Applying MLLM-Guided Image Editing (MGIE) to Video Editing

Alisha Chulani

March 18, 2024

Abstract

This paper will outline the progress made to implement MLLM-Guided Editing (MLIE) to video editing processes. This paper will first provide a background of existing MLIE models that are used and manipulated for video editing. The paper will also provide an overview of the setup for video editing, which utilizes an existing HuggingFace model and a python script for short videos thus far [FHD+24]. Lastly, the paper will detail shortcomings of the current process as well as next steps for this project.

1 Introduction

The process of image editing using complicated software tools can often take excessive time and repeated steps in order to make small changes such as color, light editing, and insertion or removal of objects. The combination of instruction based editing and a diffusion model is opening up options for easy-to-use image editing using simply human input dictation rather than the fine-tuning of multiple settings on an editing application.

However, the problem with instruction based editing is that the model is unable to hold control over the human input, which means that there is no requirement for a certain level of specificity (or generality). Thus, the implementation of multimodal large language models (MLLMs) into instruction based editing allows for more accurate editing systems that can adjust the human input before completing the editing process.

This project looks to extend this application of MLLMs combined with instruction based editing to videos rather than simple pictures. This is possible through the use of existing image editing models and python scripts that can go through videos frame-by-frame to feed through the model.

2 Existing Model

This section will detail the model that is used in the video editing project. The project is based off of the recent conference paper at ICLR 2024, GUIDING INSTRUCTION-BASED IMAGE EDITING VIA MULTIMODAL LARGE LANGUAGE MODELS by Tsu-Jui Fu, Wenze Hu, Xianzhi Du, William Yang Wang, Yinfei Yang, and Zhe Gan, using models from Apple and UC Santa Barbara.

2.1 Theory

The MLIE model is broken up into three processes. The model first takes in a human edit input and an image, then uses MLLMs to process the image and tune the command, then uses latent imagination to complete the edits. There is a lengthy mathematics background that follows with each of these processes which are briefly summarized in this section. First, the MLLM is used to process the image and convert it into a language form. The mathematics of this process is shown below, as included in [FHD+24]:

$$C = x_1, x_2, ..., x_l$$

$$f = Enc_{vis}(V)$$

$$x_t = MLLM(x_1, ...x_t1|W(f))$$

Figure 1: This image will be referred to as the basketball image in this paper.

where C is the input image caption, f is the extracted visual features, V is the input image, and W is the adapter to project f.

Next, the mathematical process of editing the human-input to a succinct and direct command occurs. The mathematics using a pre-trained summarizer for this process is shown below:

$$E = Summ(MLLM * ([prompt, X]|W(f))) = w_1, w_2, ..., w_l$$
$$w_{t\prime} \ = MLLM(w_1, ..., w_t 1|W(f))$$
$$L_{ins} = \sum_{t=1}^{l} CELoss(w_{t\prime} w_t)$$

where E is the revised instruction and X is the original input instruction.

The last step is to complete the image editing using latent imagination and diffusion. The MLIE model uses the following mathematics to map a meaningful latent $U = u_1, u_2, ..., u_L$:

$$u_t = T(u_1, ..., u_{t1}|e_{[IMG]} + h_{[IMG]})$$

where where e is the word embedding, h is the editing intention, and T is edit head. This mathematics process allows for users to input instructions for image editing. Use cases and applications of this process are explored in the next section.

2.2 Application

The existing MLIE model can be used for a wide variety of application of a user's choosing with varying success rate. This section will include an exploration of the different capabilities of the tool, which can help inform how successful the video editing output will be.

To explore this section, we will start with three images for a test set. Each image will be placed under a different style of editing command - a color editing change, an object removal, and an artistic editing change.

The images of interest are detailed below:

1. Figure 1: an image of a player dribbling a basketball
2. Figure 2: a painting of a cottage
3. Figure 3: a sketch of people holding umbrellas under the Eiffel Tower

The same commands will be used for each of the images, as detailed below:

1. Change *item* to the color green.
2. Remove *item*.
3. Make this like a pencil sketch.

Figure 2: This image will be referred to as the cottage image in this paper.

Figure 3: This image will be referred to as the umbrellas image in this paper.

Running each of these images through the instructing based editing model yields varying results as follows. First, the color changing to green produces relatively accurate results. While the results do not necessarily look believable as one would possibly produce through editing applications, the MLIE model is able to produce correct outputs that do in fact, change items green. Some figures produced are more believable than others, as the "green" that is used to overwrite the color is the same green regardless of setting of the picture. As seen in Figures 4, 5 and 6, all the greens are the same shade, but fits best into the cottage image, less so in the umbrella image, and almost not at all in the basketball image.

The next command which deals with the removal of an object from an image also yields intriguing outputs, as shown in Figures 7, 8, and 9. The cottage removal and umbrella removal entirely did not work (the cottage is simply faded and the umbrellas are choppy), and on the other end of the spectrum, the basketball *and* the the basketball player is removed as a whole, leaving an empty chunk in the middle of the image.

Lastly, the final command looks at editing the style of the images as a whole. Changing both the real-life basketball image and the artistic renditions of a cottage and the people under the Eiffel Tower to pencil sketches results in impressive results as shown in Figures 11, 10, 12. The images each come out with distinct pencil sketches, although they vary slightly in terms of style. Certain parts of the images, such as the jersey in the basketball image, still seem out of place as they are the only parts

Figure 4: Edited Cottage image using "Change cottage to the color green" as the input.

Figure 5: Edited Basketball image using "Change basketball to the color green" as the input.

Figure 6: Edited Umbrella image using "Change umbrellas to the color green" as the input.

Figure 7: Edited Cottage image using "Remove cottage" as the input.

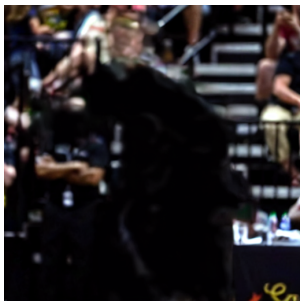

Figure 8: Edited Basketball image using "Remove basketball" as the input.

Figure 9: Edited Umbrella image using "Remove umbrellas" as the input.

Figure 10: Edited Cottage image using "Make this a pencil sketch" as the input.

done in color, but overall, the pencil sketches were successfully run across all images.

2.3 Evaluation of Existing Model

- works relatively well with color changes - removal of objects is hard as it has to fill in an unknown back

The MLIE model is useful for certain edit inputs more than others. As shown in the Application section, the MLIE model is relatively accurate for for color changes, not accurate for removal of objects, and very accurate for style editing. The edits also vary in accuracy based on the style of the input picture - real-life images are difficult to edit accurately while cartoon or simplified artistic renditions of images are more accurately edited. It can be assumed that removal of objects is poorly implemented because the MLIE model then has to make further assumptions about the background of the picture, filling in empty spaces without having full context on the image. Colors tend to be more accurately done by the MLIE models because there is no reconstruction of the image, but the MLIE model often fails to correctly detect the edges of the objects to stop filling in color (which is seen in 5). Pencil lining, or stylistic editing, however, is very accurate because the MLIE model does not need to reconstruct pieces of the image or detect edges of different objects.

Figure 11: Edited Basketball image using "Make this a pencil sketch" as the input.

Figure 12: Edited Umbrella image using "Make this a pencil sketch" as the input.

Figure 13: Single frame of car video input.

3 Video Editing Model

This section will highlight the work done to set up a video editing model as well as similar tests performed in Section 2 on videos.

3.1 Setup

The project utilizes the existing MLIE model hosted on HuggingFace. It is built off of a hand written Python script that loads in this model's API as well as a movie of the user's choosing. Simply, it loops over the movie and pulls out frames to perform the instruction-based editing process on each frame before re-linking the frames back into a video that closely resembles the original video with the user-fed edit.

3.2 Application

In order to test the success of this project, a video of a car driving along a highway with balloons on the top is used. Videos cannot be included in this paper, so a screenshot of one of the frames is shown in Figure 13 The video is tested under the same conditions as the Application section for the overview of still picture MLIE. A single screenshot of the videos of each of the ouptuts are shown in Figures 14, 15, and 16.

Complete edited videos can be found on the GitHub for this website. While choppy, the actual editing process seemed to work better for this clip than seen in the MLIE image editing exploration. While conclusions cannot be drawn as to the reasoning for this (the video shows a real-life car, so

Figure 14: Single frame of car video edited with "Change car to the color green".

Figure 15: Single frame of car video edited with "Remove balloons".

Figure 16: Single frame of car video edited with "Make this into a pencil sketch".

artistic style is not a factor), we can spot interesting trends in the editing success. Pencil sketching seems to be the most successful as edited frame by frame, but an interesting issue shows up as each frame is in a different style of "pencil sketch," as seen in Figure 17 which looks different from the initial pencil frame in Figure 16. The editing was not as accurate with the color editing, which ended up entirely removing the balloons as an unseen side effect in 14. Additionally, the removal editing was not completely accurate as pieces of leftover balloons can still be seen in Figure 15.

3.3 Drawbacks

There are multiple different drawbacks of the current video editing setup, some of which are continuations from issues in the image editing setup, and some newly arisen. First, as mentioned in the above section, the video editing cannot keep editing style consistent across frames, which can cause discontinuity between frames and make the video seem choppy and incohesive. The MLIE model does not take in the previous edited frame as an input when editing the current frame, which is a shortcoming of the MLIE model if it is to be used to complete video editing. The processing of each frame also comes with a time cost. Thus, the Python script can only access and edit every 5th frame in order to avoid taking over 10 minutes to process and using up all the GPU power freely allotted to the user in a certain time frame. If too many requests are initiated at once (ie: with a very long video, or multiple runs of the same video), the GPU times out and the user must wait for a certain amount of

Figure 17: Another singular frame of car video edited with "Make this into a pencil sketch".

time before accessing it again.

4 Conclusion and Next Steps

The current video editing system being used is useful for small editing jobs, but in the long run, would be unreasonable to use for larger-scale videos. In order to further the work done in this project, a next step would be to explore the usage of other instruction based editing tools that possibly take less GPU power and time to complete jobs. It is possible there is a trade-off between time/power needed per job and accuracy, but this trade-off can be explored depending on the style of video/request. It is also possible to create a local copy of the MLIE model on HuggingFace, and to explore the possibilities of allowing past inputs (such as previous frames) to influence the current edited frame output. GPU power can be purchased for a low price on HuggingFace, so if possible, this option can be explored in order to promote cohesiveness throughout the videos. Overall, this video editing technique is a good introduction to make simple editing processes more accessible to different groups of people. The ease with which editing is available means that tools like this can be widely distributed to encourage artistic and creative explorations by people around the world.

References

[FHD+24] Tsu-Jui Fu, Wenze Hu, Xianzhi Du, William Yang Wang, Yinfei Yang, Zhe Gan, UC Santa Barbara, and Apple. Guiding instruction-based image editing via multimodal large language models. *ICLR*, 2024.