Free-Bloom: Zero-Shot Text-to-Video Generator with LLM Director and LDM Animator Supplemental Material

Anonymous Author(s)
Affiliation
Address
email

In this section, we provide additional discussions, details, and experiments to further support our contributions. The content is organized as

- Section A - Discussions
- Section B - Joint Noise Derivation
- Section C - Implementation Details
  - Section C.1 - Serial Prompting
  - Section C.2 - Test Set for Quantitative Results
  - Section C.3 - Details of User Study
  - Section C.4 - Code Used and License
- Section D - Additional Experiments
  - Section D.1 - Qualitative Results (comparisons and more visualizations)
  - Section D.2 - User Study Quantitative Comparisons
  - Section D.3 - Analysis on Joint Noise Sampling
  - Section D.4 - Extensions (long video story, personalization, making an image move)

A Discussions

Limitations and Future Work. We look forward to further research on this method. While our method offers the advantage of being training-free and not requiring extra training data, it highly depends on the large foundation models LLMs [2, 1, 7] and LDMs [8]. Consequently, it would inherit the limitations of those large pre-trained models. For example, LDMs often struggle with generating images containing detailed faces and limbs, specific text, multiple objects, interactions between objects, etc, therefore our method has the same weakness. Moreover, LDMs are often sensitive to seed selections of initial noises [10], so when the initial frame is of low quality, our method tends to result in relatively poor performance as well. Additionally, although our method demonstrates improved temporal consistency to other zero-shot methods, we found that it is still challenging to maintain high temporal coherency between frames in the zero-shot setting. However, leveraging video data proves to be an effective solution for acquiring temporal priors. Therefore, how to combine the strengths of zero-shot methods and trained methods is a promising direction for future research.

Societal Impacts. It should be acknowledged that there can be ethical impacts like other generative models. As we adopt ChatGPT [7] and Stable Diffusion v1.5 [8], our method may inherit the bias of those two models. Also, although the results of our method of direct text-to-video generation are still a step away from convincingly photo-realistic videos, the risk of abuse, for example, generating fake, harmful, or discriminating content, should be aware.

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

B Joint Noise Derivation

First, let us consider the distribution of **unified noise**. It is composed by initial noise $\mathbf{x}_T^* \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$ for each frame and can be represented as $\mathbf{x}_T^{1:f} = [\mathbf{x}_T^*, \cdots, \mathbf{x}_T^*]^T$. The noise of any two frames in $\mathbf{x}_T^{1:f}$ have the same values, and therefore their covariance is

$$\operatorname{cov}(\mathbf{x}_T^*, \mathbf{x}_T^*) = \mathbf{D}(\mathbf{x}_T^*) = \mathbf{I}_n. \tag{1}$$

Thus the unified noise follows the distribution as

$$p(\mathbf{x}_T^{1:f}) = \mathcal{N}(\mathbf{0}, \mathbf{J}_f \otimes \mathbf{I}_n), \tag{2}$$

where $\mathbf{J}_f$ represents the all-one matrix of size $f \times f$ and $\otimes$ denotes the Kronecker product. The specific form of $\mathbf{J}_f \otimes \mathbf{I}_n$ is

$$\mathbf{J}_f \otimes \mathbf{I}_n = \begin{bmatrix} \mathbf{I}_n & \cdots & \mathbf{I}_n \\ \vdots & \ddots & \vdots \\ \mathbf{I}_n & \cdots & \mathbf{I}_n \end{bmatrix} \begin{matrix} \overset{f \times \mathbf{I}_n}{\longleftrightarrow} \\ \\ \Big\updownarrow f \times \mathbf{I}_n \end{matrix} \tag{3}$$

Second, let us consider the distribution of **individual noise**, in which each frame is independently sampled. Therefore, the covariance between any two frames is $\mathbf{0}$, and the distribution still follows a standard normal distribution:

$$p(\boldsymbol{\delta}_T^{1:f}) = \mathcal{N}(\mathbf{0}, \mathbf{I}_{nf}) \tag{4}$$

According to Section 4.2, the mixed noise is defined as $\tilde{\mathbf{x}}_T^{1:f} := \cos(\frac{\pi}{2}\lambda)\mathbf{x}_T^{1:f} + \sin(\frac{\pi}{2}\lambda)\boldsymbol{\delta}_T^{1:f}$. Since $\mathbf{x}_T^{1:f}$ and $\boldsymbol{\delta}_T^{1:f}$ are independently sampled, the sum of the two still follows a normal distribution, with a mean of $\mathbf{0}$ and a variance of

$$\begin{aligned} \sin^2(\frac{\pi}{2}\lambda)\mathbf{I}_{nf} + \cos^2(\frac{\pi}{2}\lambda)\mathbf{J}_f \otimes \mathbf{I}_n) &= \sin^2(\frac{\pi}{2}\lambda)\begin{bmatrix} \mathbf{I}_n & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ \mathbf{0} & \cdots & 0 & \mathbf{I}_n \end{bmatrix} + \cos^2(\frac{\pi}{2}\lambda)\begin{bmatrix} \mathbf{I}_n & \cdots & \mathbf{I}_n \\ \vdots & \ddots & \vdots \\ \mathbf{I}_n & \cdots & \mathbf{I}_n \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{I}_n & \cos^2(\frac{\pi}{2}\lambda)\mathbf{I}_n & \cdots & \cos^2(\frac{\pi}{2}\lambda)\mathbf{I}_n \\ \cos^2(\frac{\pi}{2}\lambda)\mathbf{I}_n & \ddots & & \cos^2(\frac{\pi}{2}\lambda)\mathbf{I}_n \\ \vdots & & \ddots & \vdots \\ \cos^2(\frac{\pi}{2}\lambda)\mathbf{I}_n & \cos^2(\frac{\pi}{2}\lambda)\mathbf{I}_n & \cdots & \mathbf{I}_n \end{bmatrix} \\ &= \mathbf{I}_{nf} + \cos^2(\frac{\pi}{2}\lambda)((\mathbf{J}_f - \mathbf{I}_f) \otimes \mathbf{I}_n) \end{aligned} \tag{5}$$

Thus, variable $\tilde{\mathbf{x}}_T^{1:f}$ follows the following distribution.

$$\begin{aligned} p(\tilde{\mathbf{x}}_T^{1:f}) &= \mathcal{N}(\mathbf{0}, \sin^2(\frac{\pi}{2}\lambda)\mathbf{I}_{nf} + \cos^2(\frac{\pi}{2}\lambda)\mathbf{J}_f \otimes \mathbf{I}_n)) \\ &= \mathcal{N}(\mathbf{0}, \mathbf{I}_{nf} + \cos^2(\frac{\pi}{2}\lambda)((\mathbf{J}_f - \mathbf{I}_f) \otimes \mathbf{I}_n)) \end{aligned} \tag{6}$$

In this distribution, without given noises of other frames, for any frame noise $\tilde{\mathbf{x}}_T^i$, it still follows a standard normal distribution that $p(\tilde{\mathbf{x}}_T^i) = \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$.

C Implementation Details

C.1 Serial Prompting

We first prompt ChatGPT [7] with the following instruction:

• *I would like you to play the role of the describer of each frame of the video as a director of a movie. The content of each video should be concise and only clearly describe the subject. Each sentence in the video is independent. Every sentence needs to include the subject's appearance and*

actions, please describe the main actions of the object and the extent of the actions in as much detail as possible. The sentences of each picture are independent, and each sentence should describe what exists in the picture. Each frame is described in only one sentence. Suppose there is a video about "[INPUT PROMPT]" *and there are* "[F]" *frames in the video. Describe the content of each frame separately. Please be straightforward and do not use a narrative style.*

Then, we use the following prompt:

• *Now perform Coreference Resolution on the above sentence, replace reflexive pronouns with their original vocabulary, and eliminate the discourse cohesion. Keep the meaning the same. The sentence for each frame should be able to fully express all the visual information of the frame. Also, the linguistic structure of each sentence should be simple and similar.*

C.2 Test Set for Quantitative Results

We list some prompts from our test set in Table 1, in which some prompts are from the webpage of Text2Video-Zero [5] and some are designed by ourselves that incorporate more complex event content.

Table 1: Prompts for Test Set.

A cluster of flowers blooms	Astronaut riding a horse
Use pan to fire an egg	Iron man is surfing
Volcano eruption	The ice cube is melting
A dog is walking down the street	A panda is walking down the street
Light a match then the match goes out	The Santa flying through the sky
River freezes	Two supermen are fighting
Two men shake hands	A bear dancing on times square
Teddy bear jumps into water	An astronaut is waving his hands on the moon
The growth of a sapling	An egg hatch into a chick
A dancing mickey	Teddy bear is greeting

C.3 Details of User Study

We conduct a user study to understand how humans would evaluate the current text-to-video methods. The survey contains a total of 20 prompts with each prompt having 4 videos output from VideoFusion [6], LVDM [3], Text2Video-Zero [5], and Ours. For each prompt, we ask raters to answer the following four questions:

• How would you rate the temporal coherence and smoothness of the videos? Please assign a score for their **continuity**. (*Temporal Coherence*)

• How would you rate the quality and fidelity of the individual frames in the videos? Please assign a score for the **visual quality**. (*Fidelity*)

• How well does the video depict the content described in the text? Please assign a score for its **content** representation. (*Semantic Coherence*)

• Based on your **overall perception**, please rank the videos. (*Rank*)

Figure 9 presents an example interface of our survey. We received valid responses from a total of 80 individuals from both industry and academia.

C.4 Code Used and License

All used codes and their licenses are listed in Table 2.

Table 2: The used codes and license.

URL	Citation	License
https://github.com/showlab/Tune-A-Video	[14]	Apache License 2.0
https://github.com/google/prompt-to-prompt	[4]	Apache License 2.0
https://github.com/huggingface/diffusers	[13]	Apache License 2.0
https://github.com/Picsart-AI-Research/Text2Video-Zero	[5]	CreativeML Open RAIL-M
https://github.com/VideoCrafter/VideoCrafter	[3]	(Hugging Face Space) MIT
https://github.com/modelscope/modelscope/	[6]	Apache License 2.0

D Additional Experiments

D.1 Qualitative Results

We showcase more visualization of the generated video in this section. In Figure 1 and Figure 2, we present the full comparison with Text2Video-Zero [5], VideoFusion [6], and LVDM [3]. In Figure 3, we randomly generate multiple videos with respect to the same prompts. In Figure 4, we demonstrate more results of our interpolation empowerment module.

D.2 User Study Quantitative Comparisons

Table 3: **User Study Comparisons.**

		User Study			
Method	Training-Free	Rank	Fidelity	Temporal	Semantic
Ours vs. LVDM [3]		55.00%	85.28%	53.33%	82.27%
Ours vs. VideoFusion [6]		63.06%	82.50%	44.44%	78.33%
Ours vs. T2V-Zero [5]	✓	73.61%	87.78%	80.00%	85.28%

For the user study part in the quantitative results, we also present the comparison-based results here in Table 3. Specifically, for the ranking column, the number denotes the percentage of participants who prefer our method and rank us before another. For the dimensions of fidelity, temporal coherence, and semantic coherence, the numbers indicate the percentage of participants who believe that our generated videos are better to that of another method in that dimension.

D.3 Analysis on Joint Noise Sampling

We additionally analyze the effect of our noise sampling method. In part A of Figure 5, we sample the noise the same as equation $\tilde{\mathbf{x}}_T^{1:f} \coloneqq \cos(\frac{\pi}{2}\lambda)\mathbf{x}_T^{1:f} + \sin(\frac{\pi}{2}\lambda)\boldsymbol{\delta}_T^{1:f}$ using $\sin\cos$ weighting. While in part B, we modify the weight of the unified noise and the individual noise as

$$\tilde{\mathbf{x}}_T^{1:f} \coloneqq (1-\lambda)\mathbf{x}_T^{1:f} + \lambda\boldsymbol{\delta}_T^{1:f} \tag{7}$$

However, this would disrupt the single-frame noise from following normal Gaussian Distribution. As we can observe, in this way, LDM fails to generate reasonable images.

D.4 Extensions

In this section, we demonstrate the extensibility of our approach with more examples. Inspired by Phenaki [12] which can generate story-based conditional videos based on a sequence of prompts, we also apply our method to the same task, which is presented in Figure 6. In Figure 7, we showcase some results of combining DreamBooth [9] to include personalized concepts in the generated videos. In Figure 8, we showcase the results of generating videos based on the given first frame by leveraging DDIM inversion [11].

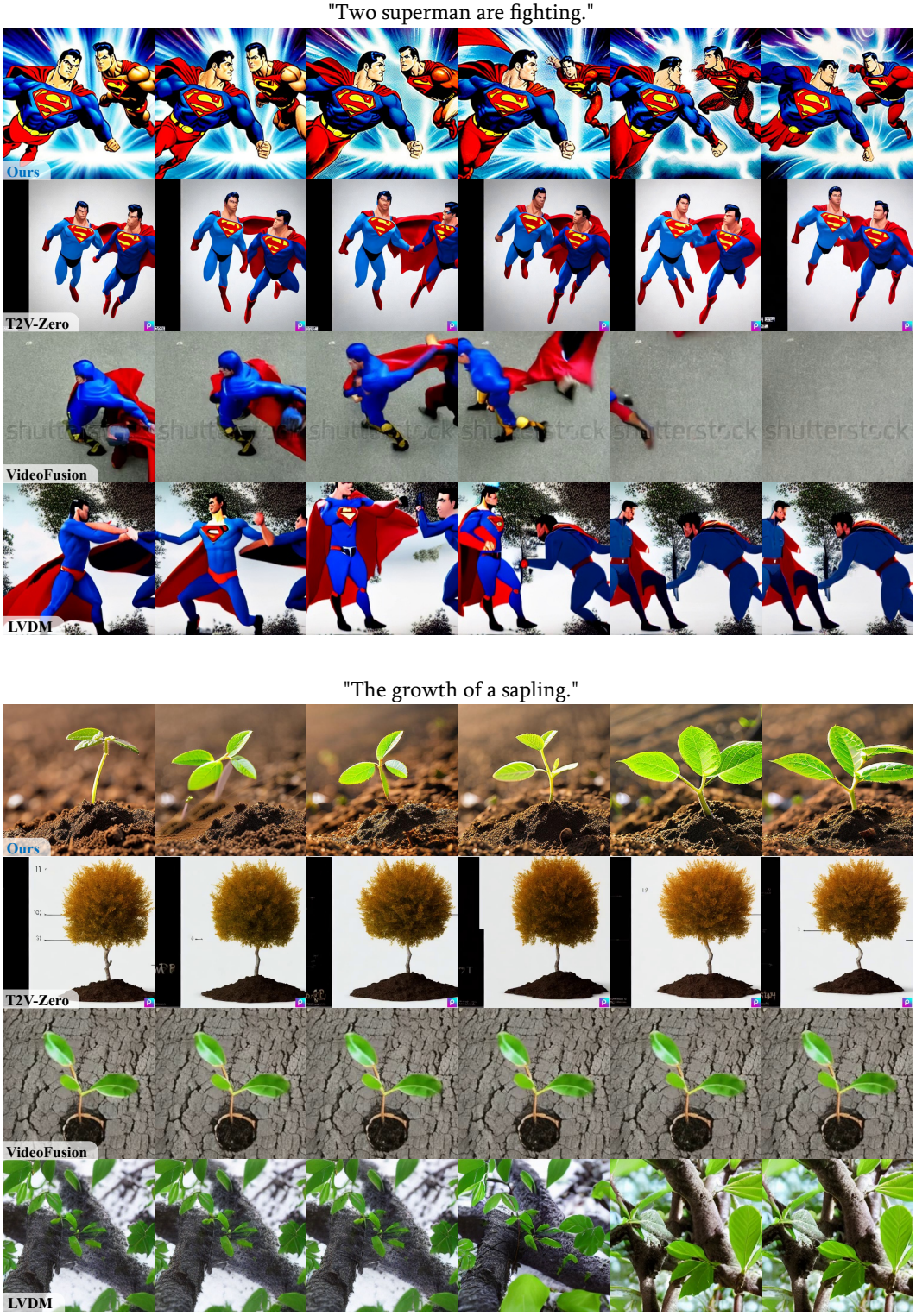

Figure 1: **Additional Qualitative Comparisons.** In the case of *"Two supermen are fighting"*, the LLM vividly decomposes the process of fighting into frames, with the fifth frame depicting *"colliding in a dazzling display of sparks and force"*, which is captured in our result. In the case of *"the growth of a sapling"*, our result clearly presents the gradual sprouting of a small sapling.

"Light a match then the match goes out."

"A teddy bear is greeting."

Figure 2: **Additional Qualitative Comparisons.** In the case of *"light a match then the match goes out"*, our method successfully depicts the entire process of a match from lightning, burning to extinguishing. In the case of *"a teddy bear is greeting"*, we exploit the world knowledge of LLM [7] to translate a greeting into a series of specific actions such as waving and smiling.

"A flower gradully blooms."

"Volcano eruption."

"A dancing mickey."

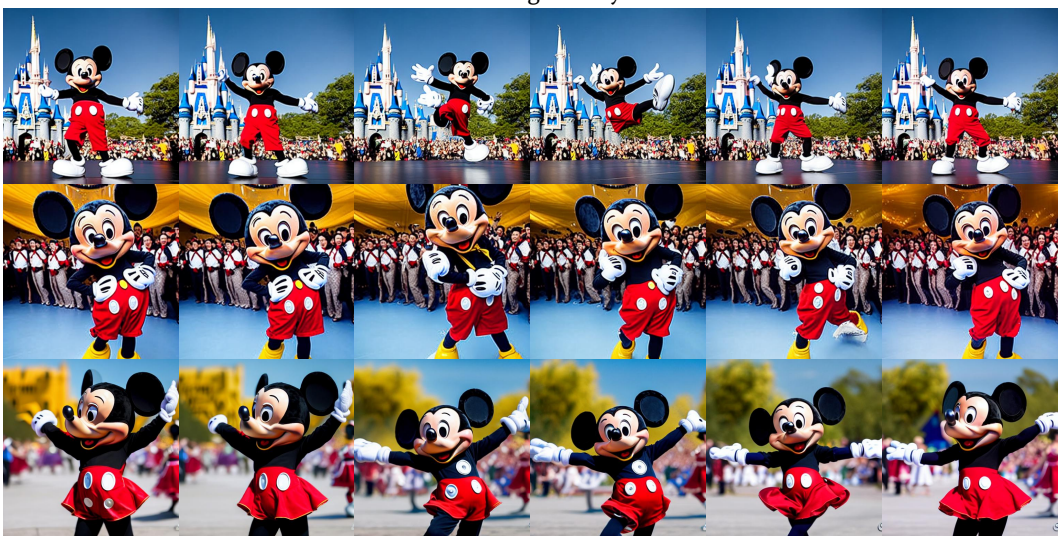

Figure 3: **Additional Qualitative Results.** Multiple results based on the same prompts are shown.

Figure 4: **Qualitative Results from Interpolation Module.** We interpolate 4 frames between each pair of original neighboring frames. Our interpolation module enables smooth transitions between two key states.

(A) $\cos(\frac{\pi}{2}\lambda)\mathbf{x}_T^{1:f} + \sin(\frac{\pi}{2}\lambda)\boldsymbol{\delta}_T^{1:f}$

(B) $(1-\lambda)\mathbf{x}_T^{1:f} + \lambda\boldsymbol{\delta}_T^{1:f}$

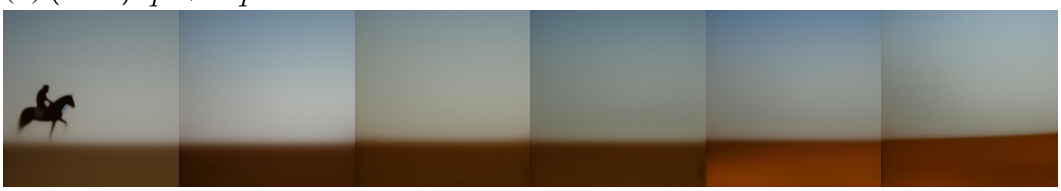

Figure 5: **Analysis on Joint Noise Sampling.** Without our proposed sampling, the initial noise at each single frame would not follow normal Gaussian distribution, resulting in corrupting frames.

1st prompt: "A flower is blooming"

2nd prompt: "The flower in the rain"

3rd prompt: "The flower gradually freezes"

Figure 6: **Extension - Long Video Story.** Our method can generate the long video story based on a sequence of prompts.

"<*ccorgi dog*> is sitting down."

"A princess is waving her hands, <*modern disney style*> ."

"<*sks mr potato head*> is dancing."

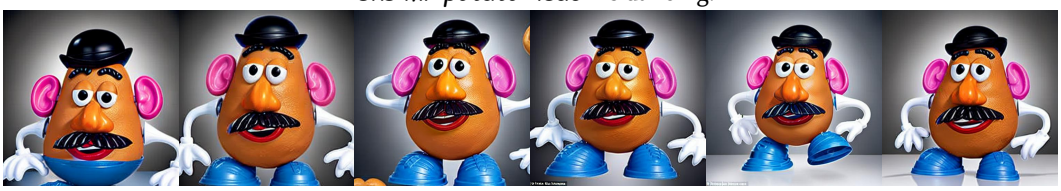

Figure 7: **Extension - Personalization.** Our method can generate videos with user-specific concepts. The tokens of "ccorgi dog", "modern disney style", and "sks mr potato head" are from their respective personalized models ccorgi-dog, Mo di Diffusion, and Mr Potato Head.

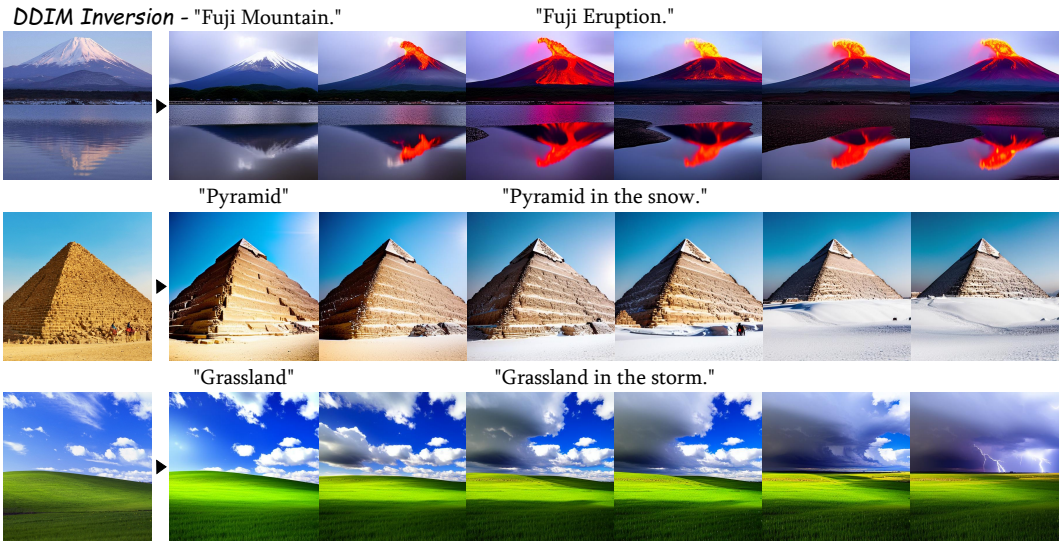

Figure 8: **Extension - Making an Image Move.** Our method can generate videos based on the first frame and its corresponding prompt by combining our method with DDIM inversion [11].

User Study on Text-to-Video Generation

Overview:
Thank you very much for helping us with our research! In this survey, we will present text descriptions along with corresponding videos, which are generated by four state-of-the-art **text-to-video AI models**. Each question will include a text prompt and the corresponding generated videos. Your task is to rank its performance across different dimensions. Additionally, please provide a final comprehensive ranking based on your preferences.

This survey consists of 5 questions and will approximately take 10 min of your time.

Sign in to Google to save your progress. Learn more

* Indicates required question

Below are four videos generated from "A cluster of flowers blooms"

A B C D

How would you rate the temporal coherence and smoothness of the videos? Please assign a score for their **continuity**. *

	1	2	3	4	5
A	○	○	○	○	○
B	○	○	○	○	○
C	○	○	○	○	○
D	○	○	○	○	○

Figure 9: Interface of surveys from the user study.

References

[1] 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.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

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

[4] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

[5] 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.

[6] Zhengxiong Luo, Dayou Chen, Yingya Zhang, Yan Huang, Liang Wang, Yujun Shen, Deli Zhao, Jingren Zhou, and Tieniu Tan. Videofusion: Decomposed diffusion models for high-quality video generation. *arXiv e-prints*, pages arXiv–2303, 2023.

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

[8] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[9] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arXiv preprint arXiv:2208.12242*, 2022.

[10] Dvir Samuel, Rami Ben-Ari, Simon Raviv, Nir Darshan, and Gal Chechik. It is all about where you start: Text-to-image generation with seed selection. *arXiv preprint arXiv:2304.14530*, 2023.

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

[12] 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.

[13] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. `https://github.com/huggingface/diffusers`, 2022.

[14] 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.