arXiv preprint arXiv:2212.00774*, 2022.

Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., and Le, Q. V. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

Wu, C., Liang, J., Ji, L., Yang, F., Fang, Y., Jiang, D., and Duan, N. 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*, pp. 720–736. Springer, 2022a.

Wu, J. Z., Ge, Y., Wang, X., Lei, W., Gu, Y., Hsu, W., Shan, Y., Qie, X., and Shou, M. Z. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. *arXiv preprint arXiv:2212.11565*, 2022b.

Xu, J., Wang, X., Cheng, W., Cao, Y.-P., Shan, Y., Qie, X., and Gao, S. Dream3d: Zero-shot text-to-3d synthesis using 3d shape prior and text-to-image diffusion models. *arXiv preprint arXiv:2212.14704*, 2022.

Zhou, D., Wang, W., Yan, H., Lv, W., Zhu, Y., and Feng, J. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*, 2022a.

Zhou, Y., Muresanu, A. I., Han, Z., Paster, K., Pitis, S., Chan, H., and Ba, J. Large language models are human-level prompt engineers. *arXiv preprint arXiv:2211.01910*, 2022b.

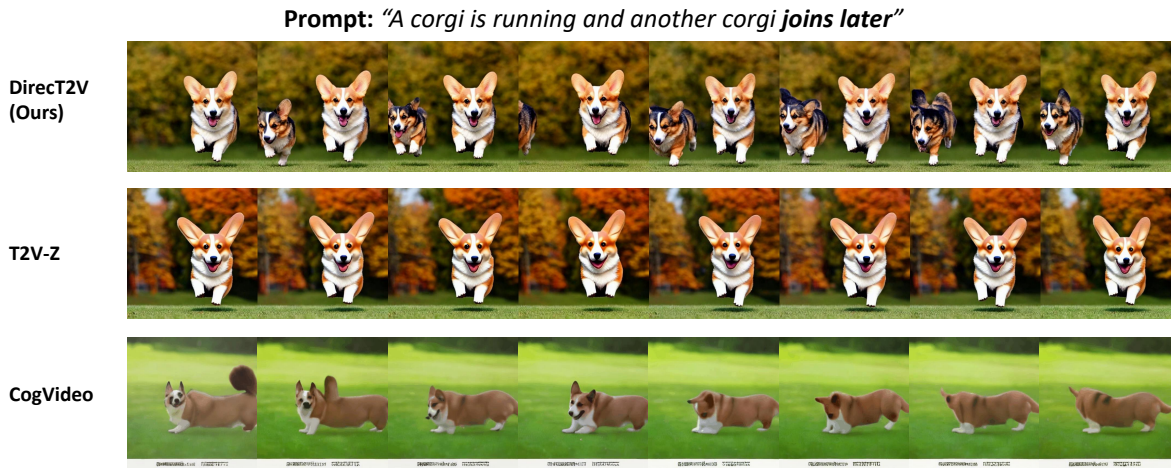

Figure 6. **Zero-shot video generation.** DirecT2V, using LLMs as frame-level directors, enables zero-shot narrative text-to-video generation, while current zero-shot (*e.g.*, Text2Video-Zero (Khachatryan et al., 2023)) or tuned (*e.g.*, CogVideo (Hong et al., 2022)) baselines do not contain high-level stories, *e.g.*, joining of the second corgi.

A. Background

Numerous works in text-to-image (T2I) field, which include GLIDE (Nichol et al., 2021), Dall-E 2 (Ramesh et al., 2022), latent diffusion models (LDM) (Rombach et al., 2022) and Imagen (Saharia et al., 2022), have been actively employing diffusion models for their high-fidelity generation. In this section, we first explain details of LDM (Rombach et al., 2022) whose methods are adopted in Stable Diffusion.

LDM (Rombach et al., 2022) is a diffusion model that performs the forward and reverse process within the latent space of an autoencoder denoted as $D(E(\cdot))$, where $E(\cdot)$ and $D(\cdot)$ 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 noise is progressively added to z_0 as follows:

$$q(z_t|z_{t-1}) = \mathcal{N}(z_t; \sqrt{1-\beta_t} z_{t-1}, \beta_t I), \tag{5}$$

where $t = 1, \ldots, T$, and $q(z_t|z_{t-1})$ denotes 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:

$$p_\theta(z_{t-1}|z_t) = \mathcal{N}(z_{t-1}; \mu_\theta(z_t, t), \Sigma_\theta(z_t, t)), \tag{6}$$

where $t = T, \ldots, 1$. This 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 (Rombach et al., 2022) instead predicts the reparametrization $\epsilon_\theta(z_t, t)$. To achieve text-conditioned image sampling, the text embedding of a user prompt ω is conditioned along the intermediate features via the cross-attention layers (Vaswani et al., 2017; Rombach et al., 2022), resulting in the conditional prediction term $\epsilon_\theta(z_t, t, \omega)$, and classifier-free guidance (Ho & Salimans, 2021) is leveraged for better alignment to user prompts.

B. Zero-shot video generation results with motion dynamics

Our framework can naturally be extended to provide motion dynamics (Khachatryan et al., 2023), 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. 11 and Fig. 12, and the video is available on our homepage.

C. Implementation details

In this work, we employ GPT-4 (OpenAI, 2023) as our instruction-tuned LLM and T2V-Z (Khachatryan et al., 2023), utilizing a single NVIDIA 3090 RTX GPU for efficient video sampling. For the generation process, we employ the PNDM

User Prompt "A nuclear explosion" Action Object Setting Camera

Frame 1 (00:00) A desert landscape extends to the horizon under a clear, azure sky. At the center of the frame, a tall, metallic tower stands isolated, with a distinctively shaped, ominous device at the top. The camera angle is wide, capturing the vast emptiness of the surroundings.
Frame 2 (00:01) The tower's top, housing the nuclear device, starts to glow with a bright, white light. The surrounding desert maintains its stark contrast, silent and untouched. The camera angle remains wide, focusing on the impending change.
Frame 3 (00:02) The blinding light from the device intensifies, casting long, dark shadows across the desert landscape. Sand particles begin to stir, reacting to the imminent force. The camera has slightly zoomed in on the tower, anticipating the blast.
...
Frame 8 (00:07) The mushroom cloud dominates the sky, the base of the explosion obscured by the continuing dust and debris. The shockwave is now dissipating, leaving behind a transformed landscape. The camera captures the final scene from a distance, showing the aftermath of the nuclear explosion.

User Prompt "A corgi is running and another corgi joins later" Action Object Setting Camera

Frame 1 (00:00) A bright, sunny park, trees lining the edges, grass freshly cut. In the center, a single corgi, with a red and white coat, is joyfully bounding forward, its tail in the air, ears flapping.
Frame 2 (00:01) Close-up on the corgi's paws, kicking up small tufts of grass as it continues its sprint. The park's greenery and distant laughter of children form the blurred background.
Frame 3 (00:02) From a low side-angle, the corgi leaps over a small puddle, water droplets scattering in the sunlight. Its tongue is out, eyes focused ahead, showing pure exhilaration.
...
Frame 8 (00:07) Aerial view of the park, the two corgis now running side by side, leaving a trail of stirred grass behind them. Their joyful chase continues under the sunny sky.

Figure 7. **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.

scheduler (Liu et al., 2022), which is a member of the deterministic diffusion samplers family (Karras et al., 2022; Song et al., 2021; Liu et al., 2022). We employ the same scheduling mechanism as T2V-Z (Khachatryan et al., 2023) and configure the scheduler parameters of the diffusion models with $T = 100$ and $T' = 96$. Furthermore, we adopt the classifier-free guidance (Ho & Salimans, 2021), 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 present examples of the full prompts used to produce the results shown in Fig. 7. These examples are depicted in Fig. 15. Additionally, in Fig. 16, we provide examples of the frame-wise prompts utilized in creating videos.

D. User study

We also conducted a user study on the faithfulness to the user prompt, realism, and narrative of the generated videos, comparing them with the state-of-the-art zero-shot text-to-video baseline (Khachatryan et al., 2023). The results are shown in Table 1, which demonstrate that ours is preferable to human evaluators.

Table 1. User study.

Method	Faithfulness	Realism	Narrative
T2V-Z	21.39%	19.47%	20.67%
Ours	**78.61%**	**80.53%**	**79.33%**

In each case, two videos were provided, showing our results compared to the baseline. The identity of the videos (such as which video is ours) was completely concealed, and the placement of the videos was also randomized. This questionnaire was distributed over three days to local communities and universities, and stakeholders of the study were strictly excluded. We aggregated 416 answers from 52 participants, each of whom was asked to answer three questions about eight assigned cases. The questions were as follows:

- The videos in the first and second rows are the results generated from different models.
  - Which video is better aligned with the text displayed at the top of the videos?
  - Which video appears more realistic and natural?
  - Which video better demonstrates the narrative given in the text?

E. Ablation study

Rotational value mapping. In Fig. 8, 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

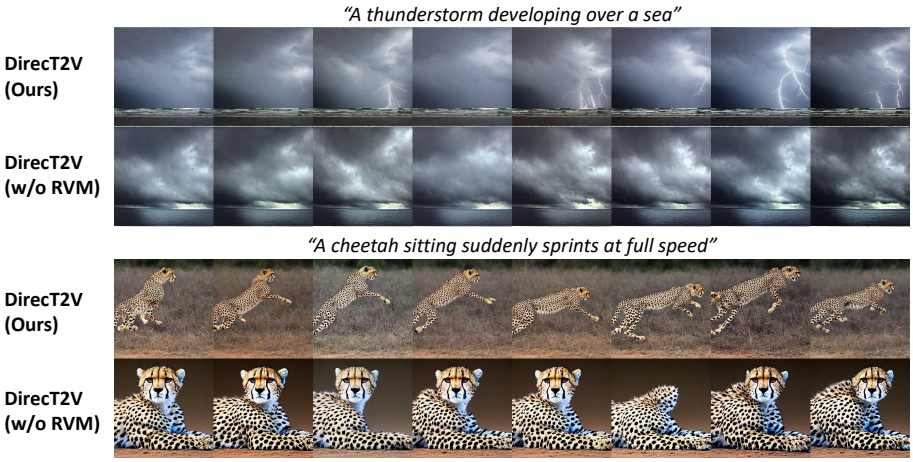

Figure 8. **Ablation study on RVM.** With frame-wise directing, RVM is essential for achieving narrative consistency, since its ablated counterpart (without RVM) does not reflect the frame-wise prompt well. Note that the same frame-wise prompts are used for both, while they show different results because RVM mixes various contexts across the frames to better reflect the upsampled prompts.

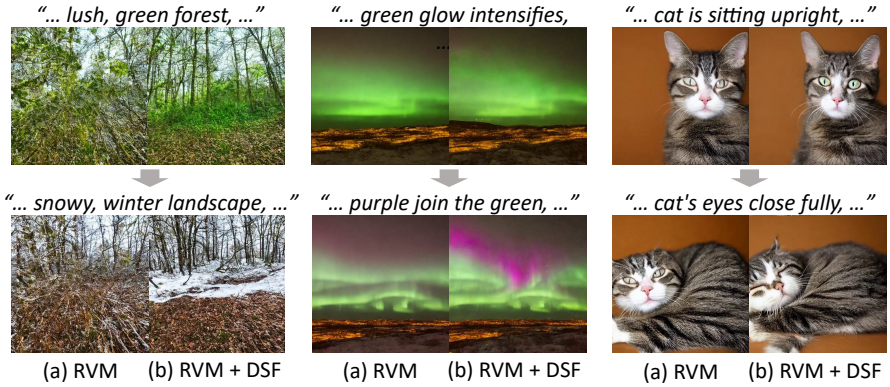

Figure 9. **Ablation study on DSF.** For each prompts, results on the left are only with RVM, while the right are with RVM and Dual-Softmax Filtering (DSF).

selection, *i.e.*, utilizing only the keys and values of the initial frame, as seen in (Khachatryan et al., 2023).

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. 9, 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.

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

Figure 10. **Examples of results from the different value mappings.**

from a simple prompt.

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.

Comparison of GPT-4, GPT-3.5, and Bard. We compare the frame-level directing abilities of GPT-4 (OpenAI, 2023), GPT-3.5 (OpenAI, 2022), and Bard (Google, 2023). In Fig. 14, we present the conversations with those LLMs. We find that they generally follow the instruction prompt well.

F. Related work

Incorporating large language models. Large language models (LLMs), such as GPT-3 (Brown et al., 2020), PaLM (Chowdhery et al., 2022), and BLOOM (Scao et al., 2022), have been shown to be effective in a wide range of tasks, *e.g.*, decision making (Li et al., 2022b), program synthesis (Austin et al., 2021), and prompt engineering (Zhou et al., 2022b). Notably, their zero-shot capabilities have demonstrated strong generalization power that almost resembles the linguistic ability of humans. By transferring such knowledge, numerous methods (Saharia et al., 2022; Li et al., 2022a; Koizumi et al., 2020; Brown et al., 2020) 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 (Stiennon et al., 2020; Christiano et al., 2017), enabled accurate manipulation of LLMs that aligns with human intent. Another line of works, including (Brooks et al., 2023; Hao et al., 2022), have proposed to combine pre-trained language models with diffusion-based generative models, aiming to generate prompts that produce more reliable results. However, the capability of LLMs to recognize the time variation of a video scene or to generate time-varying prompts for a single user prompt has rarely been explored.

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 (Wu et al., 2022a; Villegas et al., 2022; Ho et al., 2022a;b). NÜWA (Wu et al., 2022a) introduces a 3D transformer (Vaswani et al., 2017)-based encoder-decoder framework and aims to tackle various tasks, including text-to-video generation, while Phenaki (Villegas et al., 2022) presents a bidirectional masked transformer for a video creation from arbitrary-length text prompt sequences. Similarly, Imagen Video (Ho et al., 2022a) leverages diffusion models for cascading pipeline (Ho et al., 2022b) 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 (Hong et al., 2022) builds upon CogView2 (Ding et al., 2022), a text-to-image model, and employs a multi-frame-rate hierarchical training strategy, encouraging text and video alignment. Make-a-video (Singer et al., 2022) tackles T2V task more efficiently

by using the synthetic data for self-supervision. Another line of works (Ho et al., 2022c; Zhou et al., 2022a) that exploit LDM (Rombach et al., 2022) enables high-resolution video generation by introducing temporal tuning technique that efficiently fine-tunes the parameters. Taking a step forward, Text2Video-Zero (T2V-Z) (Khachatryan et al., 2023) 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. 6, the time-dependent dynamics are often disregarded and only limitedly expressed. In this work, we obtain time-varying prompts from a single user prompt using an instruction-tuned LLM (Ouyang et al., 2022; OpenAI, 2023; Google, 2023; Wei et al., 2021), and use them to synthesize videos that successfully capture both static and dynamic components.

G. Discussion

Concurrent work. In a short time since this paper, there have been a large number of works on text-to-video using image or video diffusion models (Huang et al., 2023; Blattmann et al., 2023; Esser et al., 2023; Guo et al., 2023; An et al., 2023; Luo et al., 2023; Chen et al., 2023b; Li et al., 2023a). Among these, there is a concurrent endeavor that also leverages LLMs for zero-shot text-to-video generation, Free-Bloom (Huang et al., 2023), which serves as supporting evidence for our motivation to use LLMs in this task. Free-Bloom adopts the previously known attention mechanism, which we have experimented with in our ablation study, with the key difference being its cessation of references to other frames at a timestep. Furthermore, it employs a different noise sampling technique. Overall, their work is mostly orthogonal to ours in terms of the attention mechanism and video generation.

Limitations. The performance of the proposed method, DirecT2V, may vary depending on the instruction-tuned large language models (LLMs) (OpenAI, 2023; 2022; Google, 2023) (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 (Saharia et al., 2022; Li et al., 2023b). A potential solution to this problem could involve the use of the encoder from an even larger language model (Saharia et al., 2022).

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.

Impact statement. 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 (Khachatryan et al., 2023; Singer et al., 2022; Hong et al., 2022; Singer et al., 2022; Ho et al., 2022a; Villegas et al., 2022; Wu et al., 2022a; Zhou et al., 2022a). This philosophy can be extended to other high-level visual tasks, such as zero-shot text-to-3D (Hong et al., 2023; Poole et al., 2022; Wang et al., 2022; Metzer et al., 2022; Lin et al., 2022; Chen et al., 2023a; Seo et al., 2023b; Xu et al., 2022; Seo et al., 2023a). 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.

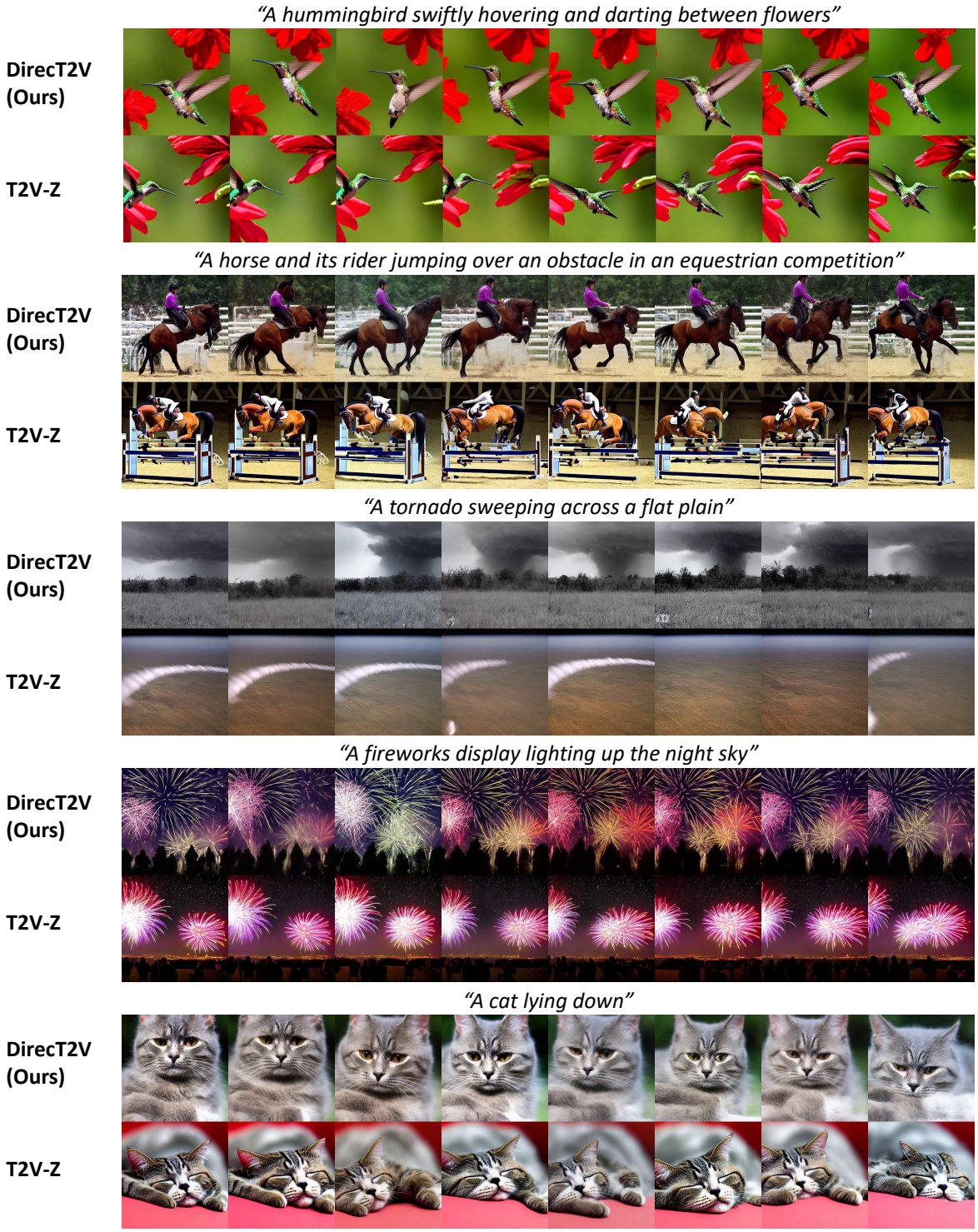

Figure 11. **Zero-shot video generation results with motion dynamics (Khachatryan et al., 2023).** We compare our method with T2V-Z (Khachatryan et al., 2023).

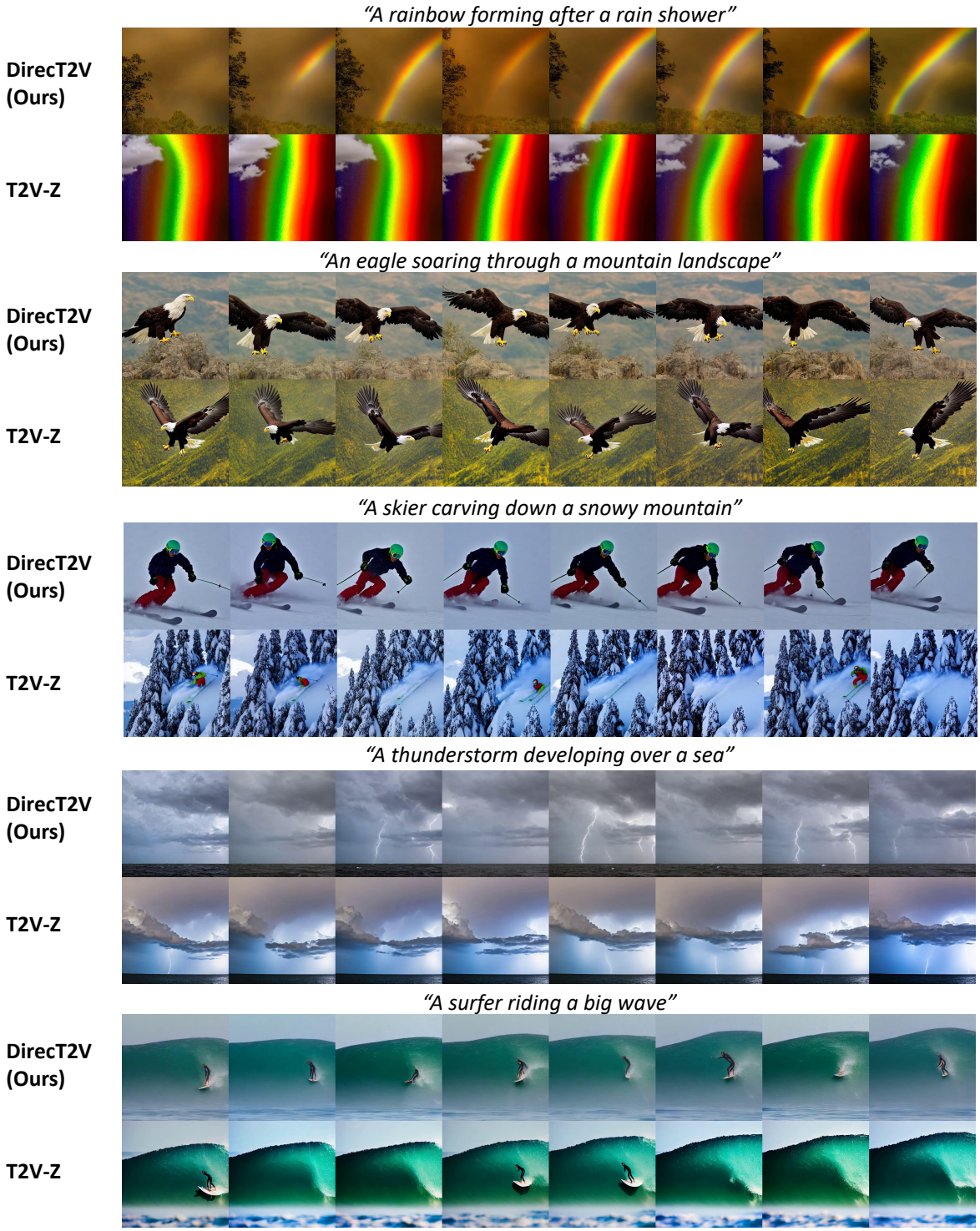

Figure 12. **Zero-shot video generation results with motion dynamics (Khachatryan et al., 2023).** We compare our method with T2V-Z (Khachatryan et al., 2023).

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.**

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 (OpenAI, 2023), GPT-3.5 (OpenAI, 2022), and Bard (Google, 2023).**

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 15. **The complete instruction for Fig. 7, 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 16. **Examples of frame-level prompts directed by LLMs.**