

# Instruction-based Image Manipulation by Watching How Things Move

## Supplementary Material

### 1. Dataset Construction

In this section, we provide more details about our data construction pipeline and the resulting dataset. Sec. 1.1 explains how we prompt multi-modal large language models (MLLMs) to generate instructions. Sec. 1.2 compares different MLLMs for instruction generation. Despite our careful design of data filtering and prompting, a small amount of low-quality data remains in our training dataset, as discussed in Sec. 1.3.

#### 1.1. MLLM Prompting

We leverage the multimodal large language model Pixtral [2] to generate editing instructions for motion-filtered frames sampled from video data. The prompt template used for this process is illustrated in Fig. 2.

The task objective is first specified for the MLLM. To ensure accurate and context-appropriate instructions, we define rules for the MLLM to follow. Specifically, the MLLM is instructed to focus on primary objects in the input images and identify changes related to attributes such as shape, size, position, perspective, camera viewpoint, and other features. Additionally, constraints are imposed to ensure that editing instructions are based solely on the content of the source image, using absolute terms or terms relative to the source image, without referencing details from the target image. This restriction prevents information leakage from the target image, which is unavailable to the model during training or inference. To enhance practicality, the MLLM is directed to begin editing instructions with action-oriented verbs. For cases where changes between input images are subtle or where transformations are too complex to articulate as simple editing instructions, the MLLM outputs N/A, and these cases are filtered out.

This carefully designed prompting process allows the MLLM to generate precise and detailed editing instructions directly from input images. As shown in Fig. 4, the constructed dataset includes high-quality images paired with accurate editing instructions that effectively describe the transformations from source to target images.

We conducted an ablation study to evaluate the importance of different components in our prompt design, focusing on the guidelines and examples. As shown in Fig. 1, removing either element results in less natural and less reliable instructions. These results highlight the value of a structured prompt template in guiding the MLLM to generate high-quality outputs.



Figure 1. Instructions generated by MLLMs with different prompting templates.

#### 1.2. Comparison between Different MLLMs

We employ the recently released open-source multimodal large language model Pixtral [2] to generate image editing instructions, taking advantage of its robust multimodal processing capabilities. We also evaluate other widely adopted MLLMs, including LLaVA [6] and GPT-4o [1]. Sample outputs from these models are illustrated in Fig. 3. All three models perform well in straightforward scenarios (such as the first sample that involves simple pose and expression changes). However, in more complex cases (*e.g.*, the third sample), LLaVA struggles to produce accurate instructions for transformations requiring advanced understanding and reasoning. In contrast, both Pixtral and GPT-4o demonstrate superior performance, generating more precise and contextually accurate instructions. Given the need to construct a large-scale dataset, we ultimately decided to utilize the open-sourced model Pixtral for generating the editing instructions.

#### 1.3. Failure Cases

Despite achieving high-quality data pairs from video through the proposed dataset construction pipeline, a small amount of failure cases persist within the dataset. These failures primarily arise from two aspects: significant content changes and incorrect understanding or reasoning by MLLMs. In the first case, although motion filtering is applied during the frame selection process, some edge cases remain unfiltered. These instances involve frames with weak correspondence among the image contents, preventing the MLLM from generating accurate editing instructions (as illustrated in Fig. 5(a)). The second issue stems from limitations in the internal reasoning capabilities of MLLMs. These models occasionally misinterpret the input image content (Fig. 5(b)), leading to two types of errors: (1) failure to identify the primary transformations, resulting in inaccurate instructions, and (2) generation of instructions irrelevant to the input images.

Interesting future work could address these challenges. One direction is to enhance the motion filtering procedure

**Task Prompt:**

You are provided with two images sampled from a video. Assume an image editor accepts text instructions to modify the first image. Your task is to output an instruction for the editor to transform the first image to resemble the second image. The instruction must be concise, clear, and no longer than 20 words.

**Guidelines:****1. Focus on the Main Elements:**

Describe changes to key objects in the image, in terms of shape, shape, position, orientation, perspective, and background, *etc.*

**2. Use Only the First Image for Context:**

Generate the instruction based solely on the first image. Do not use descriptive references to the second image.

**3. Use Absolute or Relative Terminology Regarding the First Image:**

Specify changes directly and concretely, referencing the first image for orientation or context when necessary. Sequential instructions are allowed, as long as they remain concise and directly address the editing process. For example: Rotate the object clockwise by 30 degrees, then move it to the top-left corner. Make the woman look to her left.

**4. Action-Oriented Language:**

Start the instruction with verbs like Make, Have, Change, Adjust, Move, *etc.* Alternatively, propose edits with How about.

**5. Fallback for Subtle or Complex Transformations:**

If the transformation from the first image to the second image is too subtle to identify or too complex to achieve, output N/A.

**Response Format:**

```
{
  "instruction": "<Your concise editing instruction>"
}
```

**Example:**

Context: A person is looking straight in the first image and slightly turned to their left in the second image.

Response:

```
{
  "instruction": "Make the person turn slightly to their left."
}
```

Figure 2. Multimodal LLM prompting process.

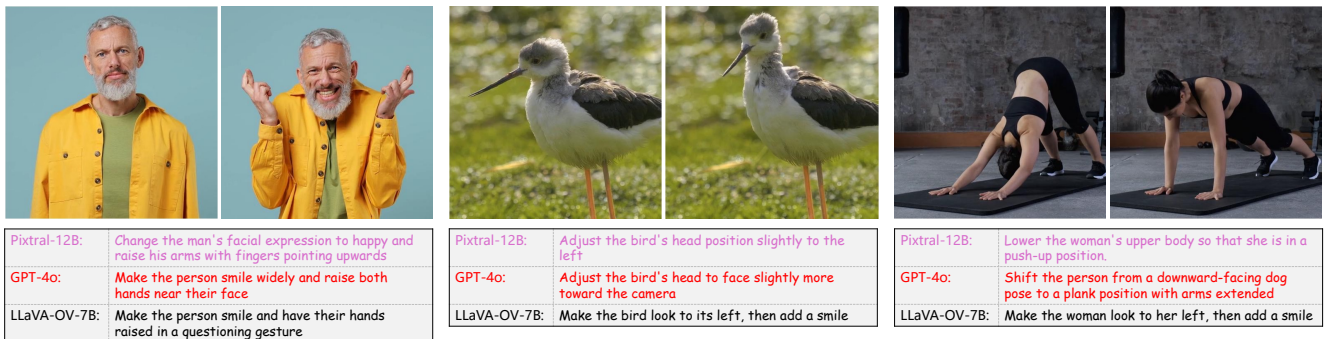


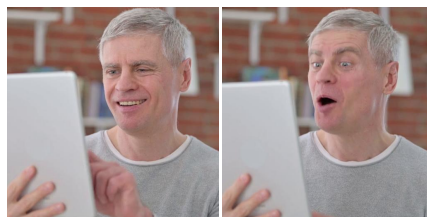
Figure 3. Image editing instructions generated by different multimodal LLMs.



Rotate the robot to the left so that its left hand is visible and its body is slightly turned



Make the man smile



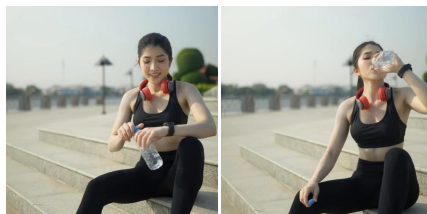
Change the man's expression to surprise



Move the hands to the right side of the keyboard



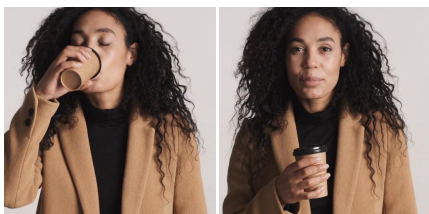
Remove the hands from the neck and place them by the sides



Change the person's action to drinking from the water bottle



Make the person wave their right hand



Change the person's pose to have them holding a coffee cup in their right hand and smile



Lower the person's body closer to the ground



Rotate the bird so that it faces to the right



Adjust the seal's head to look upwards and slightly to the left



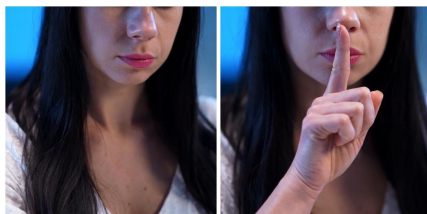
Fill the bowl with batter



Make the cat's face closer to the camera



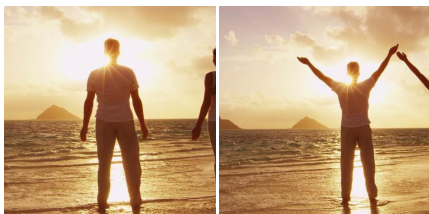
Change the cat's orientation to the right



Add a finger near the lips, making a 'shushing' gesture



Remove the bubbles from the drink



Raise the person's arms upward towards the sun



Remove the person's hands from the image

Figure 4. Samples in the constructed dataset.



to ensure the selection of more suitable frames for image editing. Additionally, refining the prompting process for MLLMs and exploring the use of more advanced, larger-scale MLLMs could improve the accuracy and relevance of the generated editing instructions.

## 2. Evaluation Details

### 2.1. Benchmark

We constructed a benchmark specifically designed for the editing types discussed in our main paper. The images in the benchmark were sourced from prior research works [4, 7, 9] and publicly available internet datasets. For each image, we manually created non-rigid editing instructions, such as “Change the view to the side” or “Make the man give a thumbs-up.” During the benchmarking process, each method generates four edited images per instruction, with different random seeds. The evaluation metrics included CLIP-I (calculated as the average across the four images), which quantifies appearance and identity preservation, and CLIP-D and CLIP-Inst (calculated as the maximum across the four images), which evaluate alignment with the text instructions.

### 2.2. User Study

To further validate the effectiveness of the proposed method, we conducted a user study. Specifically, we selected the best sample from the four generated images of each model to represent its output, and a total of 20 samples were chosen to form the complete user study. Each case in the study consists of a source image, a target editing instruction, and the edited results from Imagic [4], Instruct-Pix2Pix [3], MagicBrush [10], and our model. An example question from the user study is shown in Fig. 6. To eliminate potential bias, the cases and response options within each case were randomly shuffled during the answering process. In total, we collected 800 responses from 40 professional participants. The results showed that the majority of participants preferred the results generated by our method, providing additional evidence for the effectiveness of the proposed approach.

## 3. Additional Results

### 3.1. Results of the Model Trained with Mixed Datasets

Our dataset is constructed from video data. Therefore, the model trained solely on this dataset does not support artistic edits, such as global stylization, color changes, or local replacements, which is enabled by previous instruction-based image editing works [3, 8, 11]. This can be simply mitigated by combining our dataset with existing datasets. To demonstrate this, we combined the InstructPix2Pix dataset with

our dataset during training. As demonstrated in Fig. 7, the model trained on the mixed dataset successfully supports artistic edits, such as altering textures and styles while still being able to support non-rigid editing which is challenging with previous approaches. However, unlike our dataset, since InstructPix2Pix dataset contains lower-quality target images generated from zero-shot editing models, we observe a slight drop in terms of the model’s ability to preserve the original content when dealing with edits that require structural changes.

### 3.2. Results of Larger Base Models

In the main paper, we present results from a model finetuned on Stable Diffusion V1.5. However, our proposed method is not limited to any specific text-to-image (T2I) diffusion model, allowing for flexibility with more advanced models. For instance, we can finetune a larger and more powerful T2I model on our dataset using the proposed Spatial Conditioning strategy to achieve higher-quality, high-resolution image edits. FLUX<sup>1</sup> is a model [5] with 12B parameters, offering state-of-the-art performance image generation with top of the line prompt following, visual quality, image detail and output diversity. As demonstrated in Fig. 8(c), when finetuned with our approach, it can generate high-resolution ( $1024 \times 1024$ ) edits while maintaining the fine details of the source image, even at elevated resolutions.

### 3.3. Additional Qualitative Results

Additional edited results from the model presented in the main paper are shown in Fig. 8. Our method demonstrates strong performance in editing both real and synthetic images. Moreover, with the incorporation of masks or spatial controls, it seamlessly achieves precise editing.

## References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1
- [2] Praveesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Devendra Chaplot, Jessica Chudnovsky, Saurabh Garg, Theophile Gervet, Soham Ghosh, Amélie Héliou, Paul Jacob, et al. Pixtral 12b. *arXiv preprint arXiv:2410.07073*, 2024. 1
- [3] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 4
- [4] Bahjat Kavar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic:

<sup>1</sup><https://huggingface.co/black-forest-labs/FLUX.1-dev>

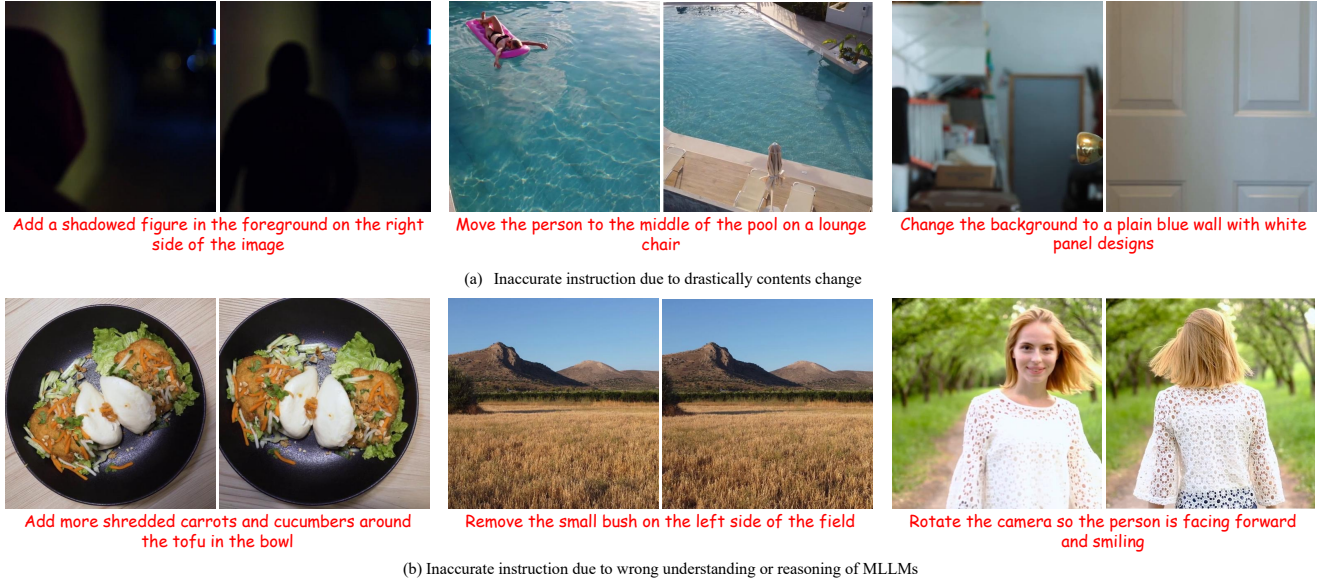







Figure 5. Inaccurate instructions due to (a) drastic content change and (b) wrong understanding and reasoning of MLLMs.

Please review the images edited according to user instructions below. For each case, select the best result among the different methods based on image consistency, instruction accuracy, and image quality.

Matrix Single Choice

\* 01 Which edited image follows the editing instruction while having the best quality and preserve the original content

Make the dog look at the camera

Best ☐ ☐ ☐ ☐

Figure 6. Illustration of the user study.

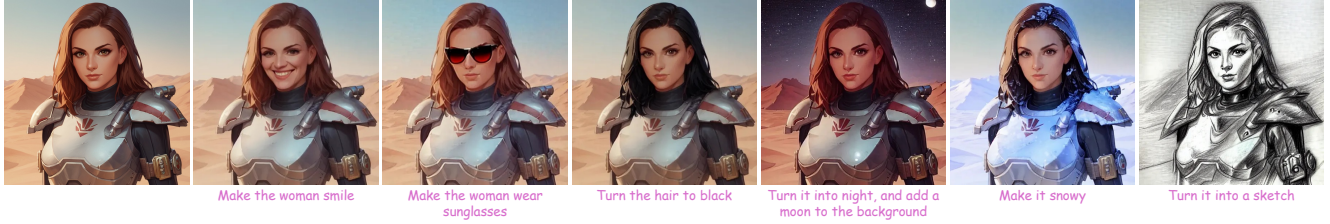
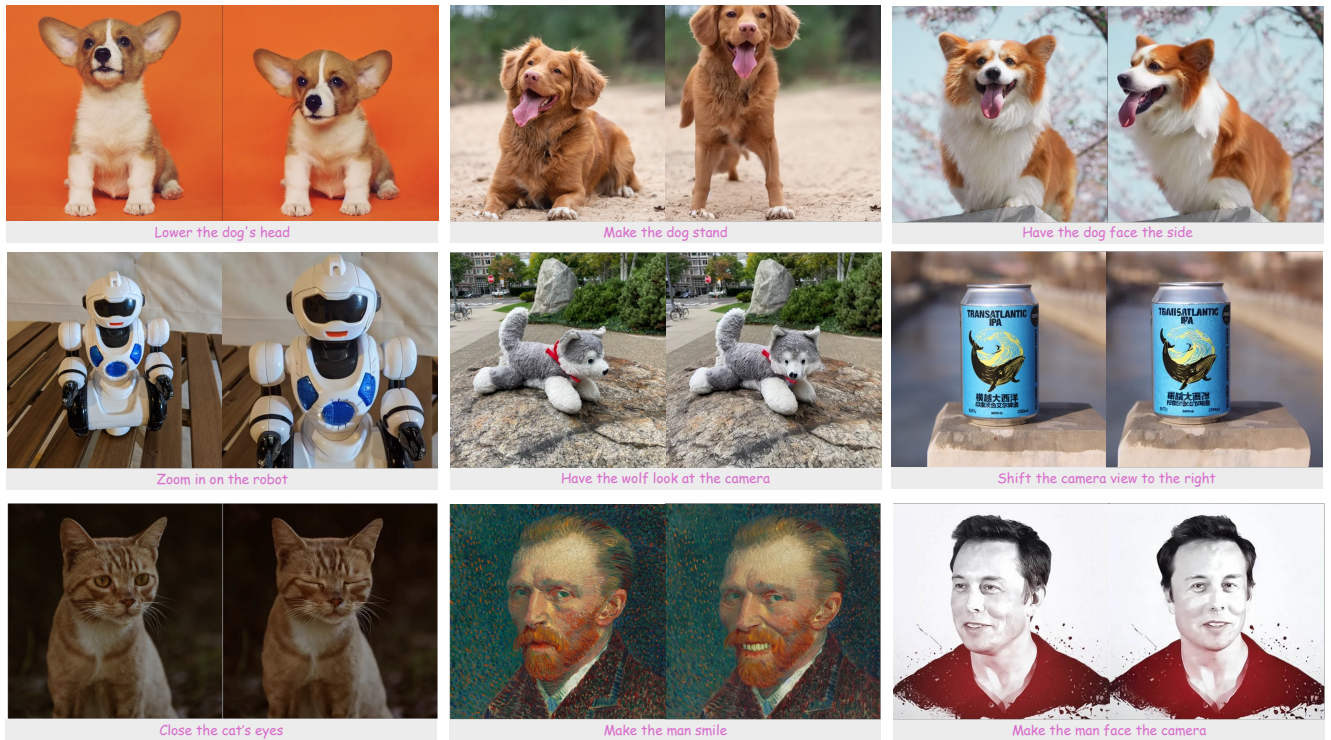


Figure 7. Results of the model trained on the mixed dataset (ours and InstructPix2Pix dataset).

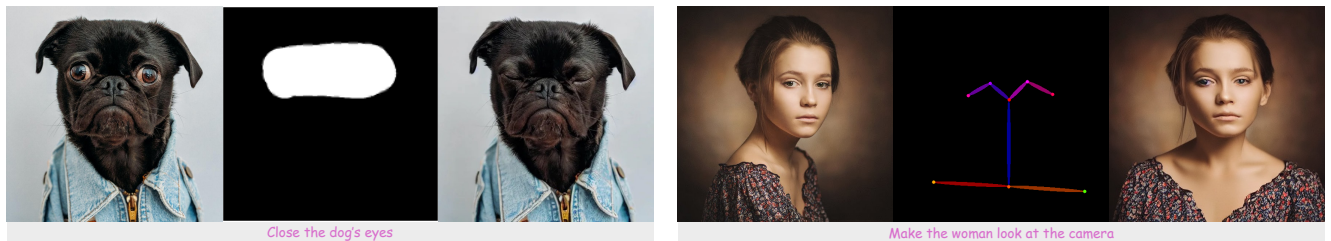
Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6007–6017, 2023. 4

- [5] Black Forest Labs. Flux.1, 2024. <https://blackforestlabs.ai/announcing-black-forest-labs/>. 4
- [6] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in Neural Information Processing Systems*, 36, 2024. 1
- [7] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023. 4
- [8] Shelly Sheynin, Adam Polyak, Uriel Singer, Yuval Kirstain, Amit Zohar, Oron Ashual, Devi Parikh, and Yaniv Taigman. Emu edit: Precise image editing via recognition and generation tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8871–8879, 2024. 4
- [9] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1921–1930, 2023. 4
- [10] Kai Zhang, Lingbo Mo, Wenhui Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. *Advances in Neural Information Processing Systems*, 36, 2024. 4
- [11] Haozhe Zhao, Xiaojian Ma, Liang Chen, Shuzheng Si, Rujie Wu, Kaikai An, Peiyu Yu, Minjia Zhang, Qing Li, and Baobao Chang. Ultraedit: Instruction-based fine-grained image editing at scale. *arXiv preprint arXiv:2407.05282*, 2024. 4





(a) Additional qualitative results



(b) Additional results with local mask and spatial controls



(c) Results of larger base model FLUX.1 Dev

Figure 8. Additional editing results using (a) text-only instructions, (b) local masks or spatial controls, and (c) fine-tuned larger and more powerful base T2I models, such as FLUX.1, for high-quality image editing.