清華大学 交叉信息研究院
Institute for Interdisciplinary Information Sciences, Tsinghua University

CLEVAM-DM: Consistent Local Edits in Videos via Attention Manipulation in Diffusion Models

Liu Hanzuo[1,*] Liu Kehan[1,*] Wang Zehua[1,*]
[1]IIIS, Tsinghua University

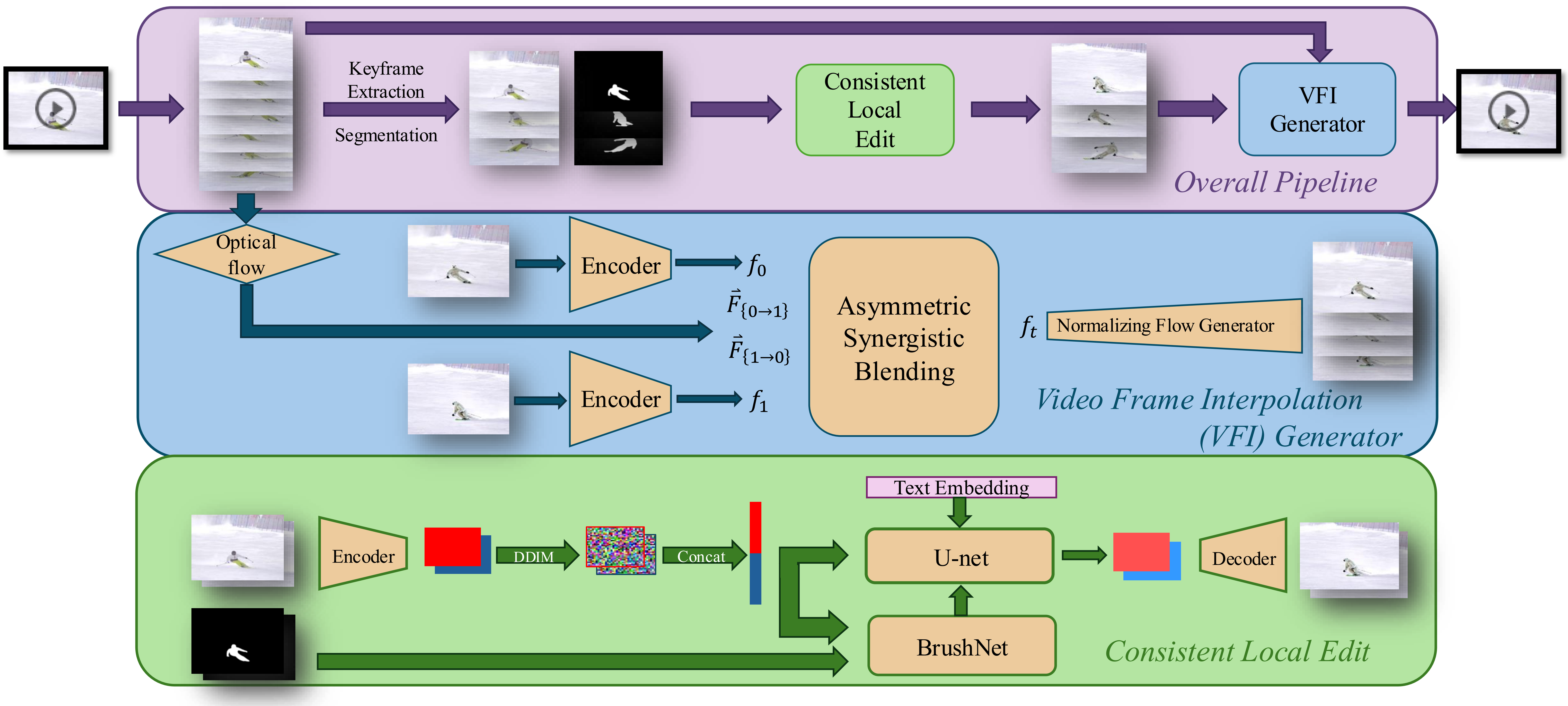

Figure 1: Overview of the CLEVAM-DM workflow. **Top:** The overall pipeline extracts keyframes, applies consistent local edits, and synthesizes intermediate frames. **Middle:** Video frame interpolation module utilizes the optical flow of the original video, and generates smooth transitions between edited keyframes. **Bottom:** Consistent local editing is achieved via inpainting and attention sharing in diffusion models.

Part I

Introduction

Edit videos with precision and consistency!

- **Problem:** Local video edits often cause flickering and inconsistency.
- **Goal:** Make local changes to video frames that look natural and stay consistent over time.
- **Our Solution:** **CLEVAM-DM**—a pipeline that combines advanced inpainting, attention manipulation, and smart frame interpolation.

Why is this important?

- Achieve high-quality, temporally consistent edits
- Works out of the box—no extra training needed!
- Fast and efficient—runs on a single GPU and less memory limit

Part II

Methodology

Our framework operates through three sequential stages: keyframe preparation (1), consistent local edit (2 and 3), and video frame interpolation (4).

1. Keyframe Extraction and Mask Inference

- Strategic extraction of representative keyframes from the source video
- Text-guided object segmentation based on user prompts

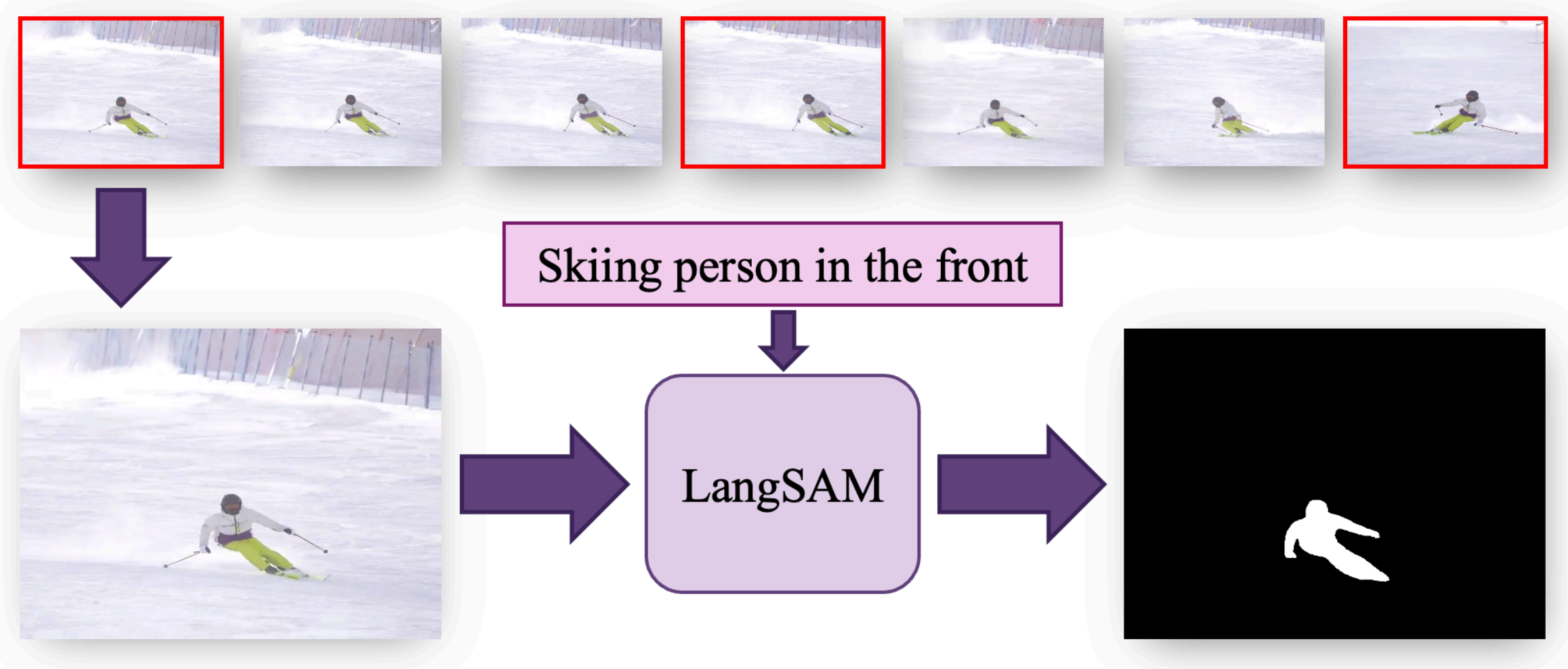

Figure 2: Keyframe segmentation (frames outlined in red) and mask inference process.

2. Local Edits through Inpainting

- Application of BrushNet [2], an inpainting model built on pretrained diffusion models, for precise local keyframe editing. It removes text cross-attention from the U-Net and instead inputs a concatenation of noisy latents, masked image latents, and a resized mask.
- Use of DDIM-inverted latent as the initial input to the unet. This helps preserve information in the original image and improves alignment.
- The use of BrushNet and DDIM-inverted latents offers a 'free lunch'—allowing edits to either the foreground object or the background simply by inverting the mask.

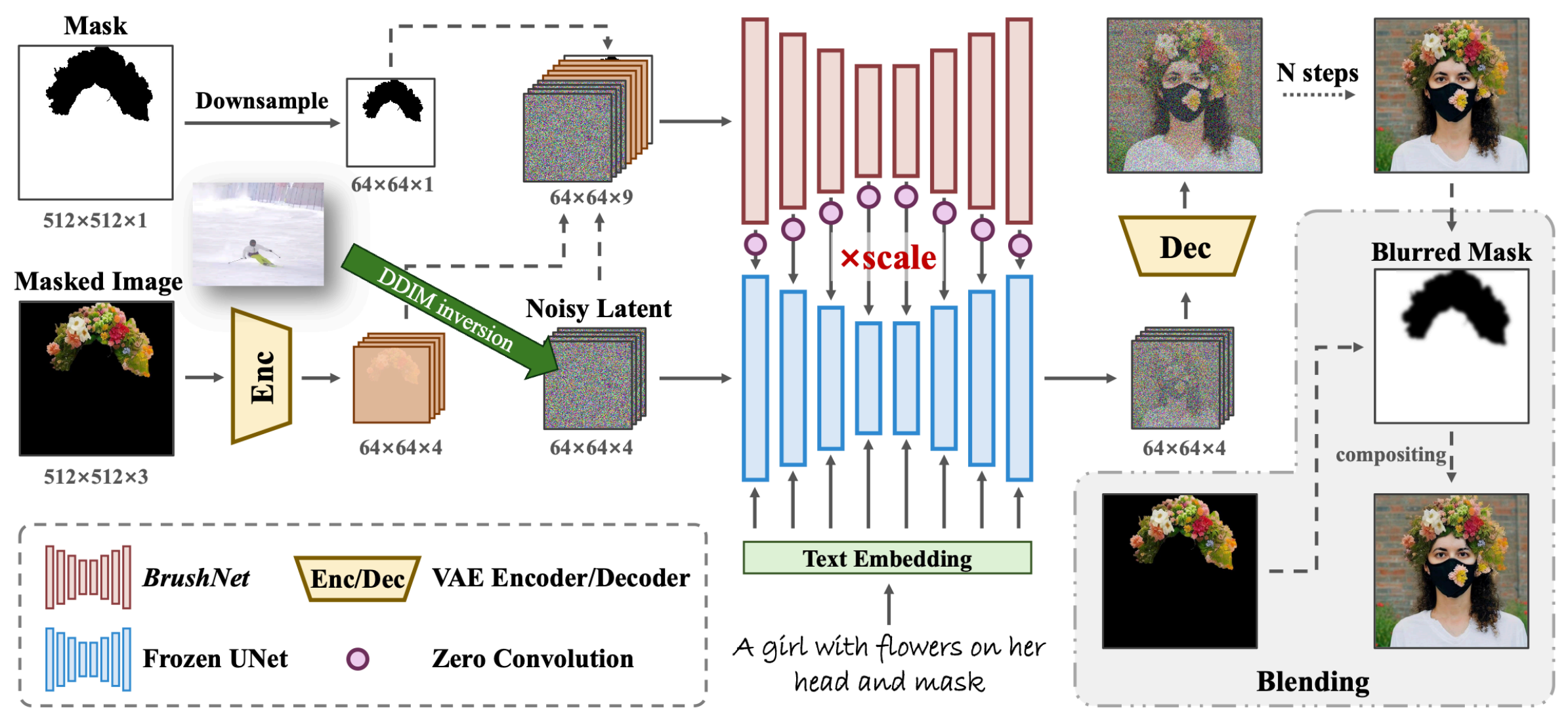

Figure 3: BrushNet model architecture for local video editing. We use the DDIM inversion of the original video frames as initial latents.

3. Temporal Consistency Enhancement

We implement full attention sharing [1] across all attention layers in the BrushNet model to ensure temporal coherence:

Formally, let Q_i, K_i, and V_i be the queries, keys, and values projected from deep features ϕ_i of frame $\mathcal{I}_i$ in the sequence. The attention update for ϕ_i is given by:

$$\text{Attention}(Q_i, K_{1\ldots n}, V_{1\ldots n}),$$

where $K_{1\ldots n} = \begin{bmatrix} K_1 \\ K_2 \\ \vdots \\ K_n \end{bmatrix}$ and $V_{1\ldots n} = \begin{bmatrix} V_1 \\ V_2 \\ \vdots \\ V_n \end{bmatrix}$.

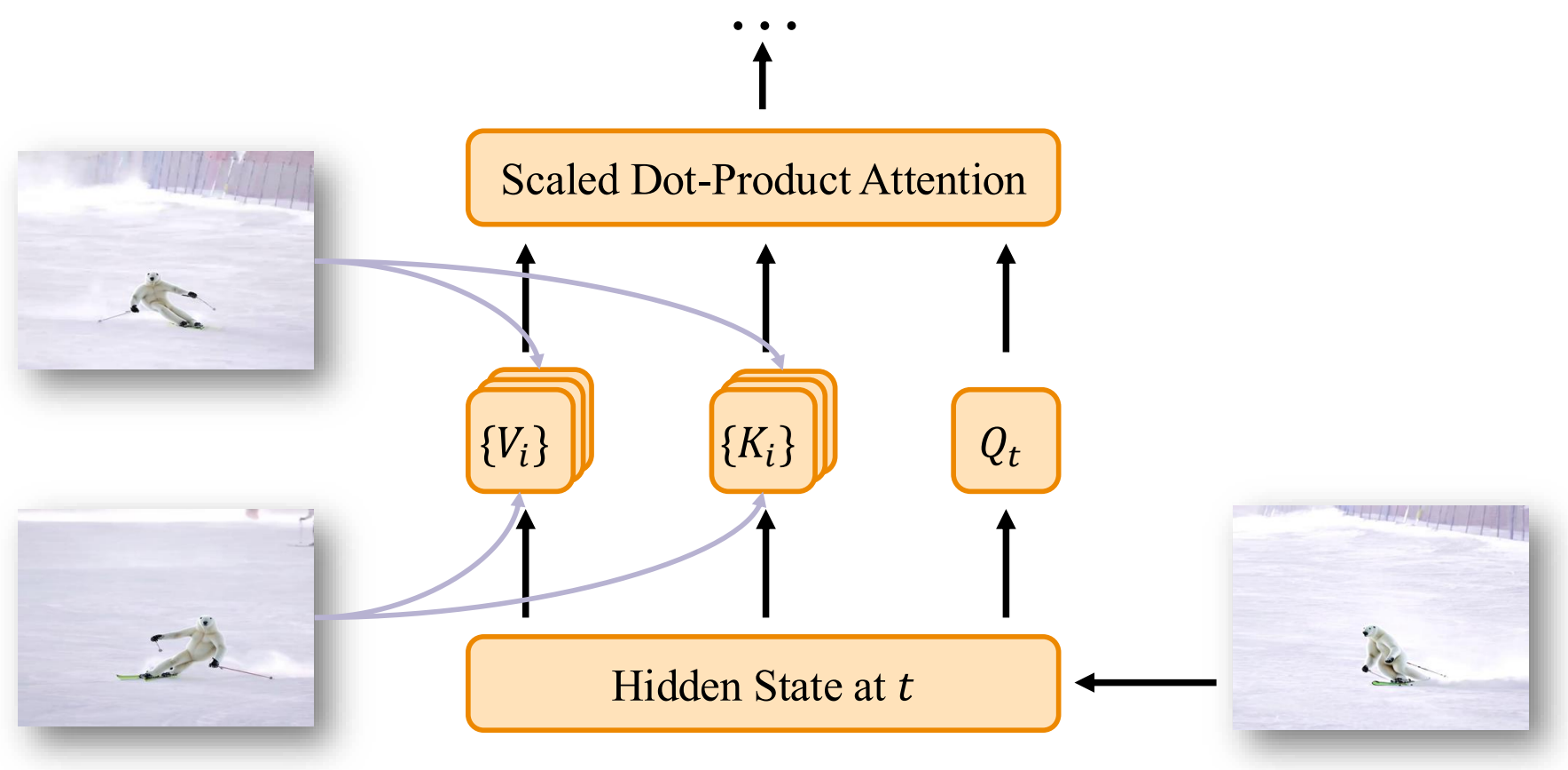

Figure 4: Temporal consistency enhancement through full attention sharing in the BrushNet model.

4. Video Frame Interpolation

Our approach combines multiple techniques for high-quality frame interpolation:

- Optical flow computation between processed keyframes
- Application of PerVFI [3] for intermediate frame synthesis
- Selective preservation of unedited regions to minimize artifacts via copying the unmasked region of the original frames
- Integration of spatial information from the original video to handle occlusions

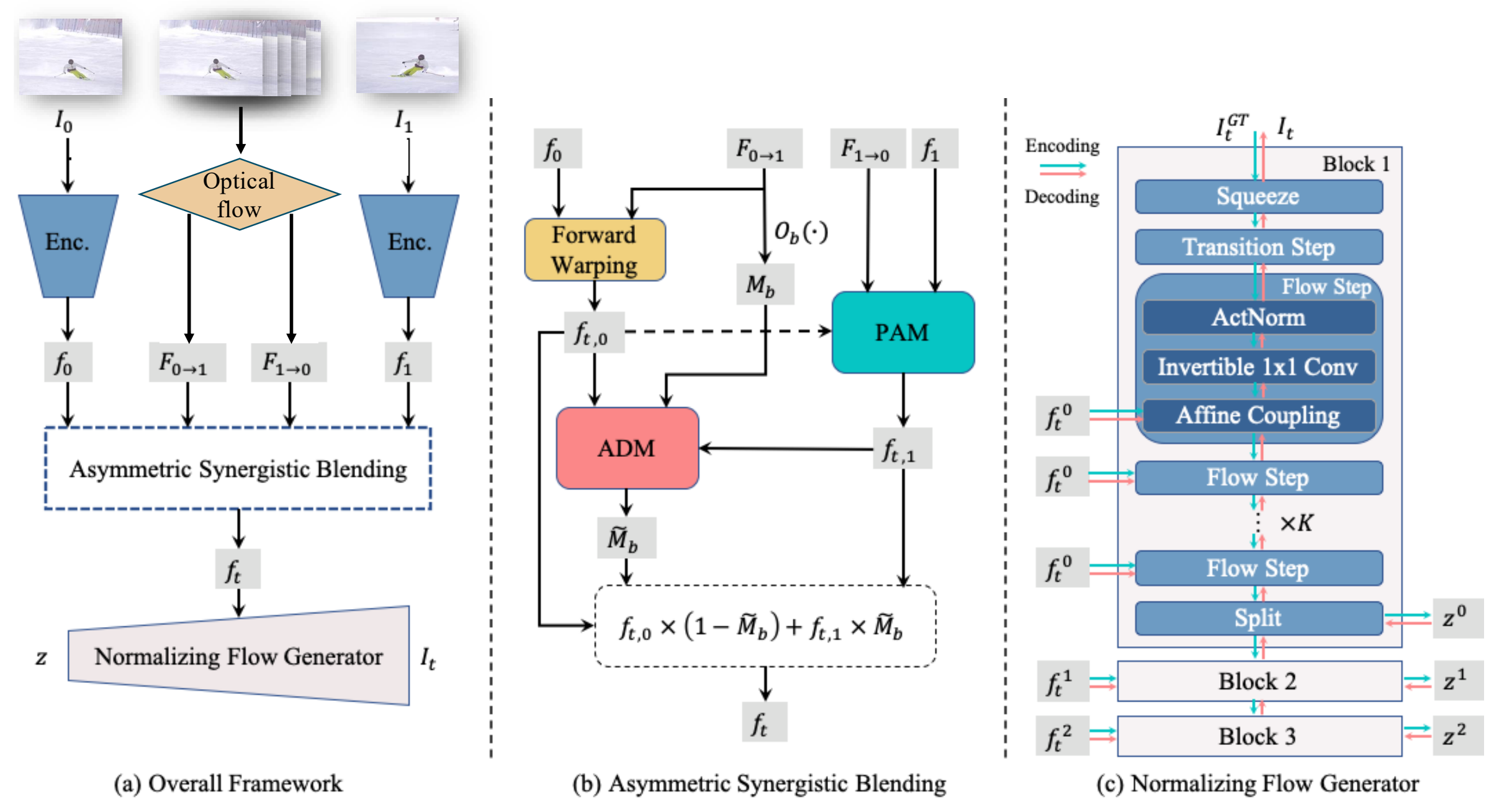

Figure 5: Frame interpolation process using optical flow from the original video and PerVFI.

Part III

Results

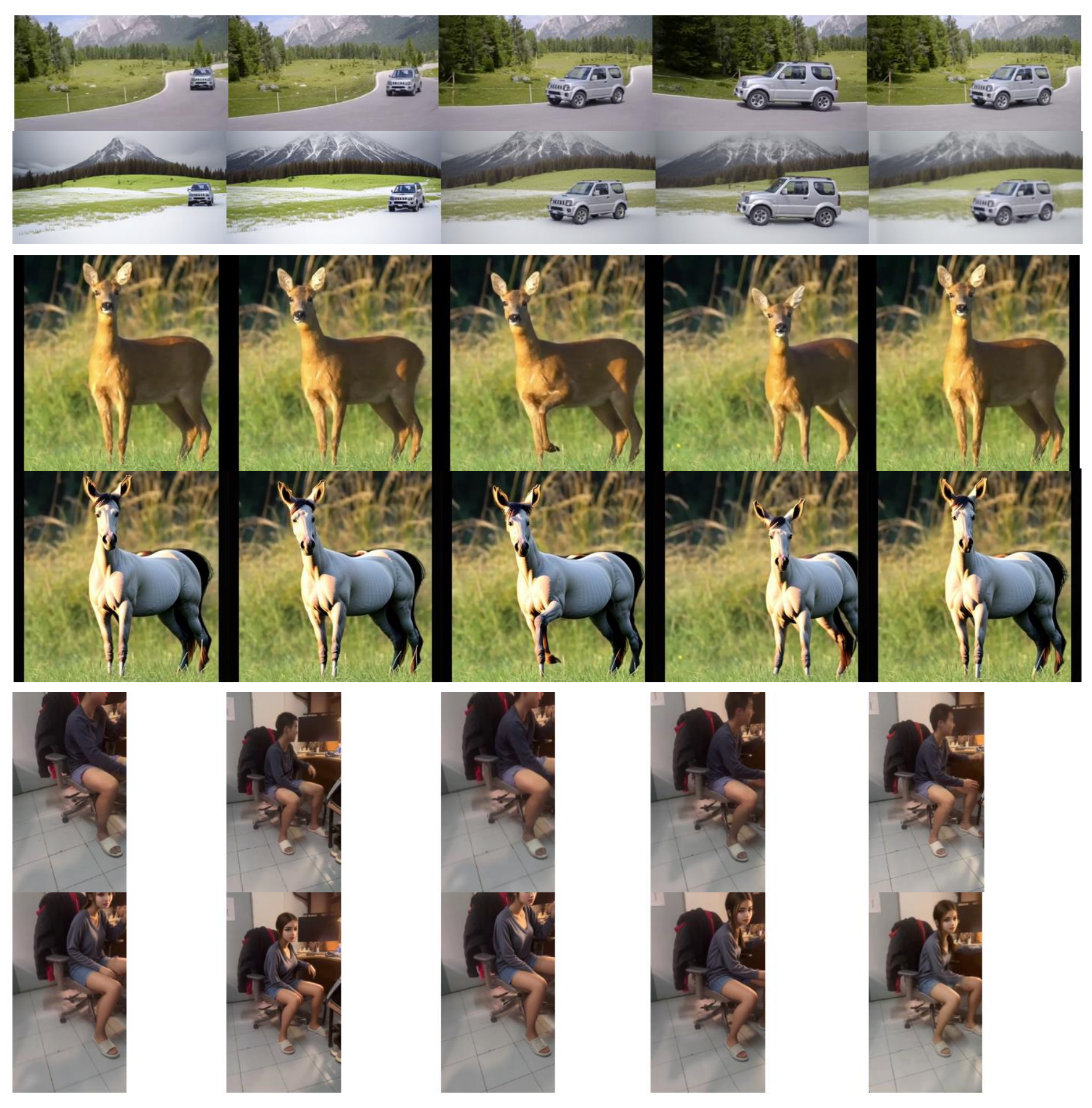

Figure 6: Results of CLEVAM-DM. The top rows are from the original video, and the bottom rows are from the processed video. The first four processed frames are keyframes, and the last column are interpolated frames. Shoutout to our roommate for making a guest appearance!

Interested in a hands-on demonstration? Reach out to us for access to our live Gradio demo (hopefully)!

References

[1] Amir Hertz, Andrey Voynov, Shlomi Fruchter, and Daniel Cohen-Or. Style aligned image generation via shared attention, 2024.

[2] Xuan Ju, Xian Liu, Xintao Wang, Yuxuan Bian, Ying Shan, and Qiang Xu. Brushnet: A plug-and-play image inpainting model with decomposed dual-branch diffusion. In *European Conference on Computer Vision*, pages 150–168. Springer, 2024.

[3] Guangyang Wu, Xin Tao, Changlin Li, Wenyi Wang, Xiaohong Liu, and Qingqing Zheng. Perception-oriented video frame interpolation via asymmetric blending. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2753–2762, 2024.