Avatar Video Editing Using Diffusion

Rishi Jain

Avatar editing using text-conditioned diffusion is a unique problem for a few reasons. While diffusion models have been incredibly successful for image generation and editing, they have yet to achieve similar results for video. Specifically, some problems that arise involve difficulty producing temporally consistent frames over long periods, maintaining foreground/background details, and synthesizing identifiable faces.

I. Long Video Editing

Diffusion for text-conditioned video editing/synthesis is a relatively new field, with many different approaches being researched in the last year. One approach that many projects take is by fine-tuning image diffusion models along with additional modules for video generation. Some approaches add a temporal dimension to components of pre-trained image diffusion models or train additional spatio-temporal attention modules. Other training-based approaches require fine-tuning on the original video itself. Due to these methods being under review, model weights are largely not available and training requires large video datasets/high compute requirements.

The alternative zero-shot methods are more attractive for the purposes of the assignment. However, most methods of text-guided video editing models can only process shorter sequences of frames. After exploring many options based on compute restrictions, I found that the best option was LOVECon, which expands on self-attention methods like FateZero with a windowed approach. This allows for longer video editing with cross-window attention to ensure consistency across windows while reducing the GPU memory requirement. Given that avatars for speech require higher frame rates, this option was the most computationally feasible in order to produce consistently edited long videos. Additionally, the DDIM inversion scheme maintains important color/shape consistency when performing attribute editing for foreground objects which is essential for maintaining avatar identity.

II. Controllable Video Editing

We also wanted to ensure that the edited video is plausible given the original video, allowing us to both maintain the actress identity and ensure that edits in subsequent frames are temporally consistent and smooth. To ensure this, the method I use for LOVECon is based on ControlNet, specifically fine-tuned to incorporate Canny edge detections. Conditioning the edits on Canny edge maps ensures that all the lines align in subsequent frames, details which are often lost in DDIM inversion. Empirically, Canny edge detections work best for situations where the avatar includes the actress body and additional background details. Additionally, the precise edges that this provides over other conditioning methods such as HED and pose are well-suited for the identity maintenance problem described in Section III.

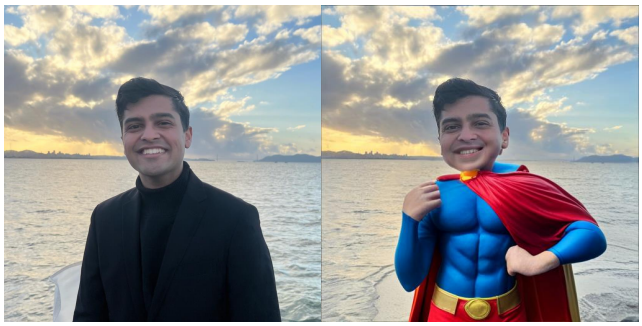

Fig. 1: Example of inpainting only the outfit. Here, the text conditioning is to show me as superman. Clearly, the face is also distorted using this method.

III. Avatar Identity Maintenance and Speech Alignment

Stable Diffusion and other image diffusion models tend to perform badly at generating realistic faces without task-specific tuning. Additionally, image editing diffusion models struggle to maintain the identity of the person depicted in the original image.

There are a few solutions I considered for this. First, I considered using an inpainting model that masks only the outfit of the actress. This method would leave the face unchanged, but experiments showed that applying a mask only to the outfit produced images that had a disconnect from the face, as seen in Figure 1. Additionally, there is evidence in other research that using an image inpainting model for video diffusion as detailed in LOVECon would not work. Next, I considered using DreamBooth to fine-tune Stable Diffusion so it has a language-aligned understanding of the actress's identity. This poses two problems. The first is that this method is notoriously hard to train well without a large corpus of diverse training images, and the other is that no text-conditioned video diffusion method can produce the articulatory facial movements needed for speech. As realistic speech is a crucial component of this assignment, this method is not viable.

My solution is to instead run the Canny-conditioned ControlNet (which maintains facial outline), and then edit the face back in frame-by-frame using the original video. This way, the facial movements remain aligned with the audio and

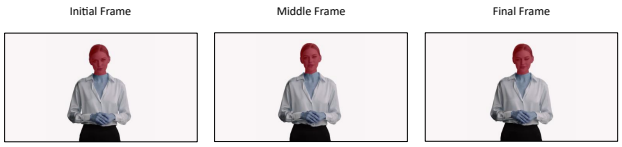

Fig. 2: Propogations of the initial frame segmentation mask using XMem.

the actress identity remains unchanged, while the changes made during diffusion are seamless with the face.

The implementation for this can be seen in Figure **??**. In the first stage, I use Segment Anything to segment the head of the actress in the first frame. Then, I use XMem to propogate this segmentation map across the remaining frames. The resulting segmentation will already be aligned with the output of the Canny-conditioned video diffusion model, and the face would be pasted over per frame.

IV. Complications

Despite my efforts, I was unable to produce the final video. Due to unforeseen technical issues out of anybody's control, my access to compute resources limited me from successfully running any of the video diffusion methods I researched. Given what was available to me, I made an attempt to try a few options. Unfortunately, the GPU memory requirement for this task, as video requires parallel processing of high dimensional images, made it impossible to run. However, I have code for the segmentation propagation and my intended solution available at . Additionally, I am happy to elaborate further on my exploration and why I arrived at the solution that I did. If I were to try my solution with a better GPU, I am confident that my solution will work well.