Towards Real-Time Text2Video via CLIP-Guided, Pixel-Level Optimization

Peter Schaldenbrand* **Zhixuan Liu*** **Jean Oh***

Abstract

We introduce an approach to generating videos based on a series of given language descriptions. Frames of the video are generated sequentially and optimized by guidance from the CLIP image-text encoder; iterating through language descriptions, weighting the current description higher than others. As opposed to optimizing through an image generator model itself, which tends to be computationally heavy, the proposed approach computes the CLIP loss directly at the pixel level, achieving general content at a speed suitable for near real-time systems. The approach can generate videos in up to 720p resolution, variable frame-rates, and arbitrary aspect ratios at a rate of 1-2 frames per second. Please visit our website to view videos and access our open-source code: `https://pschaldenbrand.github.io/text2video/`.

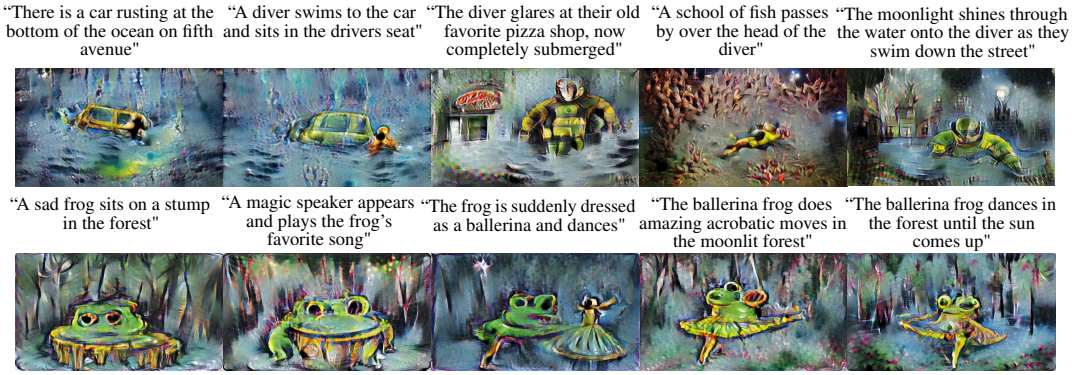

Figure 1: Frames sampled from videos generated by our approach produced with a series of language descriptions expressing narratives.

Introduction Animation is a compelling medium allowing unlimited degrees of visual expression while requiring its creator both the artistic skills and a tremendous amount of time to produce an artifact. Many people do not have the time and resources to learn animation software or create by hand; however, most can describe the visual elements of a story with words. To bridge this gap, we introduce a method for generating animated videos using natural language input.

Existing CLIP-Guided Text2Video and Text2Image approaches utilize pretrained image generator models, such as Diffusion. Using a pretrained image generator model ensures that the output will appear in distribution to the generator's training data. While this can produce highly-realistic imagery, it constrains the variety of producible content. Optimizing through large generator models is also time consuming. A single frame of video generated with Disco Diffusion [1] takes on the order of 5 minutes, 17 seconds for animation adaptations of Stable Diffusion [7], and 1 minute for CogVideo [4] while our model-free approach generates 1-2 frames per second.

Approach To achieve real-time text2video generation, we propose a two-step approach; (1) quickly and noisily generating semantic content, then (2) refining image textures in a post-processing step.

*The Robotics Institute, Carnegie Mellon University. {pschalde,zhixuan2,hyaejino}@andrew.cmu.edu

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

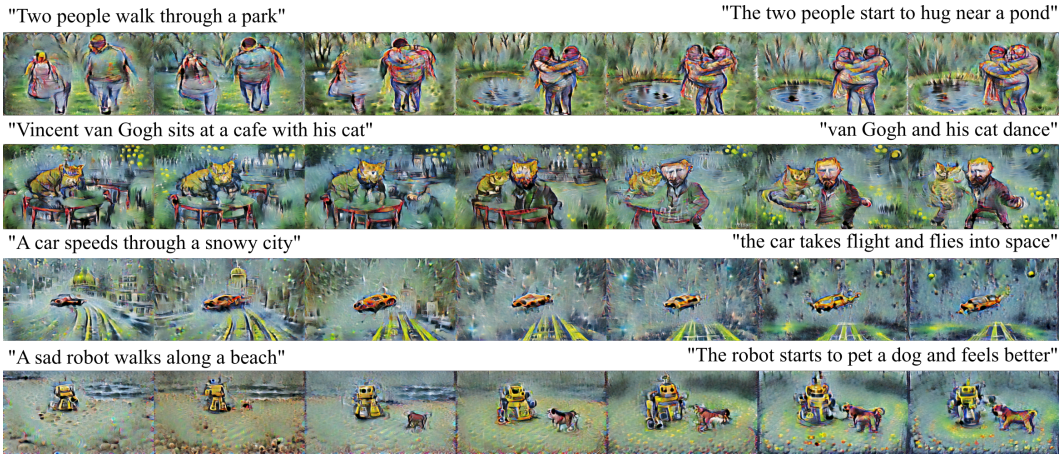

Figure 2: Samples from 60 frame videos generated using two text prompts, each in $\sim$ 100 seconds.

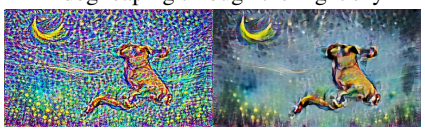

Figure 3: Our trained CycleGAN [10] model denoising the directly optimized pixels (left).

$$\mathcal{L}_{text} = w_n \cos(E(\texttt{frame}_t), E(\texttt{prompt}_n)) + w_{n+1} \cos(E(\texttt{frame}_t), E(\texttt{prompt}_{n+1})) \quad (1)$$

$$\mathcal{L}_{stable} = w_c ||\texttt{frame}_t - \texttt{frame}_{t-1}||_1 \quad (2)$$

$$\min_{\texttt{frame}_t} \mathcal{L}_{text} + \mathcal{L}_{stable} \quad (3)$$

To generate the semantic content, we alter the pixels of the video frame directly to decrease the difference between the frame and the text prompts. We generate each frame sequentially while iterating through given text prompts to guide the content. The first frame is initialized with uniformly distributed noise. We draw on past CLIP-Guided techniques [2, 8, 3, 9] to compare the frame and the language description: the augmented frames and the text prompt are encoded using CLIP [6] (E) and compared using cosine distance (Eq. 1). For the first frame, only the first prompt is used, and for subsequent frames, a linearly interpolated weight determines how much each prompt influences the generated frame. The initial state of each frame past the first is the prior frame plus some noise. Using the prior frame encourages frame to frame consistency of the locations of the content, while adding noise ensures that the frames do not get stuck in local minima and some modifications are made between frames. An extra measure to ensure the frames do not change too drastically is adding a video stability loss (Eq. 2) with a small amount of weight.

Our objective function (Eq. 3) encourages semantic content generation but does not influence the appearance or texture producing images that are very noisy in appearance. We train CycleGAN [10] to denoise these images, training the model to translate from noisy video frames to photographs [5]. Fig. 3 shows examples of the translation. The frames processed with CycleGAN do not appear photographic, but are smoother and have more realistic colors.

A user can specify the number of frames for each text prompt they give. A temperature parameter controls strength of the video stability loss (Eq. 2) and the amount of noise to add between frames.

Results We demonstrate that our approach can generate images of general concepts in different configurations/combinations as shown in Fig. 2 with more results at: `https://pschaldenbrand.github.io/text2video/`. Empirically, our approach is 20-300 times faster than existing approaches [1, 7, 4].

Discussions The appearance of the videos can be noisy and bizarre and is currently beyond user control. In future work, we intend to add priors to ensure smoother motion in the videos and more user control over the appearance and style of the videos.

Ethical Considerations Visual content generation is controlled by CLIP [6] in this approach. CLIP is known to carry over biases from its training data which reflect many wrong and harmful ideals. It is important that the user of this work understands the biases and tendencies of CLIP. It is also important that users do not use this work to create misinformation or harmful content.

References

[1] Katherine Crowson et al. Disco diffusion. `http://discodiffusion.com/`, 2021.

[2] Kevin Frans, LB Soros, and Olaf Witkowski. Clipdraw: Exploring text-to-drawing synthesis through language-image encoders. *arXiv preprint arXiv:2106.14843*, 2021.

[3] Federico Galatolo., Mario Cimino., and Gigliola Vaglini. Generating images from caption and vice versa via clip-guided generative latent space search. *Proceedings of the International Conference on Image Processing and Vision Engineering*, 2021.

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

[5] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015.

[6] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020, 2021.

[7] 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.

[8] Peter Schaldenbrand, Zhixuan Liu, and Jean Oh. Styleclipdraw: Coupling content and style in text-to-drawing translation. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 2022.

[9] Amy Smith and Simon Colton. Clip-guided gan image generation: An artistic exploration. *Evo* 2021*, page 17, 2021.

[10] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

A Temperature Parameter

We explored the effects of the temperature parameter (Fig. 4). Temperature is a value between 0 and 100 that users specify to control how much frame-to-frame differences should be encouraged. Temperature controls the weight of the video stability loss (w_c in Eq. 2) and the standard deviation of the noise added to frames when initializing the subsequent frame with the previous. The exact relationship between temperature and these parameters was hand written based on observations of the effects of the parameters. Objects and settings with low temperature barely move frame-to-frame, while with high temperature the scene completely changes (Fig. 4).

B Resolution and Aspect Ratio

A user specifies the resolution of the generated video a priori. CLIP requires 224×224 resolution input images, however, with the augmentation step, video frames can be cropped and resized to fulfill this requirement. Our approach operates with arbitrary resolutions, but the resolution does impact both the content and the appearance of the generated video frames. Lower resolutions produce simpler and smoother video frames, while a high resolutions are noisy, sparse, and do not align to the text prompt as well as medium resolutions (Fig. 4). To investigate our post-processing model's involvement in the altering of the appearance of generated frames at different resolutions, we resized the same image to multiple resolutions prior to post-processing (Fig. 6). At low resolutions, our post processing model, CycleGAN, smooths the images greatly but has little effect at denoising with high resolutions.

Figure 4: Sampling 6 (out of 60) frames from generated videos using two input text prompts and varying the temperature parameter.

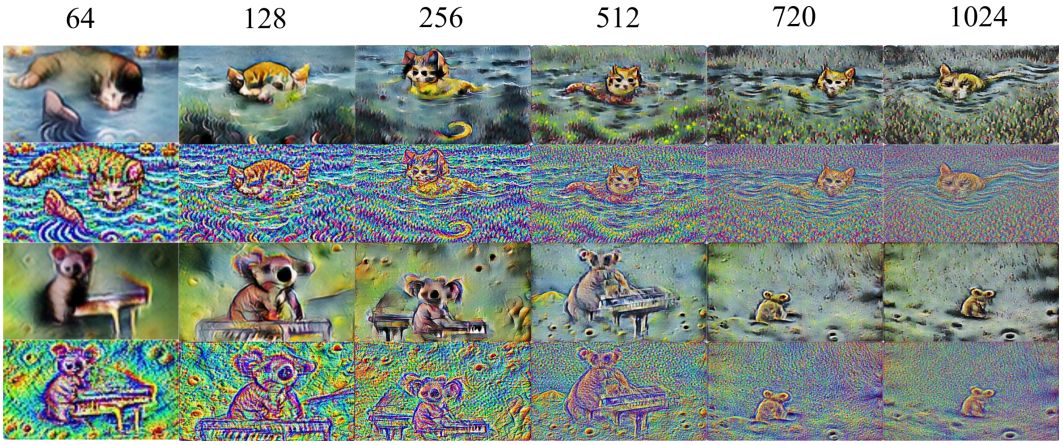

Figure 5: Generating video frames with varying resolutions. The vertical pixel resolution is shown above each generated frame with the post-processed frame displayed above the pixels that were directly optimized. The language prompts used: (top) “A cat swimming in the ocean" and (bottom) “A koala playing the piano on mars".

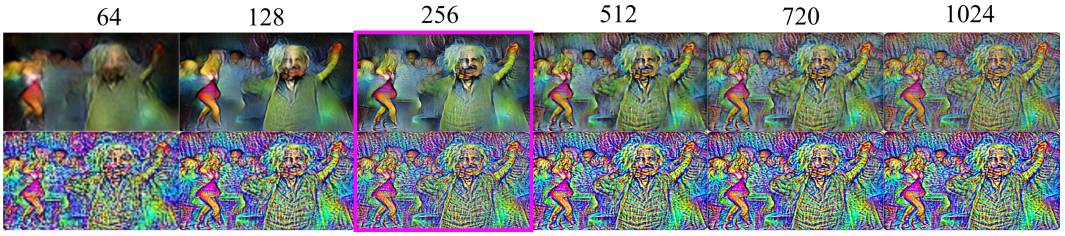

Figure 6: A video frame with a vertical height of 256 pixels was generated with the language description “Albert Einstein dancing in the club". The image, prior to post-processing with CycleGAN, was scaled to different resolutions to investigate CycleGAN’s denoising abilities on different sized images.