Prompt to Prompt to Video

Itamar Tsayag and Avraham Raviv
Bar-Ilan University
Israel
itamar.tsay@gmail.com , avrahamsapir1@gmail.com

Abstract

Building upon the recent success of the "Prompt-to-Prompt" framework for text-driven image editing, we introduce an extension, "Prompt-to-Video," which aims to generate dynamic video sequences from textual prompts using large-scale text-driven synthesis diffusion models. By leveraging the core ideas of the "Prompt-to-Prompt" framework, we detect description words in the given text and iteratively modify their temperatures to generate a series of semantically coherent images. We then apply interpolation techniques to create smooth transitions between these images, ultimately forming a cohesive video sequence. Our "Prompt-to-Video" approach expands the capabilities of text-based editing and synthesis, offering a seamless way to generate contextually relevant and visually appealing video sequences from textual prompts.

1. Introduction

In recent years, large-scale text-driven synthesis diffusion models have demonstrated impressive capabilities in generating diverse and high-quality images based on textual prompts. Building on this success, we propose an extension called "Prompt-to-Video," which aims to generate a dynamic video sequence from textual descriptions. Our method leverages a two-step process: detecting description words in the given text and then iteratively modifying their temperatures to generate a series of semantically coherent images. We then employ interpolation techniques to create smooth transitions between the generated images, ultimately forming a cohesive video sequence.

In our approach, we first analyze the text prompt to identify salient description words, which serve as the foundation for the synthesis process. We then generate multiple images by adjusting the temperatures of these words, resulting in a range of visual variations for each keyframe. To assemble the final video, we apply advanced interpolation techniques to create seamless transitions between the generated images, ensuring that the video maintains both temporal and visual coherence.

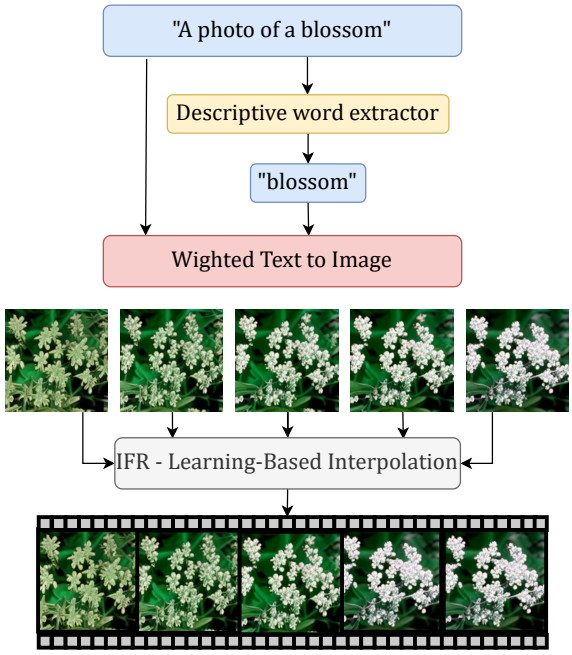

Figure 1: **An overview of our idea.**

Our "Prompt-to-Video" framework, which illustrated in Figure 1, opens up new possibilities for video generation and editing using text-based controls, enabling users to create contextually relevant and visually appealing video sequences without the need for manual editing or spatial masking. We showcase the effectiveness of our approach with a diverse set of text prompts and video synthesis models, demonstrating our framework's ability to produce high-quality video sequences that faithfully adhere to the given textual descriptions. Our code, full report and all results (more than 100 videos) are stored in

2. Related Work

Text-based Video Generation

Text-to-image (T2I) generation has experienced significant progress with methods like [19] extending unconditional Generative Adversarial Networks (GANs) to T2I generation. Later GAN variants have focused on progressive generation, better text-image alignment, and sequence-to-sequence translation using Variational Auto-Encoders (VQ-VAEs) and Transformers. Some notable examples include DALL-E [17], Make-A-Scene [3], and Parti [25]. Diffusion models like GLIDE [13], DALL-E2 [18], VQ-diffusion [4], and stable diffusion [20] have also been employed for T2I generation.

In contrast to the progress in T2I generation, text-to-video (T2V) generation has lagged behind, largely due to the lack of large-scale datasets with high-quality text-video pairs and the complexity of modeling higher-dimensional video data. Early works like Sync-DRAW [12], [14], and [**?**] focused on video generation in simple domains or specific human actions. More recent approaches such as GODIVA [23], NUWA [24], CogVideo [7], and Video Diffusion Models (VDM) [6] support more realistic scenes and leverage both image and video data.

The Make-A-Video approach stands out from previous works in several ways. First, it breaks the dependency on text-video pairs for T2V generation, making it more accessible and versatile compared to prior work that has been restricted to narrow domains or required large-scale paired text-video data. Second, it fine-tunes the T2I model for video generation, allowing for more effective adaptation of model weights compared to freezing them, as in CogVideo [7]. Third, motivated by prior work on efficient architectures for video and 3D vision tasks, the use of pseudo-3D convolution [15] and temporal attention layers allows for better temporal information fusion compared to VDM [6].

Detection of Descriptive Words

The detection of descriptive words is an important aspect of text-based image and video generation. Algorithms like the one used in [8] rely on natural language processing techniques, such as part-of-speech tagging and dependency parsing, to identify adjectives, adverbs, and other descriptive terms.

Deep learning models, such as BERT [2] and Transformer [21], have been employed to better understand the meaning of descriptive words in context. By incorporating detected descriptive words into the conditioning information for generative models, researchers have generated more accurate and contextually relevant visual content [9, 22]. This line of research has the potential to improve text-to-image and text-to-video generation methods by providing more fine-grained control over the generated content.

Interpolation

Interpolation techniques have been widely studied and employed in the context of image and video generation to create smooth transitions between keyframes or latent representations [11, 26]. These methods enable the generation of intermediate content by blending features from two or more input images or videos, producing visually coherent and smooth outputs that maintain the desired properties of the original inputs [1, 16].

3. Method and Results

The architecture consists of a word selection, a T2I model with the possibility of amplifying or attenuating the semantic effect of the selected word, and finally an interpolation method between the images.

Given a prompt, for example ”A fluffy bunny doll”, the adjectives ”fluffy” and ”bunny” are detected.

The T2I section of the architecture is based on the Prompt-to-Prompt paper [5]. Prompt-to-Prompt is an editing technique in T2I diffusion based models that preserve most of the original image. Text-to-image diffusion models consists of predicting the noise ϵ from a noisy image z_t and text embedding $\psi(P)$ along t diffusion steps. Most importantly, the interaction between the text modality and the pixel modality are in the Cross-Attention layers, where the embeddings of the visual and textual features are fused and spatial attention maps for each textual token are produced.

Amplifying or attenuating the semantic effect of a word in the generated image is done multiplying the cross-attention layer corresponding to the adjective with a weight factor. Higher weight value will produce images with a significant increase of the semantic representation of the adjective - i.e. more snow, more blossom and so on, and vice-versa.

We refer to the array of images produces as keyframes that are to be interpolated together for a high frame-per-second video. We experimented with several methods for interpolation, such as weighted image blending, and the latest IFRNet [10]. Results for four different propmts are shown in Figure 2, and more than 130 videos are available under videos folder in project page.

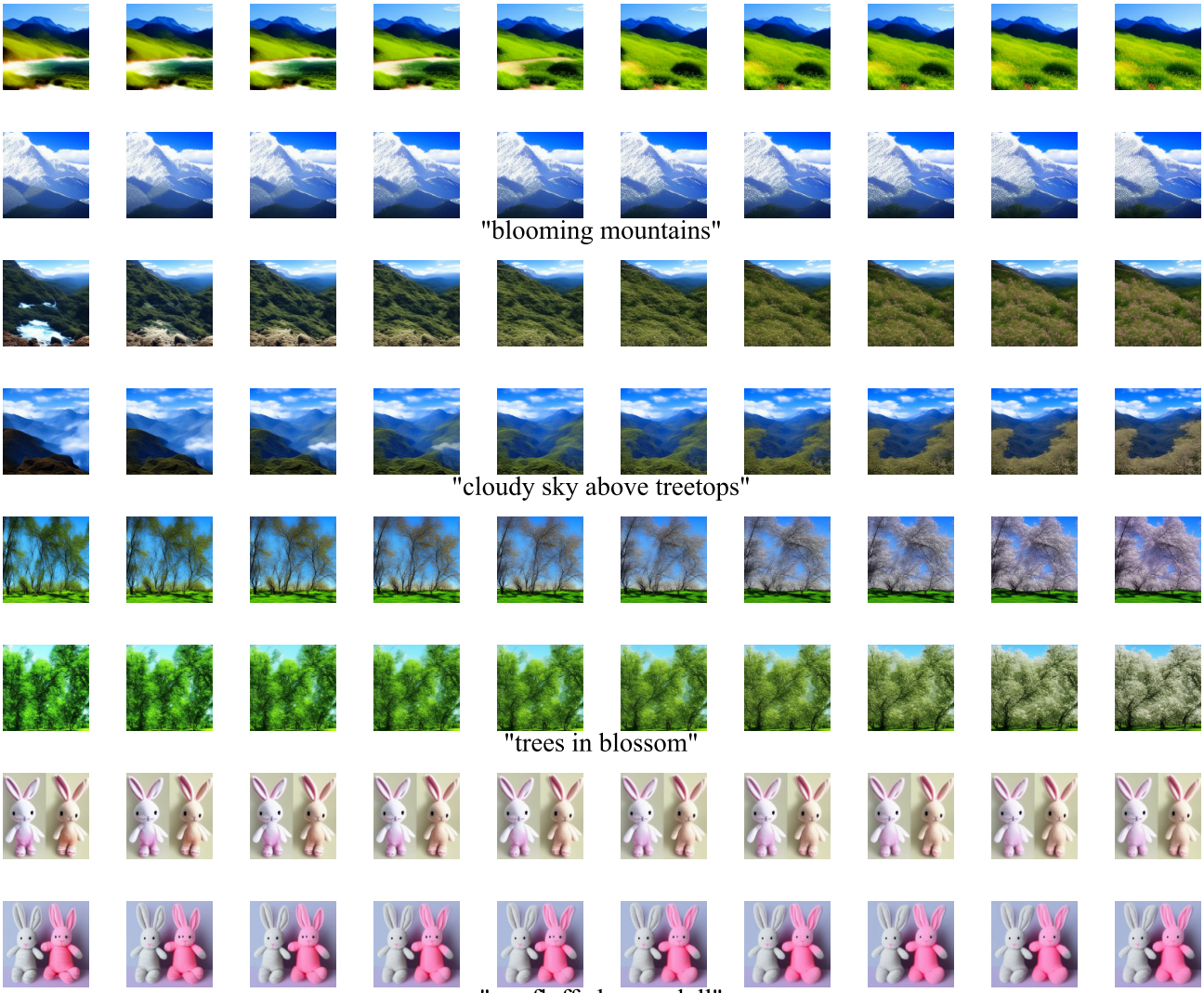

Figure 2. Each line is a different video.
The videos are 10 seconds long and we sampled 10 frames.

4. Conclusion and Future Work

This paper talks about a new method called "Prompt-to-Video" that can make videos using text instructions. The idea comes from another method called "Prompt-to-Prompt," which was used to edit pictures using text. First, the method finds important words in the text and then adjusts the weights of these embedding words to create a group of related images. After that, it uses special techniques to connect the images smoothly and make a full video. This new method makes it easy to create interesting and good-looking videos just by using text, without any need for manual editing or training.

For future work, we plan to enhance the "Prompt-to-Video" framework by allowing it to accept an image alongside the text prompt. This way, the generated video will be based on the provided image, giving users more control over the visual content and context of the resulting video sequence. By combining the advantages of image-based and text-based inputs, we aim to offer a more versatile and customizable video generation experience.

References

[1] M. Abadi, T. R. Shaham, T. Dekel, M. Irani, and W. T. Freeman. Neural network interpolation for continuous imagery effect transition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2417–2426, 2016.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2018.

[3] Eyal Gafni, Chuang Gan, Yelong Shen, and Marc'Aurelio Ranzato. Make-a-scene: Text-to-image generation with semantic layouts. *arXiv preprint arXiv:2201.03587*, 2022.

[4] Shaobo Gu, Yanghua Zhu, Deva Ramanan, and Anirudh Goyal. Vq-diffusion: Generating images with vector quantized diffusion models. *arXiv preprint arXiv:2202.08855*, 2022.

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

[6] Jonathan Ho, Aditya Ramesh, Mark Chen, Mikhail Pavlov, Pieter Abbeel, Oriol Vinyals, and Prafulla Dhariwal. Video diffusion models. *arXiv preprint arXiv:2202.07014*, 2022.

[7] Yiwen Hong, Chia-Wen Chen, Yunpeng Chen, and Jian Sun. Cogvideo: Text-to-video generation with frozen t2i priors. *arXiv preprint arXiv:2202.05649*, 2022.

[8] Matthew Honnibal and Ines Montani. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, 2017.

[9] Justin Johnson, Agrim Gupta, and Fei-Fei Li. Image generation from scene graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1219–1228, 2018.

[10] Lingtong Kong, Boyuan Jiang, Donghao Luo, Wenqing Chu, Xiaoming Huang, Ying Tai, Chengjie Wang, and Jie Yang. Ifrnet: Intermediate feature refine network for efficient frame interpolation, 2022.

[11] M. Leras, D. Ulyanov, and A. Vedaldi. Interpolation between deep image representations. *arXiv preprint arXiv:1804.08475*, 2018.

[12] Anurag Mittal, Sanja Fidler, Shikun Huang, and Raquel Urtasun. Sync-draw: Automatic video generation using deep recurrent attentive architectures. *arXiv preprint arXiv:1611.10314*, 2017.

[13] Alex Nichol, Prafulla Weiss, Pieter Abbeel, and Prafulla Dhariwal. Glide: Gumbel-softmax for image diffusion model guidance. *arXiv preprint arXiv:2111.15664*, 2021.

[14] Yingwei Pan, Ting Mei, Ting Yao, Houqiang Li, and Yong Rui. To create what you tell: Generating videos from captions. *Proceedings of the 25th ACM international conference on Multimedia*, pages 1789–1798, 2017.

[15] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4489–4497, 2017.

[16] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *Proceedings of the International Conference on Learning Representations*, 2015.

[17] Aditya Ramesh, Mikhail Pavlov, Gabriel Goyal, Scott Bileschi, Boris Klein, and Oriol Vinyals. Zero-shot text-to-image generation. *arXiv preprint arXiv:2102.12092*, 2021.

[18] Aditya Ramesh, Mikhail Pavlov, Alec Radford, Mark Chen, Oriol Vinyals, and Ilya Sutskever. Dall-e 2: Text-to-image generation with a latent space prior. *arXiv preprint arXiv:2202.03178*, 2022.

[19] Scott Reed, Zeynep Akata, Honglak Lee, and Bernt Schiele. Generative adversarial text-to-image synthesis. *arXiv preprint arXiv:1605.05396*, 2016.

[20] Roland Rombach, Eyal Gafni, Marc'Aurelio Ranzato, and Arthur Szlam. Stable diffusion. *arXiv preprint arXiv:2202.11128*, 2022.

[21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[22] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. In *Advances in Neural Information Processing Systems*, pages 1152–1164, 2018.

[23] Ruilong Wu, Boyuan Li, Xinyu Li, Yupei Li, Jun-Yan Zhu, and Cewu Lu. Godiva: Generating open-domain images and videos with attention. *arXiv preprint arXiv:2109.04027*, 2021.

[24] Ruilong Wu, Boyuan Li, Tianyang Xue, Yizhou Wang, and Cewu Lu. Nuwa: A unified representation for diverse text-to-image generation tasks. *arXiv preprint arXiv:2110.06802*, 2021.

[25] Lantao Yu, Yixuan Zhang, Hanting Cao, Hao Xie, Ting Chen, Yizhe Chen, Changyou Chen, and Yong Yu. Parti: A partitioned tokenizer for controllable and diverse text-to-image generation. *arXiv preprint arXiv:2201.05637*, 2022.

[26] Y. Zhang, Z. Murez, and P. Perona. Interpolation-aware semantic texture manipulation using a few annotated examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5971–5980, 2019.