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

Mathematics of Big Data - Spring 2024

Alisha Chulani

18 March 2024

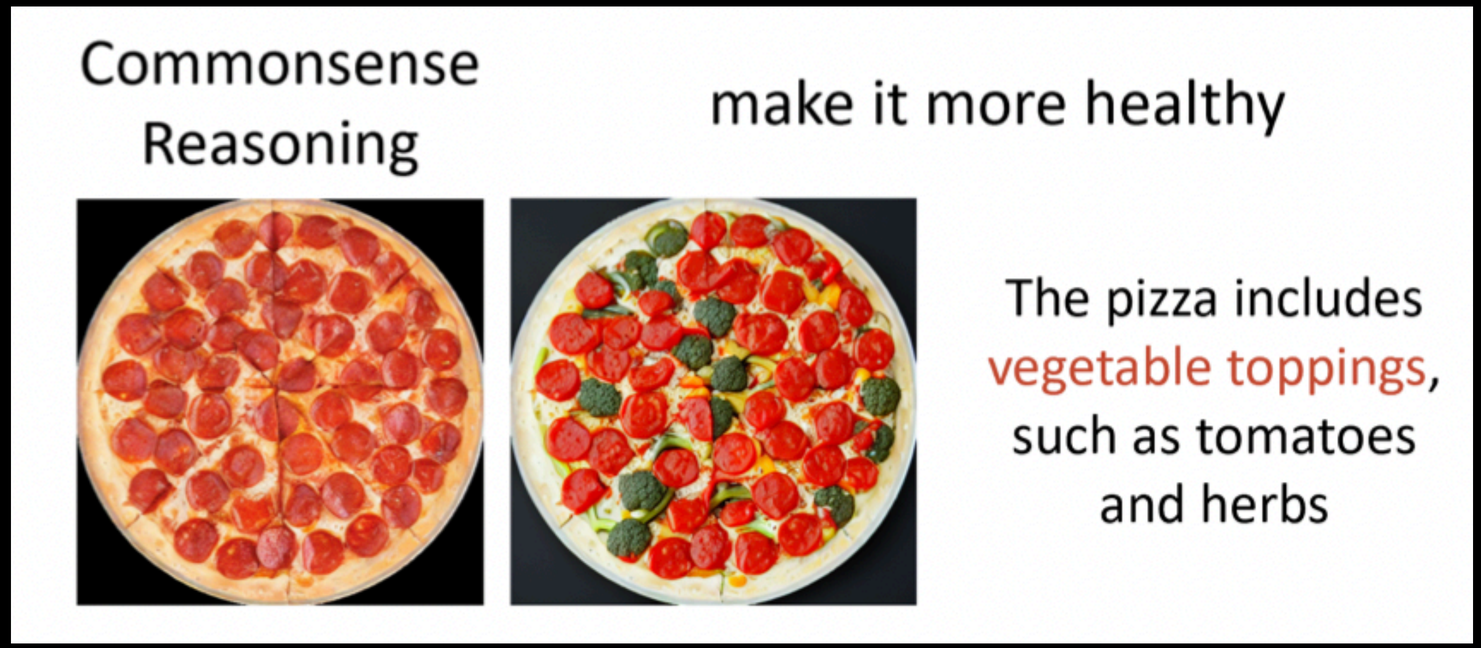
Background: Image Editing

Instruction-based editing

- Easier image editing capabilities using human-input prompts
- Can save time and money
- More accessible to different groups of people

Problems with Instruction Based Editing

- Human inputs can be unbridled
- Too specific? Not specific enough?



Solution:

MLLMs with Instruction Based Editing

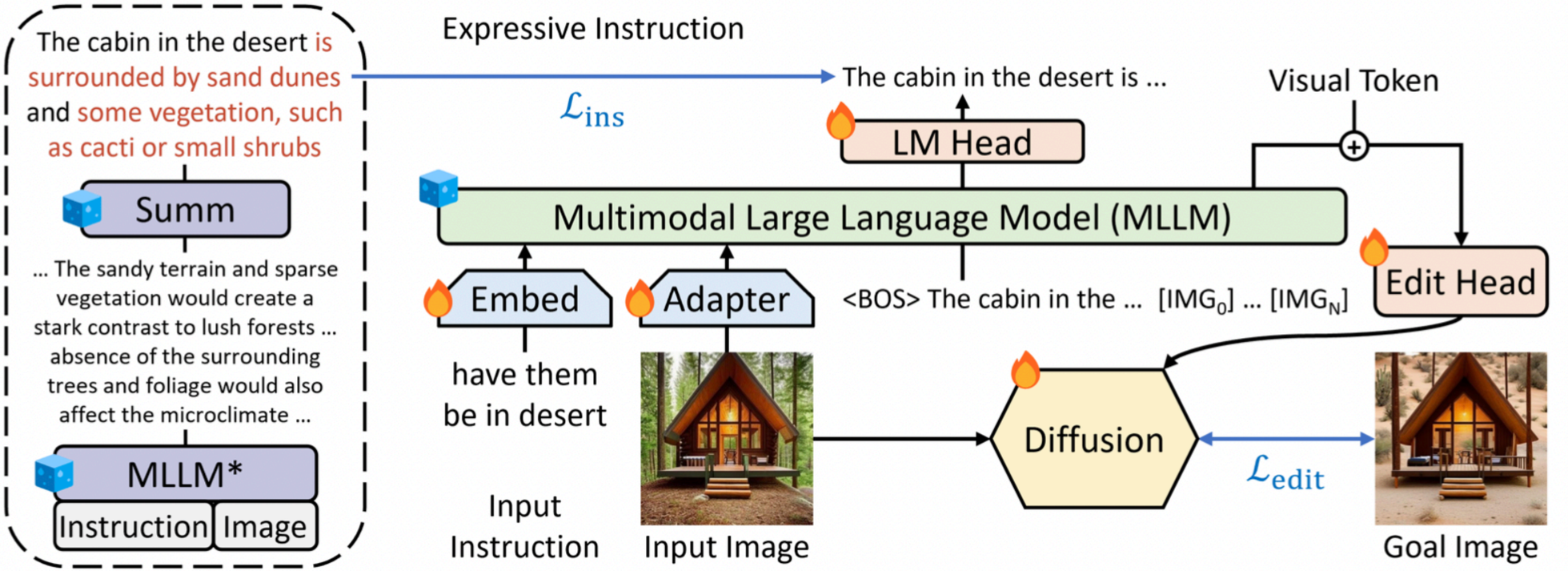


Figure 2: Overview of MLLM-Guided Image Editing (**MGIE**), which leverages MLLMs to enhance instruction-based image editing. MGIE learns to derive concise expressive instructions and provides explicit visual-related guidance for the intended goal. The diffusion model jointly trains and achieves image editing with the latent imagination through the edit head in an end-to-end manner. 🔥 and 🧊 show the module is trainable and frozen[1], respectively.

1. MLLM (LLaVa)

- f = visual features
- C = word instruction
- W = adapter to project f into language

$$
\begin{aligned}
\mathcal{C} &= \{x_1, x_2, ..., x_l\}, \\
f &= \mathbf{Enc}_{\text{vis}}(\mathcal{V}), \\
x_t &= \mathbf{MLLM}(\{x_1, ...x_{t-1}\} \mid \mathcal{W}(f)),
\end{aligned}
$$

2. Summarizer

- E = expressive instruction (summarized)
- X = long instruction

$$
\begin{aligned}
\mathcal{E} &= \mathbf{Summ}(\mathbf{MLLM}^*([\texttt{prompt}, \mathcal{X}] \mid \mathcal{W}(f))) \\
&= \{w_1, w_2, ..., w_l\}, \\
w'_t &= \mathbf{MLLM}(\{w_1, ..., w_{t-1}\} \mid \mathcal{W}(f)), \\
\mathcal{L}_{\text{ins}} &= \sum_{t=1}^{l} \mathbf{CELoss}(w'_t, w_t),
\end{aligned}
$$

3. Latent Imagination

- T = sequence to sequence model
- U = meaningful latent
- e = word embedding
- h = hidden state

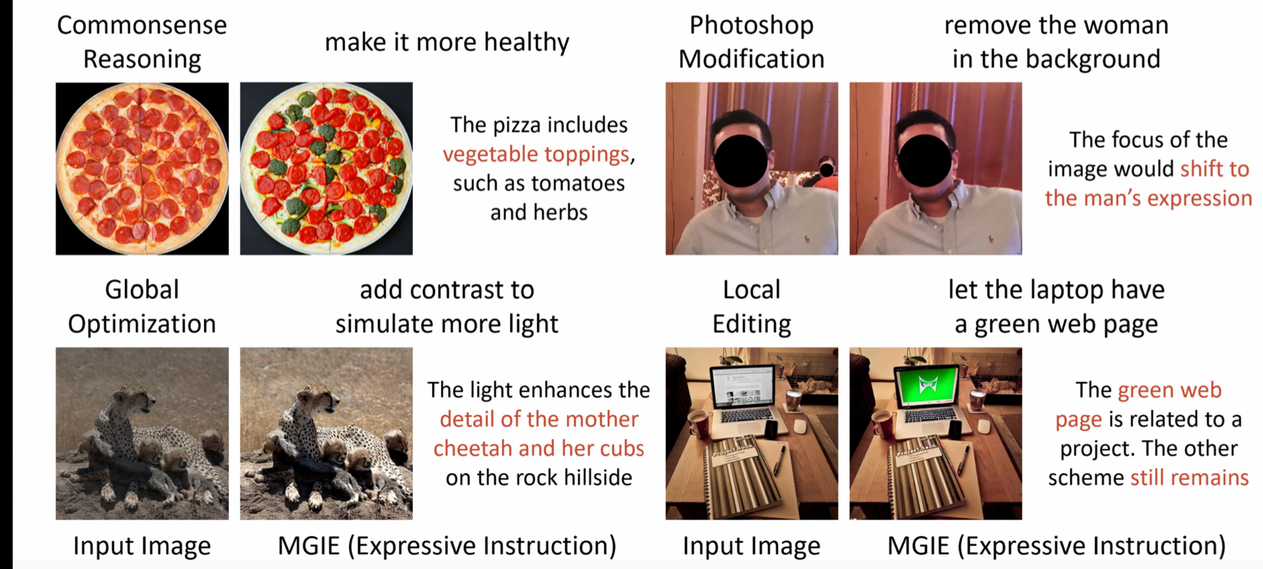
$$u_t = \mathcal{T}(\{u_1, ..., u_{t-1}\} \mid \{e_{\texttt{[IMG]}} + h_{\texttt{[IMG]}}\}),$$

Idea

- Extending instruction-based editing to videos

GUIDING INSTRUCTION-BASED IMAGE EDITING VIA MULTIMODAL LARGE LANGUAGE MODELS

Tsu-Jui Fu[1], Wenze Hu[2], Xianzhi Du[2], William Yang Wang[1], Yinfei Yang[2], Zhe Gan[2]
[1]UC Santa Barbara, [2]Apple



Initial Exploration of MLIE

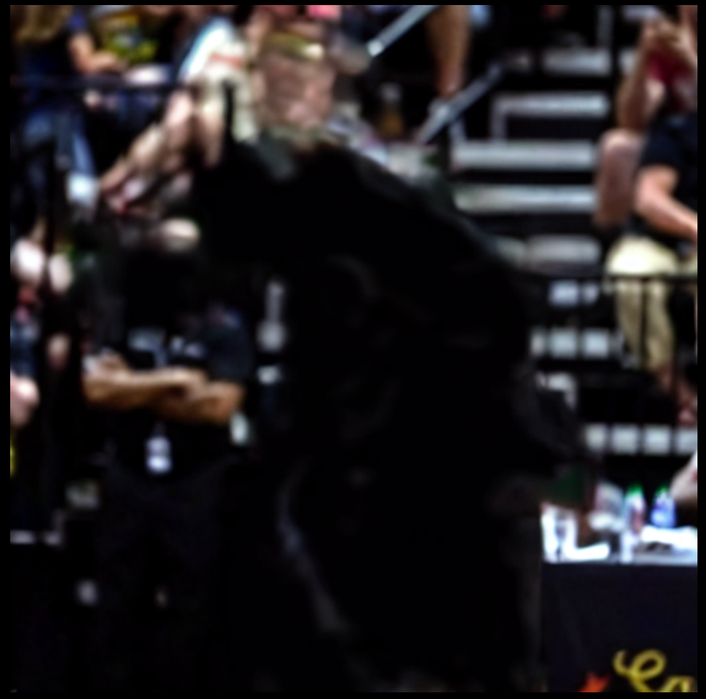
- Change *item* to the color green.
- Remove *item*.
- Make this like a pencil sketch.

Initial Exploration of MLIE

Change *item* to the color green.

Initial Exploration of MLIE

Remove *item*.



Initial Exploration of MLIE

Make this into a pencil sketch.

Evaluation

- Works well with style editing…
- Less well with color editing…
- And not very well with object removal

- Model unable to reconstruct pieces of the image when removal leaves empty gaps
- Color issues when it comes to detecting edges of objects

Initial Exploration of MLIE

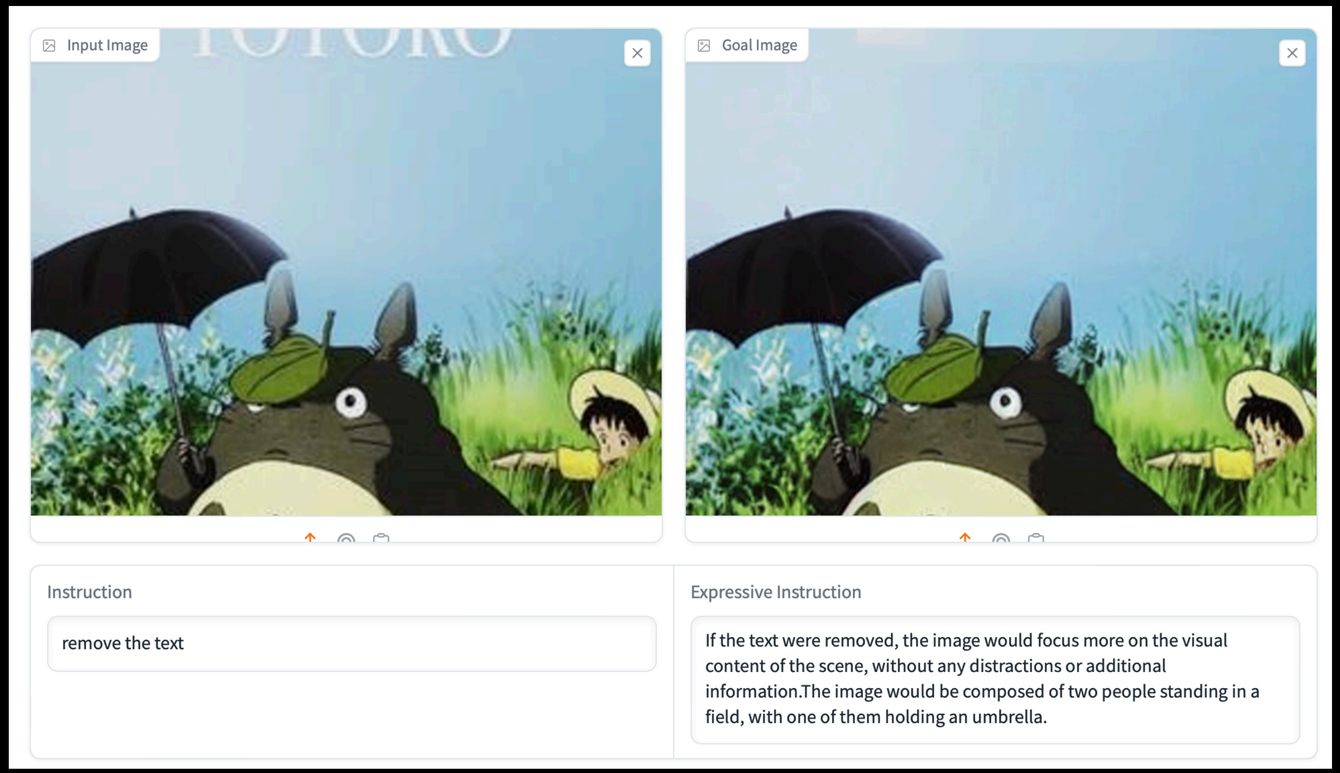
Works for certain explorations better than others…

- Can be translated into video editing too?

Project

My Idea

- Using an existing HuggingFace model
- Use a python script to extend it to edit input videos



Coding Setup

Video Examples

Run experiments similar to MLIE exploration

- Make this into pencil sketch

- Change the car to the color green:

- Remove the balloons

Next Steps for Final Project

- Speed up the process! Are there different MLIE models out there that are more efficient?
- Explore continuity for things such as pencil sketch/style editing - involves changing the model…