VEGGIE 🥦: Instructional Editing and Reasoning Video Concepts with Grounded Generation

Supplementary Material

6. Appendix

In this Appedix, we provide extra details on

- Implementation details of VEGGIE training, and evaluation and baseline evaluations.
- Extra details on our data generation pipeline, including each module's details, prompts for each promptable module, data filtering, and visualization.
- Extra visualizations for each task and the comparison with the other 6 strong baseline models.
- Limitation and future work discussion.

6.1. Implementation Details

Model Architecture. Our MLLM is initialized with LLaVA-OneVision-7B (LLaVA-OV) [38]. It is a strong MLLM consisting of Qwen2 [73] LLM with 32K context window, SigLIP [80] visual encoder, and a 2-layer-MLP projector. LLaVA-OV can handle diverse visual-language tasks (including interleaved-frame, video). It provides a good starting point for our VEGGIE to understand complex user instructions and can respond with multiple frame-wise implicit planning thanks to its long context window. Our video diffusion model is initialized from the instructional image editing model, MagicBrush [83]. We further inflated 2D convolution layers to 3D form and inserted temporal attention layers following AnimateDiff [22] to adapt videos. Our alignment network is a single-layer MLP. We set 32 grounded task tokens for each frame.

Training Details. Our MLLM is initialized with LLaVA-OneVision-7B (LLaVA-OV) [38]. Our VidDM is initialised from the instructional image editing model, MagicBrush [83] with Stable Diffusion v1.5 backbone [57]. We further inflated 2D convolution layers with temporal attention layers, following AnimateDiff [22] to adapt videos. Our VEGGIE adopts a 2-stage curriculum training strategy (Sec. 3.2). In the first stage, we fully fine-tune the 2D convolution layers in the UNet, the alignment network, and the task query tokens in the MLLM on image data, with 862M trainable parameters. In the second stage, we train all 3 dimensions in the UNet, the alignment network, the task query tokens, and a LoRA in the MLLM, leading to 1.3B trainable parameters. Both stages are trained end-to-end with only a diffusion loss. More details are in the Appendix.

We keep the VAE encoder and decoder frozen during the entire training process. In the first stage, we keep the MLLM (including visual encoder, MLP projector, and LLM) frozen, and fully fine-tune learnable grounded task queries, alignment network, and diffusion model, leading to around 800M training parameters. We set $1e^{-4}$ learning rate, and 96 batch size on each GPU. We use 16 A100 GPUs for the first stage of fine-tuning with 25K steps. In the second stage, we insert LoRA [27] modules into the LLM backbone, and inflate diffusion models by inserting extra temporal layers as in AnimateDiff [22]. We fine-tune LoRA, alignment network, learnable grounded task query tokens, and the diffusion model, leading to around 1.3B trainable parameters. We set $5e^{-4}$ learning rate, and 1 batch size with 8 gradient accumulation steps on 32 A100 GPUs. For LoRA, we set lora rank 64, lora alpha 16, and lora dropout 0.05. We train the second stage video model 2.5K step with 8 uniformly sampled frames.

Evaluation and Baseline Details. We primarily compare our model with strong instructional editing models [9, 19, 66]. Additionally, we include non-instructional editing models [10, 20, 41] for completeness, although these are not fair baselines since they are not end-to-end and rely on additional conditions, such as depth maps or intermediate captions.

We randomly sample 3 seeds for both our method and baseline methods. In our experiments, we use different classifier-free guidance scores (g_T and g_V in Sec. 3.2) for different skills. Specifically, we set $g_T = 14.5$ and $g_V = 1.5$ for grounding and reasoning segmentation, while for other editing skills, we use $g_T = 10.5$ and $g_V = 2.0$.

For baseline methods, we adopt their default settings (e.g., diffusion steps, guidance scores, frame numbers) as provided in their GitHub repositories. To ensure fair evaluation, we sample the same eight frames from each method's video editing results.

For alignment and smoothness metrics, we use CLIP-B/32 to measure text-image and image-image similarity, averaging across all frames to obtain video-level scores. For detection metrics, we use GroundingDINO (Swin-T OGC) to detect target objects frame by frame, averaging confidence scores across all frames for the final video-level metric.

For the removal task, where fewer detected objects and lower alignment with the original text prompt are desired, we compute alignment and detection metrics as $1 -$ value.

We compare the model judged best for each video sample. The agreement between human and MLLM judgments is 0.74, whereas the agreement between human and CLIP is only 0.45. We conducted 5 times of the MLLM evaluation and took an average.

Methods	Grounding			Reasoning		
	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
Segmentation Models						
HTR [50]	47.11	47.60	47.35	20.01	28.02	24.01
VideoLISA [1]	53.23	54.37	53.80	38.48	39.20	38.84
MoRA [12]	57.73	53.63	55.68	38.92	37.48	40.36
Generative Editing Models						
InstructDiff [19]	19.88	12.81	16.35	14.02	8.07	11.05
InsV2V [9]	13.89	17.37	15.63	16.89	10.45	13.67
VEGGIE (Ours)	**37.74**	**21.83**	**29.79**	**22.53**	**15.97**	**19.25**

Table 4. Comparison of video concept grounding and reasoning segmentation tasks with other instructional generative models and expert segmentation models.

6.2. Data Collection Details

As mentioned in the earlier Sec. 3.3, beyond collecting existing data, we proposed a novel data synthesis pipeline to generate instructional video data by animating images in the instructional image dataset.

Specifically, we first select images from Omni-Edit [64], an instructional image editing dataset with carefully designed tasks/skills.

We first use QWen2-VL [61] to caption the original image and give an animation prompt to animate the image via CogVideX1.5-I2V [74]. Please refer Tab. 5 and Tab. 6 to our prompt for caption and animation. After getting the animated video, we utilize AnyV2V [35] to edit the video based on the reference image (edited image from image dataset). The reference image gives a strong prior to maintaining the image dataset's high-quality edit and thus transfer it to the video via the video editing model.

Next, we filter out videos by evaluating VBench metrics [30], including aesthetic quality, motion smoothness, image quality, subject consistency, and background consistency. We set thresholds at 0.6 for aesthetic quality, 65 for imaging quality, 0.9 for motion smoothness, subject consistency, and background consistency. We provide our generated data visualization in Fig. 9.

6.3. More Quantative Results & Discussion

Video Concept Grounding & Reasoning Segmentation We include additional results on video concept grounding and reasoning segmentation in Tab. 4. VEGGIE outperforms the diffusion-based baseline by a significant margin, showcasing its superior ability to accurately locate fine-grained object references and handle complex reasoning tasks. We hypothesize that through grounded generation, VEGGIE demonstrates remarkable precision in concept editing. For example, as shown in Fig. 11 in the Appendix, VEGGIE can remove the woman without altering the nearby girl.

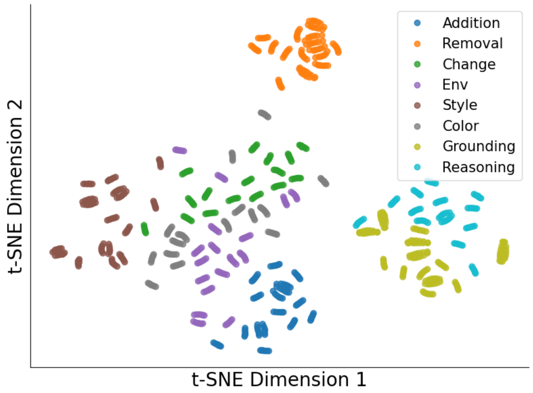


Figure 8. t-SNE Visualization of different task query distribution. Different colors represent different tasks/skills. Best view in color.

6.4. Limitation and Future Works

Our current method, VEGGIE, is built upon Stable-Diffusion 1.5, which inevitably constrains its editing quality compared to cutting-edge video generation models that rely on DiT or flow-based architectures. In addition, the video outputs we produce are relatively short, lagging behind some recent state-of-the-art methods in terms of length and temporal consistency. Furthermore, we observe increased editing artifacts when incorporating large amounts of grounding data, suggesting that multi-task data mixture strategies play a key role in maintaining high-quality edits.

Despite these limitations, our results demonstrate promising directions for improvement in terms of model design, data curation, and evaluation. Future work could explore integrating more advanced base architectures (e.g., DiT [34, 74] or flow-based models), extending the maximum video duration, developing more systematic data [28] with more advanced method [46] and carefully designed mixture strategies to balance fidelity and flexibility, and conducting scalable training. We hope our findings will inspire further research into these directions, pushing the boundaries of instructional video editing performance.

Task Query Visualization & Analysis via t-SNE. To analyze task/skill correlations, we project their grounded queries into lower-dimensional spaces using PCA and t-SNE. As shown in Fig. 8, distinct clusters form for each category (e.g., Addition), indicating effective differentiation by the model. *Reasoning* and *Grounding* appear together on the right. It may be because they both require cognitive/semantic understanding or logical reference. *Color*, *Env*, and *Change* clusters are closer to each other, indicating that the model views them as similar operations focusing on changing different visual attributes. *Style* lies in the lower-left region but remains relatively close to *Color*, *Env*, and *Change*. This proximity may reflect that "stylization" is conceptually similar to these visual attribute tasks, although it targets different

Table 5. Qwen2-VL prompt for Image caption.

Please describe this image shortly, try to capture main details in the image.
Here are some examples of image caption styles:

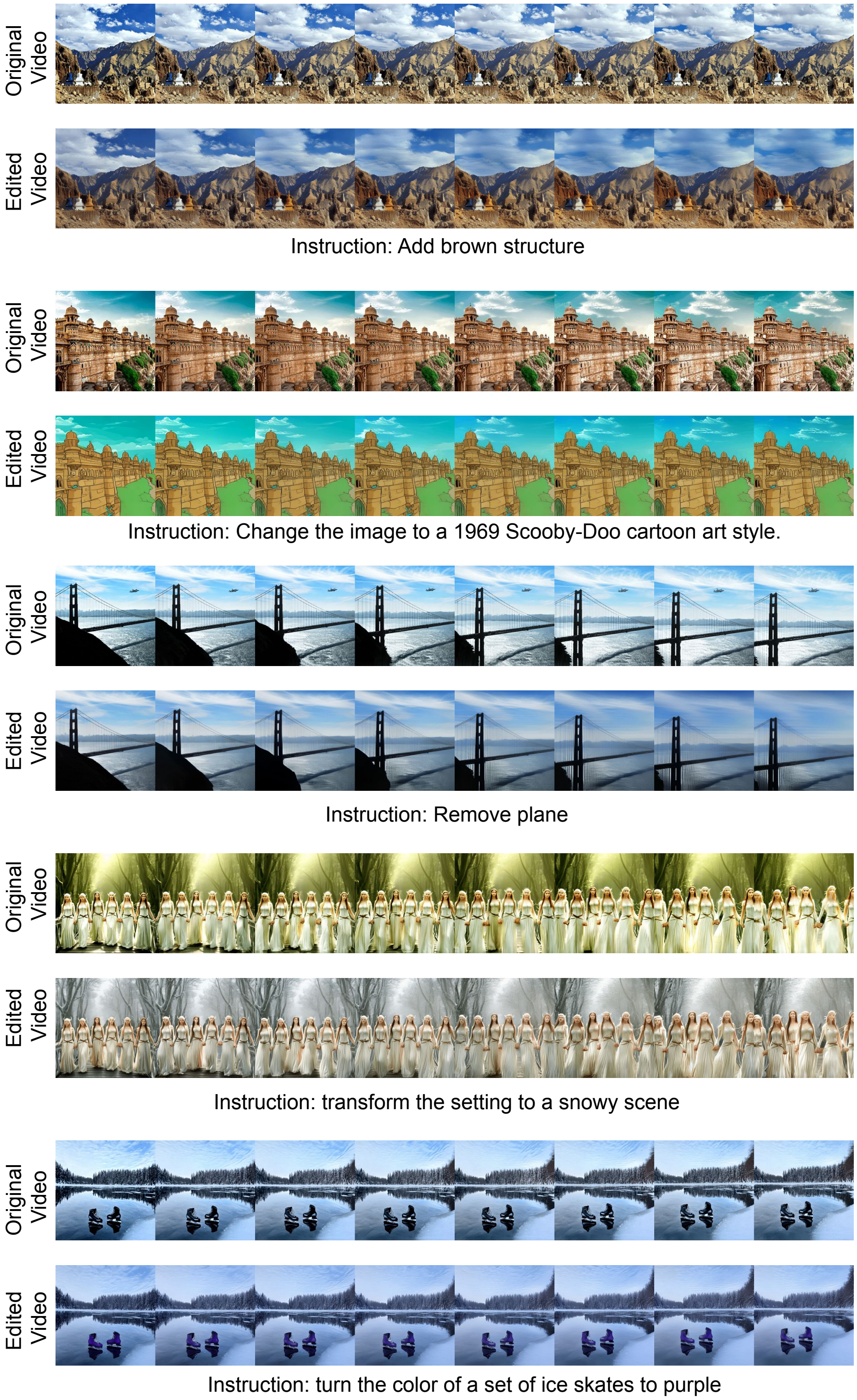
1. A Couple In A Public Display Of Affection
2. A kitten turning its head on a wooden floor
3. An Old Man Doing Exercises For The Body And Mind
4. Man Walking

Now, please describe the given image briefly in one sentence, please do not say something like 'The image shows...' or 'The image depicts...'.

transformations. *Removal* stands apart on the top, especially distant from *Addition*, indicating the model perceives them as distinct rather than inverse operations. In contrast, *Addition* lies closer to tasks like *Reasoning* and *Grounding*. It suggests that the act of adding elements may rely on similar semantic or referential processes (e.g., deciding what to add and how to reference the newly added element).

6.5. Extra Visualization

We provide extra visualization in Figs. 10 to 16



Instruction: Add brown structure

Instruction: Change the image to a 1969 Scooby-Doo cartoon art style.

Instruction: Remove plane

Instruction: transform the setting to a snowy scene

Instruction: turn the color of a set of ice skates to purple

Figure 9. Examples of our generated instructional video editing data.

Table 6. Qwen2-VL prompt for generating animation prompt.

I want to animate this image using an Image-Text-to-Video model. Your task is to generate a detailed and reasonable text prompt that describes how the image should be animated.

Guidelines:

1. Clarity & Realism - The animation description should be logical based on the given image, ensuring the movement makes sense for the scene.

2. Short & Vivid Description - Use expressive language to guide the animation model effectively, ensuring high-quality and visually engaging results.

Ensure that your animation prompt aligns with the content of the provided image and describes a visually compelling motion sequence.

Do not output animation prompts that contain objects/scenes not included in the given image.

Make sure the prompt is short in 1-2 sentences.

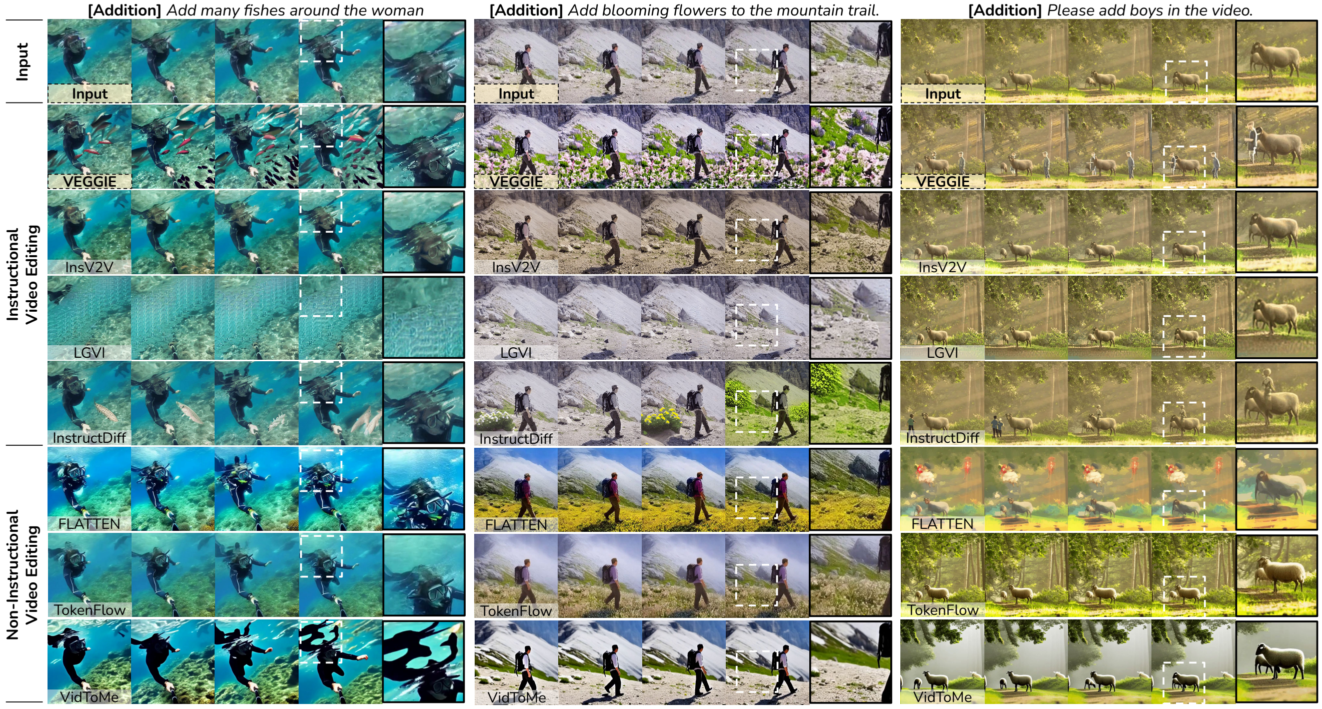

[Addition] *Add many fishes around the woman*
[Addition] *Add blooming flowers to the mountain trail.*
[Addition] *Please add boys in the video.*
Input
Instructional Video Editing
Non-Instructional Video Editing
Input
VEGGIE
InsV2V
LGVI
InstructDiff
FLATTEN
TokenFlow
VidToMe

Figure 10. More Examples of Concept Addition.

Table 7. GPT-4o prompt for MLLM-as-a-Judge for automatic instructional video editing evaluation.

User

You are an evaluator for instructional video editing tasks. Your job is to assess how well the edited video fulfills the user's specific instructions.

I will provide:
1. The original video (first GIF)
2. The edited video (second GIF)
3. The user's instruction: [user instruction]

Please evaluate the editing result using the following format:
INSTRUCTION: [Repeat the user's instruction]
EVALUATION:
- Accuracy score (1-10): [Your score]
- Quality score (1-10): [Your score]
- Appropriateness score (1-10): [Your score]
- Overall score (1-10): [Your final score]

EXPLANATION: [Provide a brief justification for your scores, highlighting specific strengths and weaknesses of the edit]
RECOMMENDATION: [Optional suggestions for improvement]

When scoring, consider:
- Accuracy: Does the edit precisely follow the given instruction? - Quality: Is the edit visually seamless and natural-looking? - Appropriateness: Does the edit maintain coherence with the original video context?

The overall scale is:
1-3: Poor - Major issues with the edit
4-6: Acceptable - Follows instruction but with noticeable flaws
7-8: Good - Clear, effective edit with minor issues
9-10: Excellent - Flawless execution of the instruction

Assistant

Scores, Explanation, Recommendation

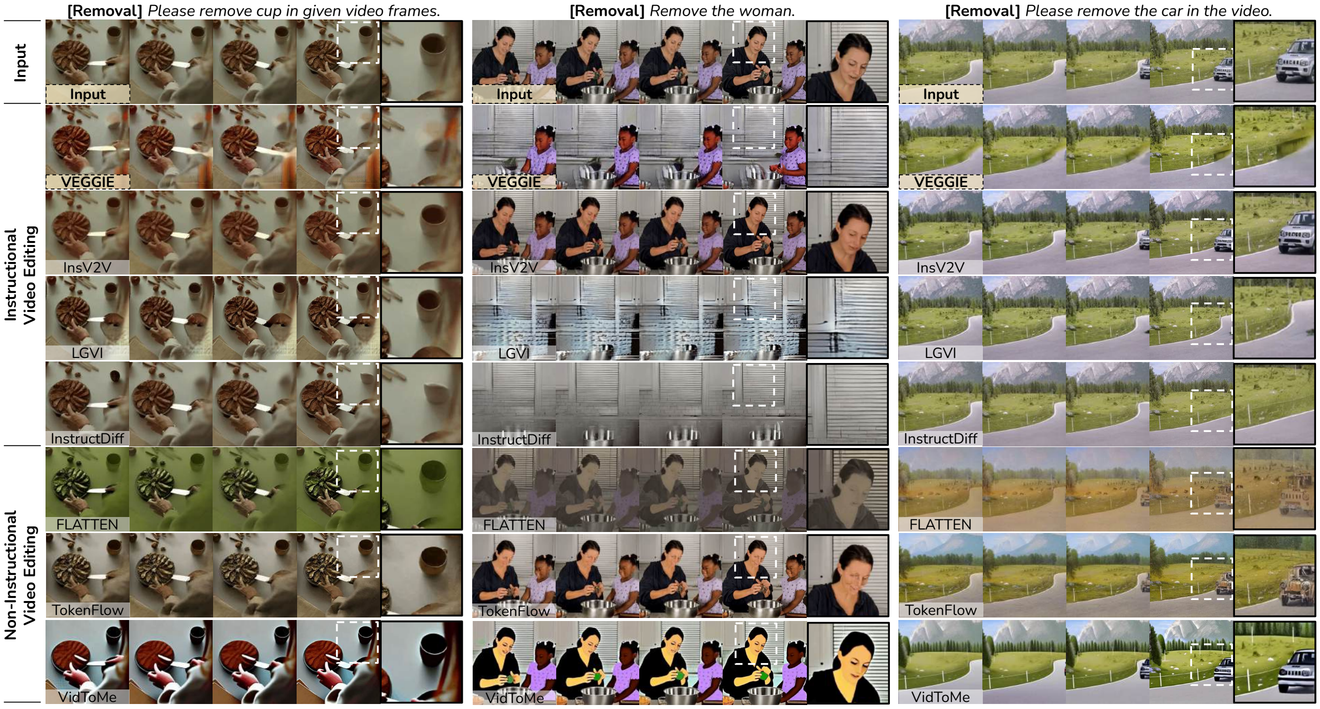

[Removal] Please remove cup in given video frames.
[Removal] Remove the woman.
[Removal] Please remove the car in the video.
Input
Instructional Video Editing
Non-Instructional Video Editing
Input
VEGGIE
InsV2V
LGVI
InstructDiff
FLATTEN
TokenFlow
VidToMe

Figure 11. More Examples of Concept Removal.

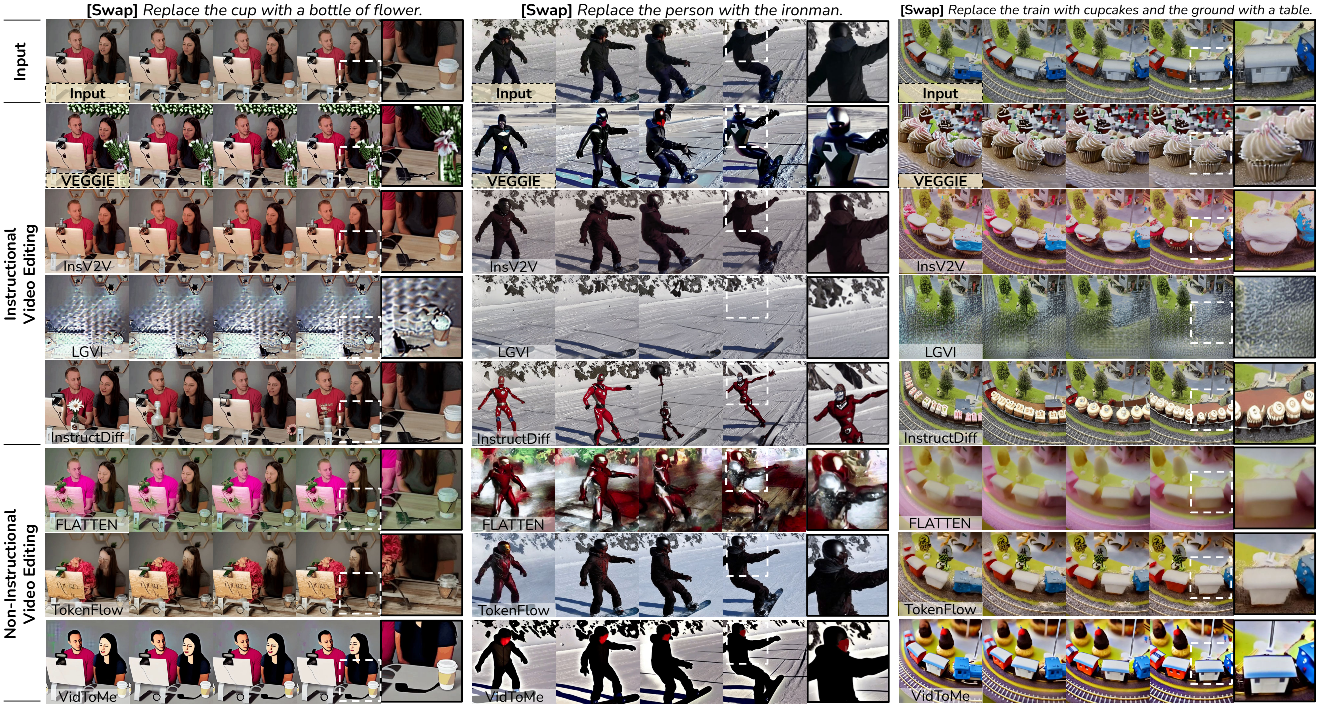

[Swap] Replace the cup with a bottle of flower.
[Swap] Replace the person with the ironman.
[Swap] Replace the train with cupcakes and the ground with a table.
Input
Instructional Video Editing
Non-Instructional Video Editing
Input
VEGGIE
InsV2V
LGVI
InstructDiff
FLATTEN
TokenFlow
VidToMe

Figure 12. More Examples of Object Changes.

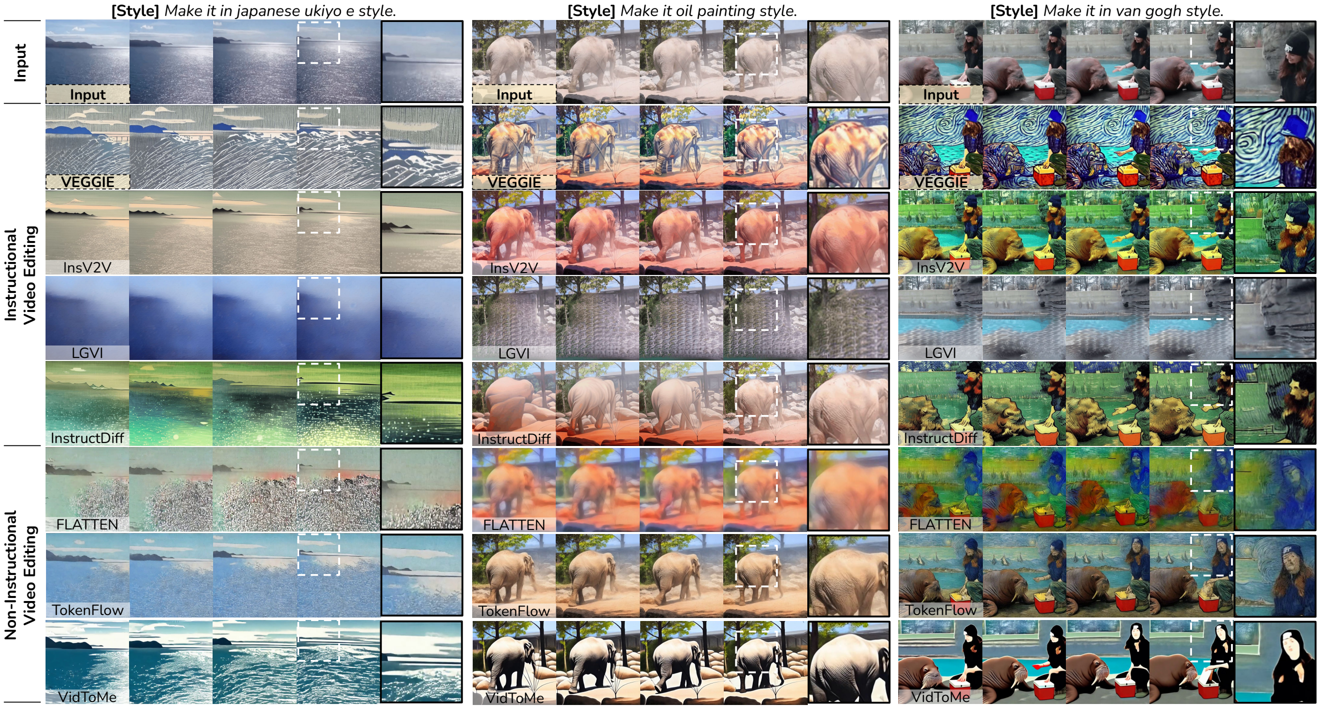

[Style] Make it in japanese ukiyo e style.
[Style] Make it oil painting style.
[Style] Make it in van gogh style.
Input
Instructional Video Editing
Non-Instructional Video Editing
Input
VEGGIE
InsV2V
LGVI
InstructDiff
FLATTEN
TokenFlow
VidToMe

Figure 13. More Examples of Stylization.

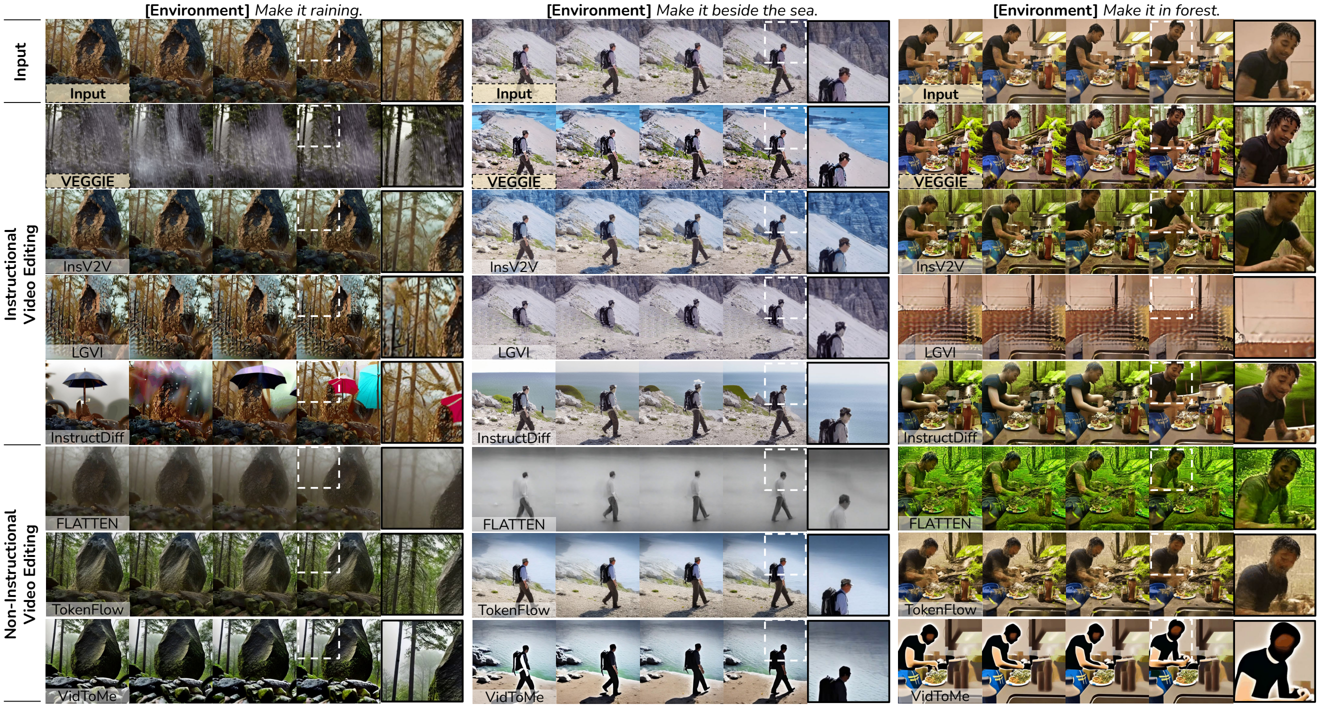

[Environment] Make it raining.
[Environment] Make it beside the sea.
[Environment] Make it in forest.
Input
Instructional Video Editing
Non-Instructional Video Editing
Input
VEGGIE
InsV2V
LGVI
InstructDiff
FLATTEN
TokenFlow
VidToMe

Figure 14. More Examples of Environment and Background Editing.

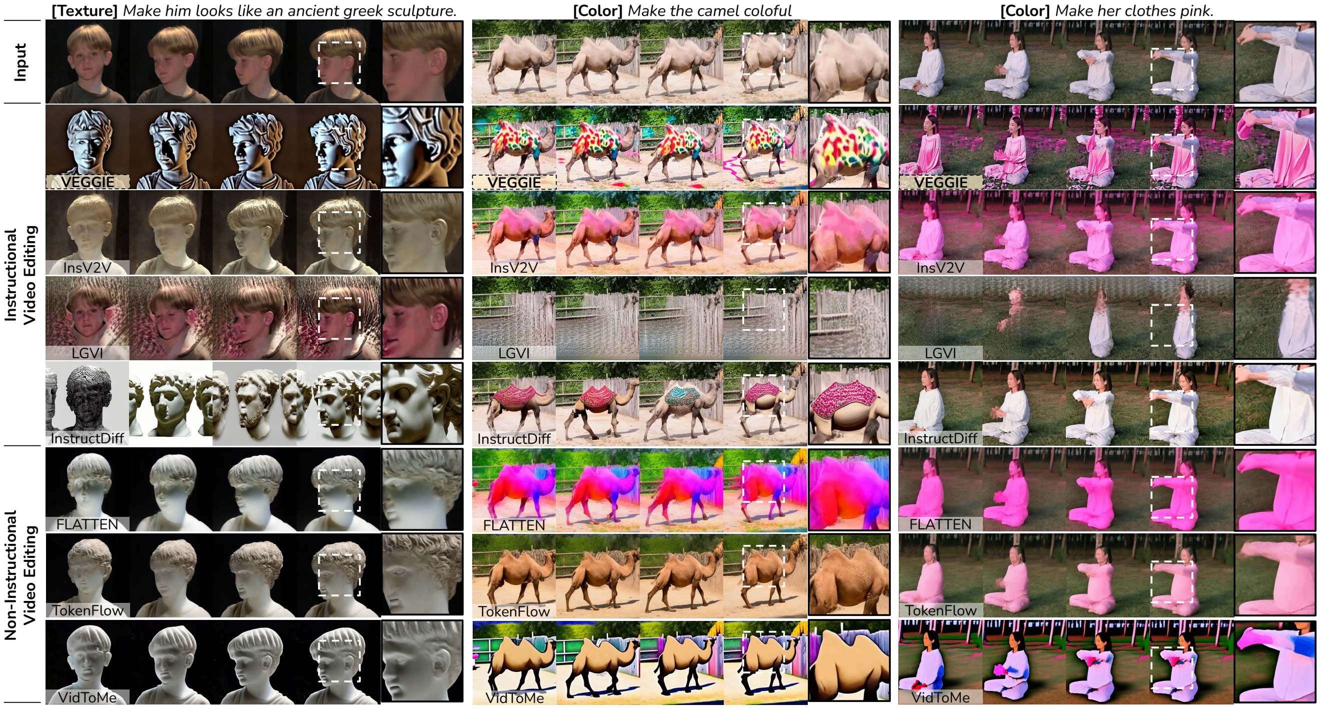

[Texture] Make him looks like an ancient greek sculpture.
[Color] Make the camel coloful
[Color] Make her clothes pink.
Input
Instructional Video Editing
Non-Instructional Video Editing
VEGGIE
InsV2V
LGVI
InstructDiff
FLATTEN
TokenFlow
VidToMe

Figure 15. More Examples of Visual Features Editing.

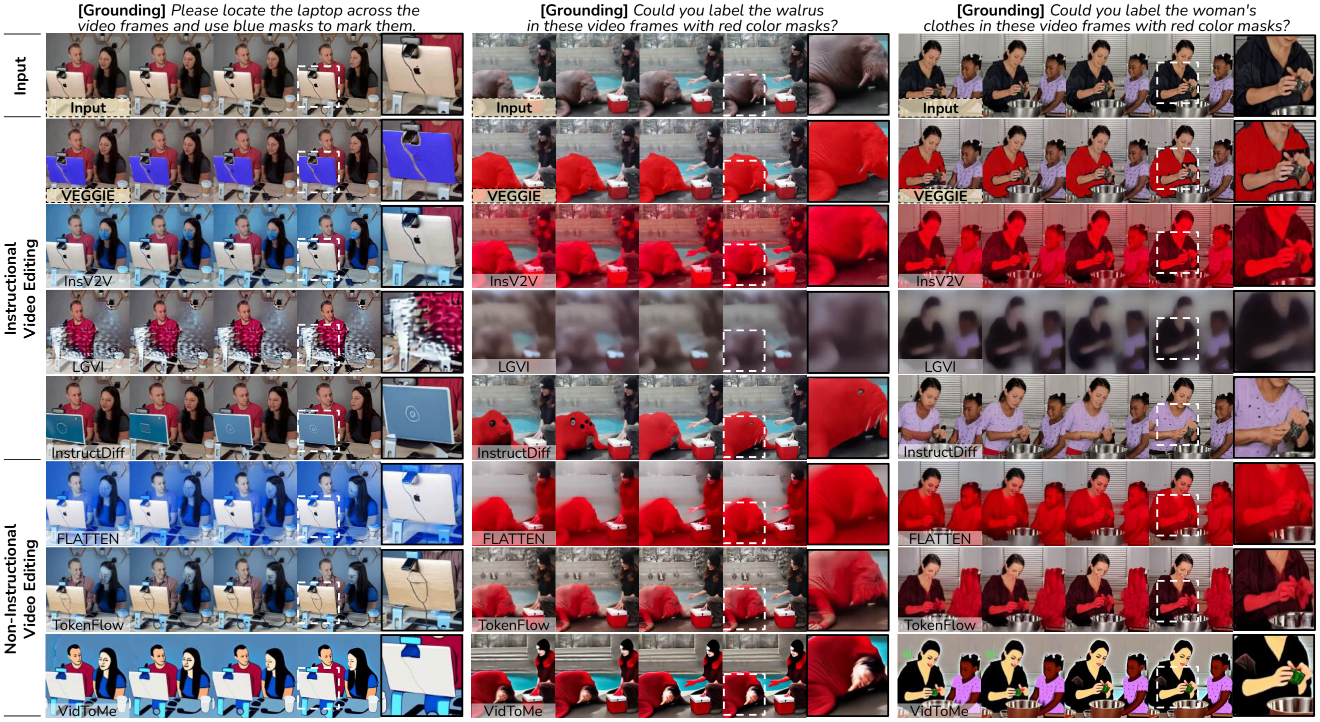

[Grounding] Please locate the laptop across the video frames and use blue masks to mark them.
[Grounding] Could you label the walrus in these video frames with red color masks?
[Grounding] Could you label the woman's clothes in these video frames with red color masks?
Input
Instructional Video Editing
Non-Instructional Video Editing
Input
VEGGIE
InsV2V
LGVI
InstructDiff
FLATTEN
TokenFlow
VidToMe

Figure 16. More Examples of Object Grounding.

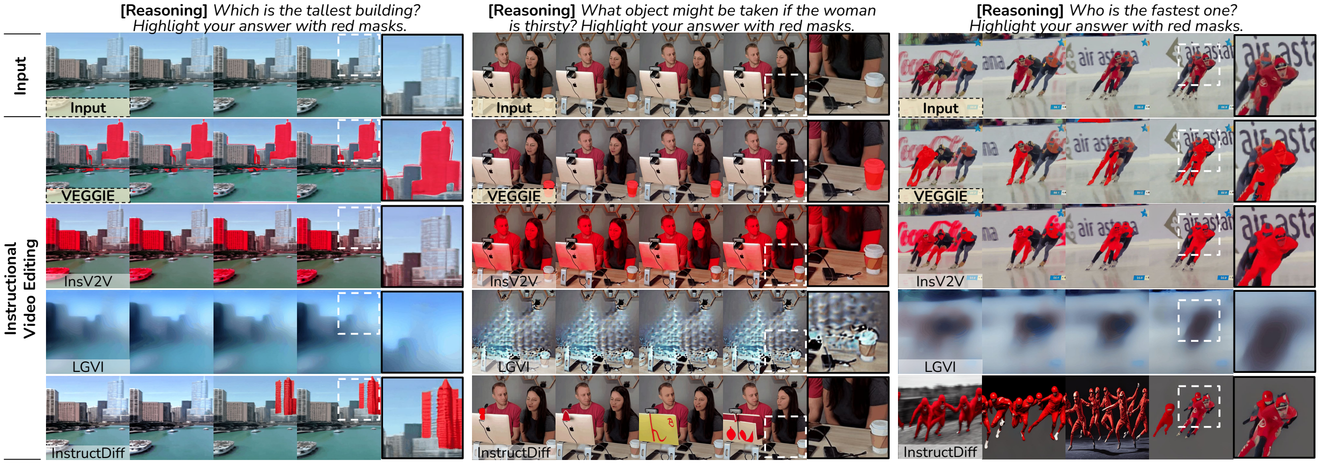

[Reasoning] *Which is the tallest building? Highlight your answer with red masks.*
[Reasoning] *What object might be taken if the woman is thirsty? Highlight your answer with red masks.*
[Reasoning] *Who is the fastest one? Highlight your answer with red masks.*
Input
Instructional Video Editing
Input
VEGGIE
InsV2V
LGVI
InstructDiff

Figure 17. More Examples of Object Reasoning Segmentation.