

Big Data Midterm Project

Alisha Chulani - March 18 2024

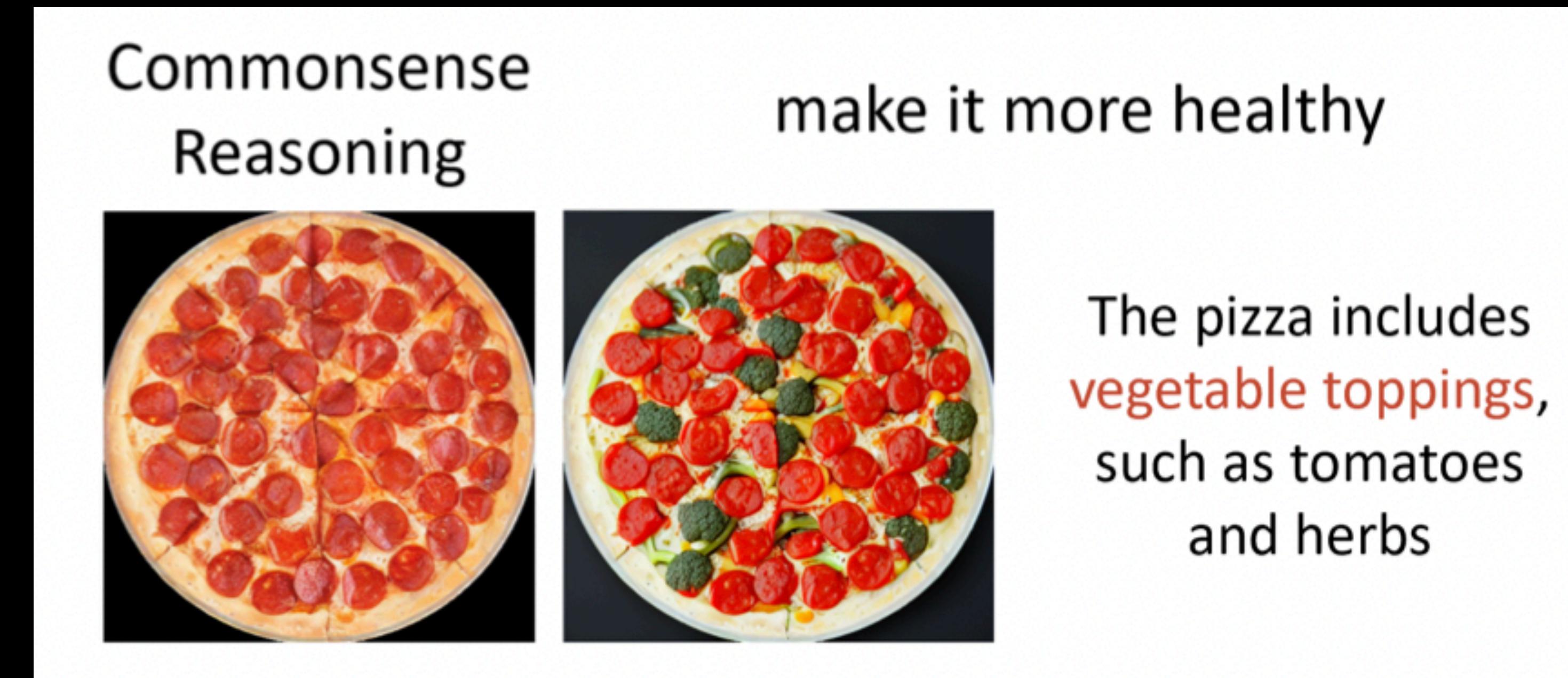
Background: Image Editing

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

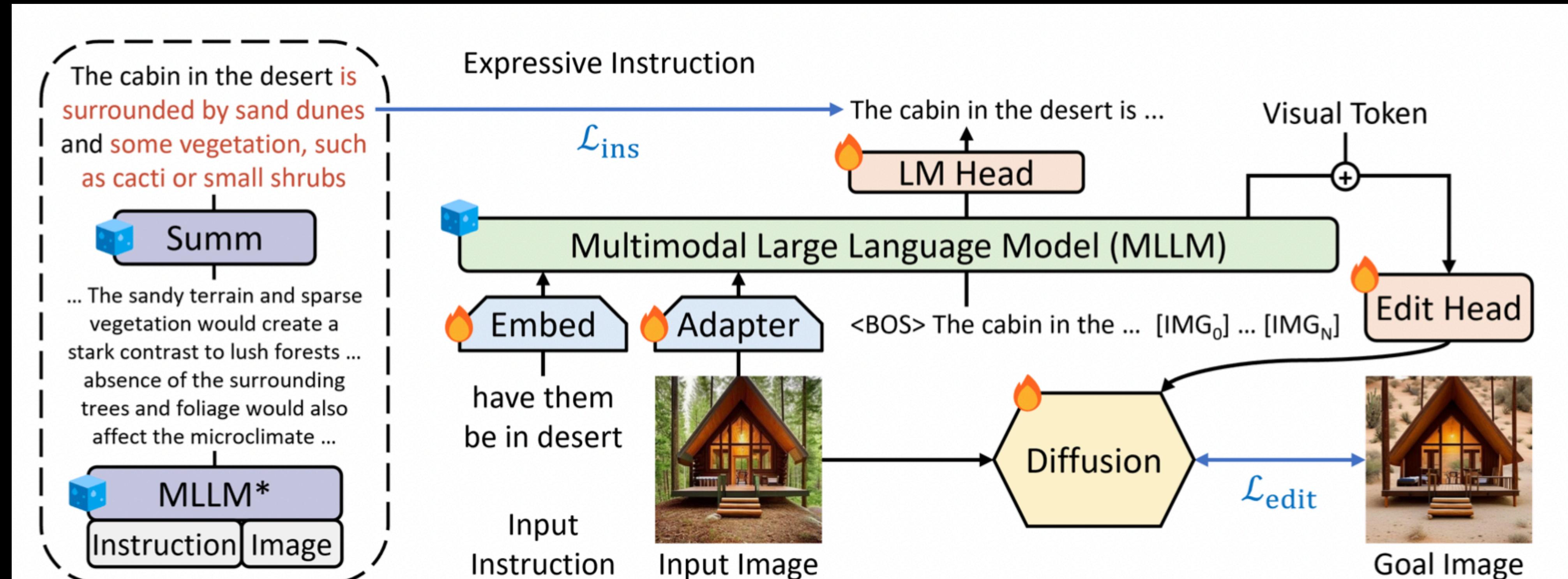


Figure 2: Overview of MLLM-Guided Image Editing (MGIE), which leverages MLLMs to enhance instruction-based image editing. MGIE learns to derive concise expressive instructions and provides explicit visual-related guidance for the intended goal. The diffusion model jointly trains and achieves image editing with the latent imagination through the edit head in an end-to-end manner. 🔥 and 🎲 show the module is trainable and frozen¹, respectively.

1. MLLM (LLaVa)

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

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

2. Summarizer

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

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

3. Latent Imagination

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

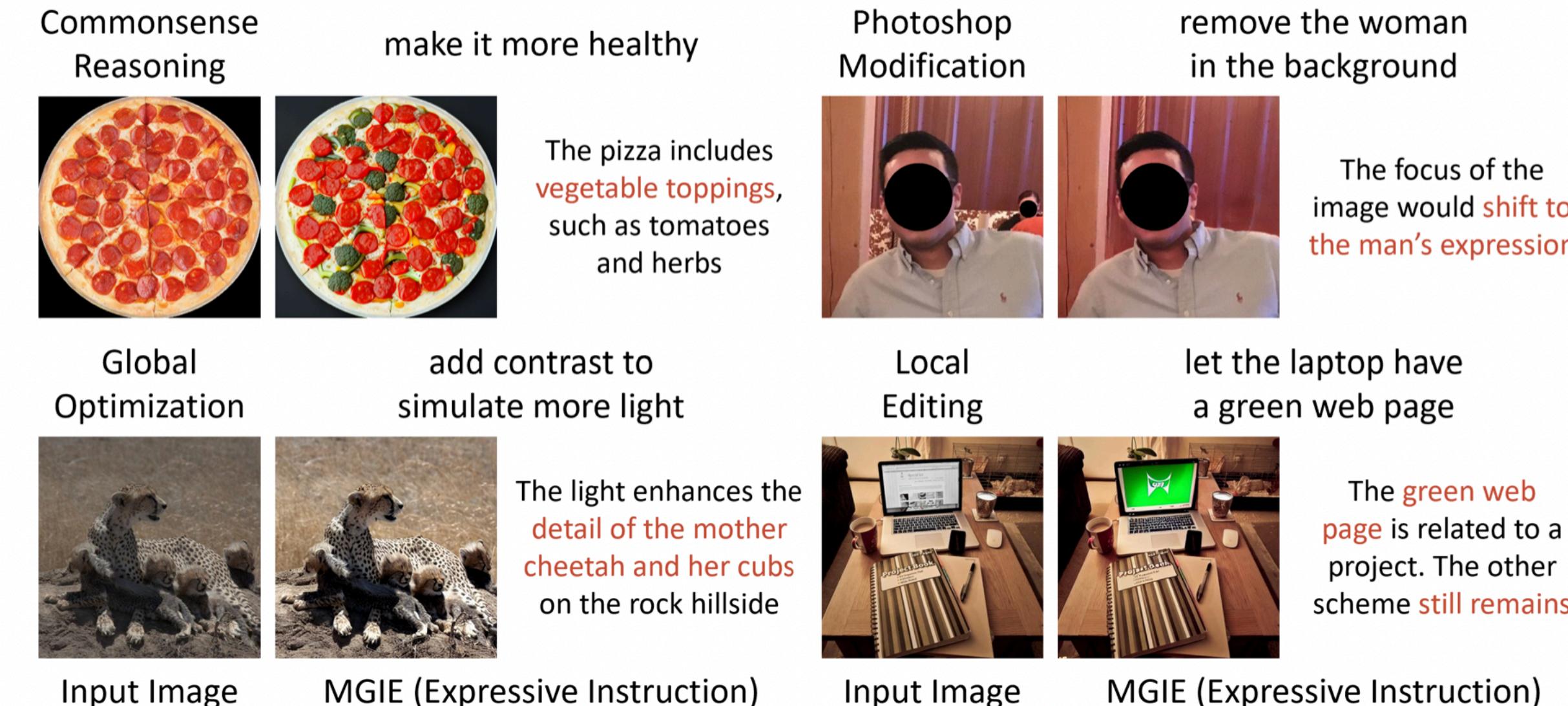
$$u_t = \mathcal{T}(\{u_1, \dots, u_{t-1}\} \mid \{e_{[\text{IMG}]} + h_{[\text{IMG}]} \}),$$

Idea

- Extending instruction-based editing to videos

GUIDING INSTRUCTION-BASED IMAGE EDITING VIA MULTIMODAL LARGE LANGUAGE MODELS

• Tsu-Jui Fu¹, Wenze Hu², Xianzhi Du², William Yang Wang¹, Yinfei Yang², Zhe Gan²
¹UC Santa Barbara, ²Apple



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

Initial Exploration of MLIE

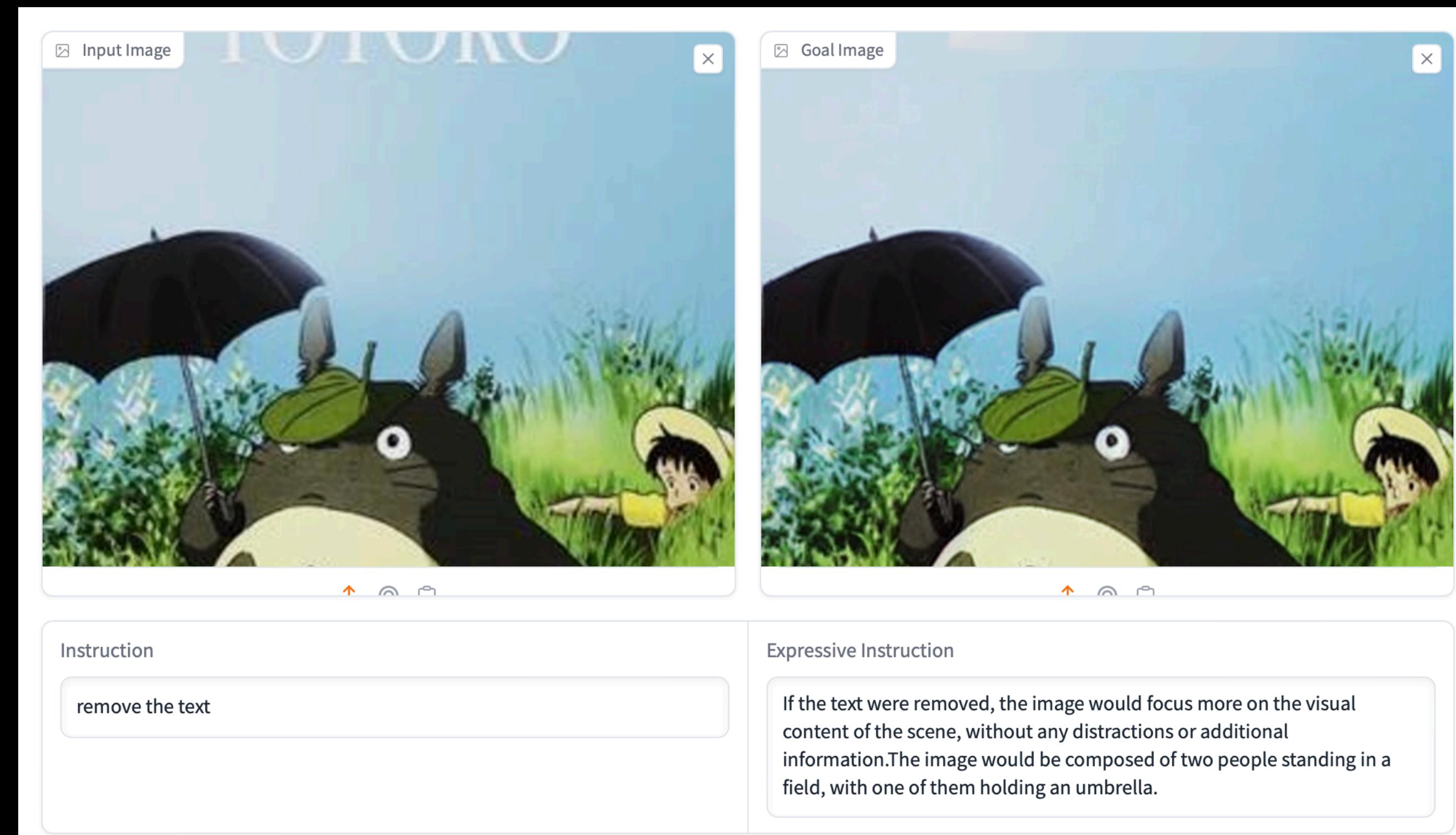
Works for certain explorations better than others...

- Can be translated into video editing too?

Project

My Idea

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

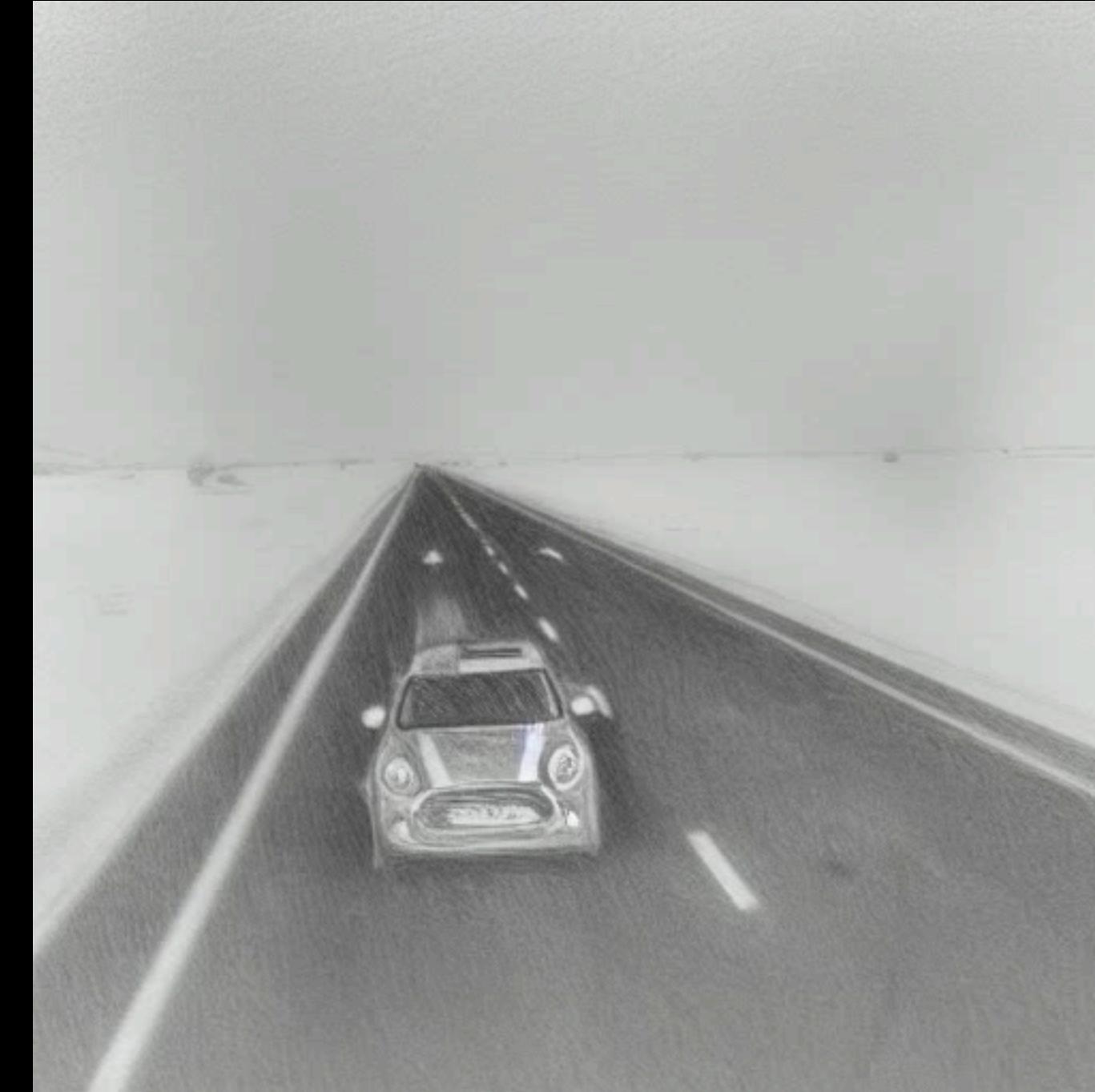


Coding Setup

Video Examples

Slow, so only small videos right now and pretty choppy

- 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...