Appendix

In our supplementary materials, we present additional details and results of FlowV2V to provide deeper insights into the proposed method. The contents are organized according to the following structure:

- **Section A:** Implementation Details of the Flow Calibration Network.
- **Section B:** Implementation Details of DAVIS-EDIT.
- **Section C:** More Results.
- **Section D:** Impact of RAFT's Estimation Error.
- **Section E:** Efficiency Study.
- **Section F:** Limitations.

A. Implementation Details of Flow Calibration Network

This section details the implementation of the flow calibration network, covering its motivation, architecture, and training strategy.

Network Architecture. The flow calibration network is a critical component in the FlowV2V system, as its performance directly affects the accuracy of flow guidance in the DF-I2V module, thereby influencing the overall consistency of the edited video. The primary objective of this network is to remove redundant regions in the input optical flow maps and ensure that the refined flow maps remain consistent with the edited first frame, ultimately enhancing the quality of video editing results. To achieve this, we draw inspiration from ProPainter.

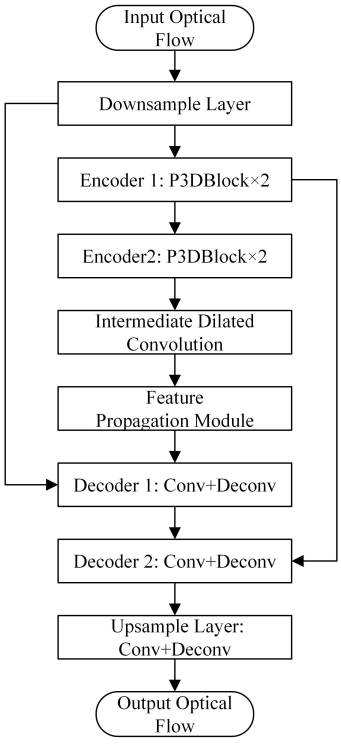

Figure 1. Structure of the network architecture.

Figure 2. More results under the facial image editing scenario.

Figure 3. More results under the video style transfer scenario.

Training Details. We train the flow calibration network on the YouTube-VOS dataset, whose training set contains 3,471 videos along with their corresponding segmentation mask annotations. We use RAFT to extract optical flow maps for our approach. After data preprocessing, the network is trained for 500k iterations with a batch size of 8. Specifically, in each training iteration, we randomly sample 10 optical flow frames and apply the random mask generation algorithm from flow-guided transformer. The AdamW optimizer is employed to update the model parameters, with the initial learning rate set to 10^{-4}. We implement our method using the PyTorch framework and train it on 4 H20-NVLink (96GB) GPUs.

B. More Results

In this section, we present additional qualitative results produced by FlowV2V, including face editing, style transfer, rotation, multi-object scenarios, instruction-based editing, text- and image-based editing.

C. Impact of RAFT's Estimation Error

In this section, we provide a detailed analysis of how the estimation errors of RAFT affect the performance of the FlowV2V model.

Textureless Regions. RAFT struggles in textureless regions where reliable motion cues are scarce. For example, when pale blue, nearly transparent tears slide down a face, the transparent liquid introduces almost no discernible texture and causes only minimal photometric variation against the background. As a result, RAFT fails to accurately estimate the optical flow in these regions.

Object Occlusion. RAFT encounters difficulties when objects become occluded or disappear between consecutive frames, as no subsequent visual information is available to support motion estimation. A typical case can be observed in the soccer

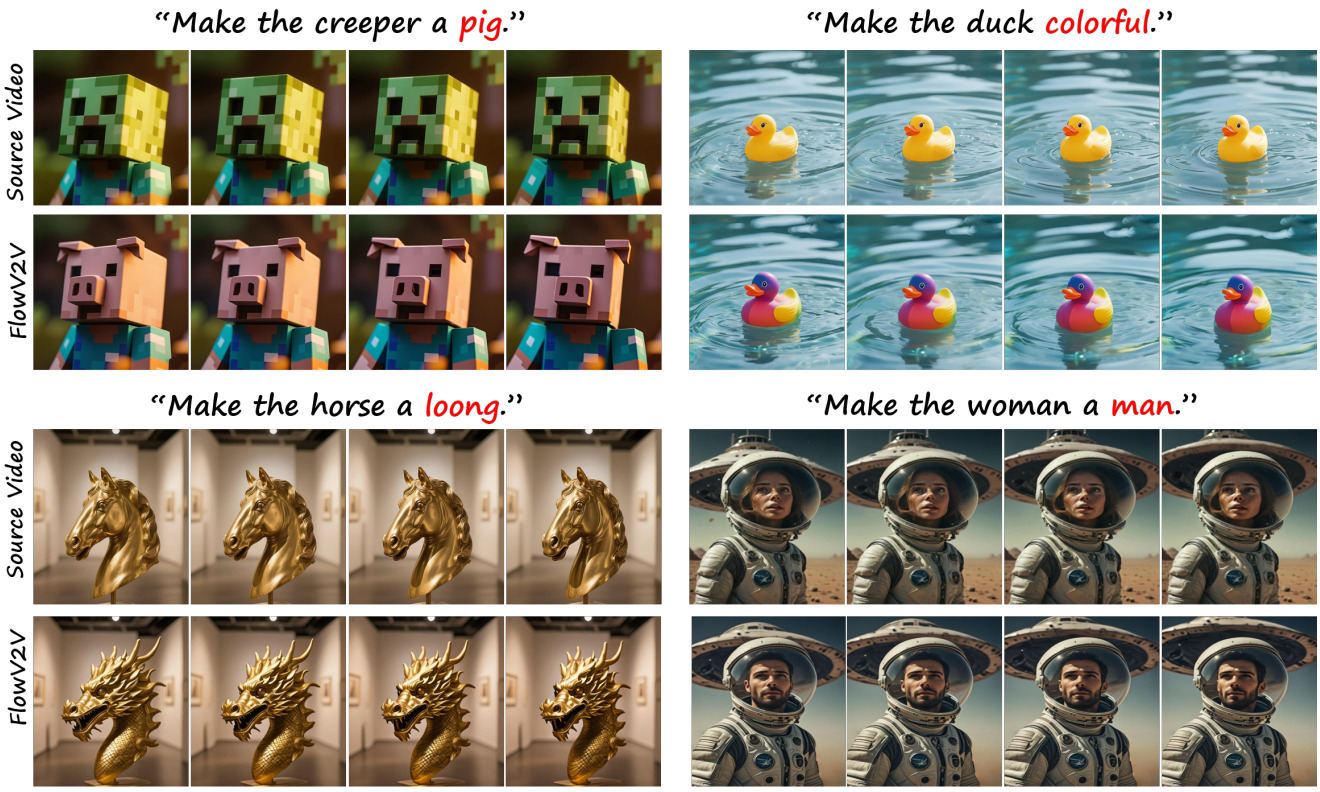

Figure 4. More results under the rotation-based editing scenario.

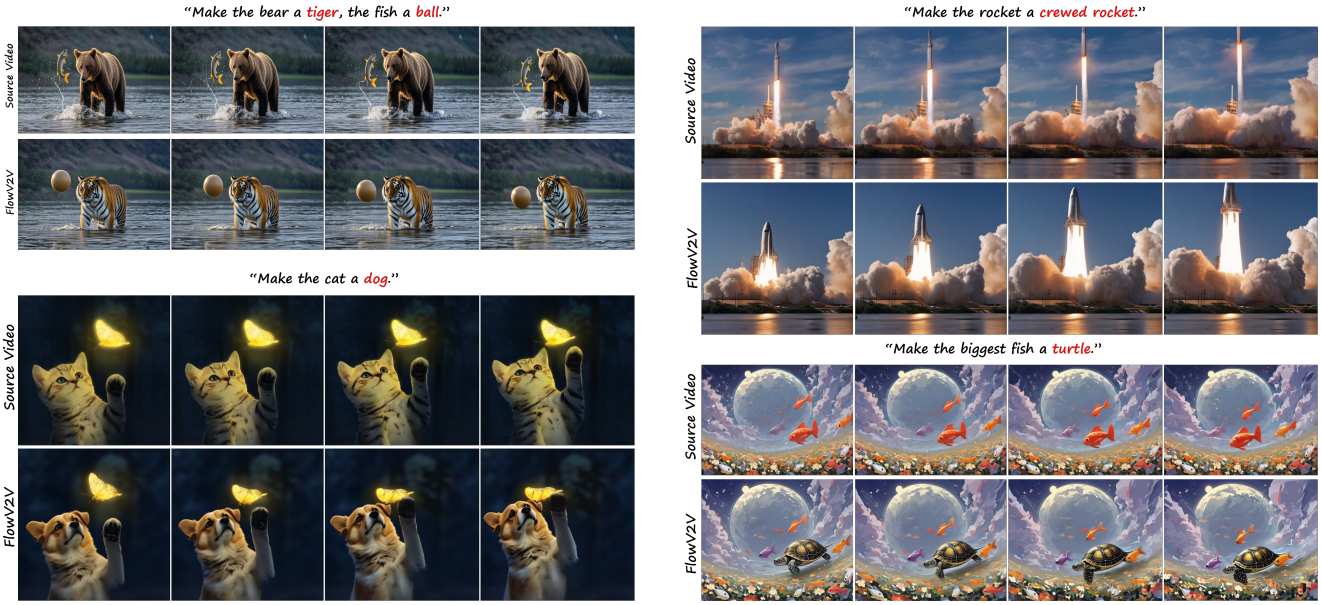

Figure 5. More results under the multi-object editing scenario.

ball example from the DAVIS-EDIT dataset, where the ball is occluded by trees, leading to inaccurate flow predictions and noticeable discrepancies between the generated and original videos.

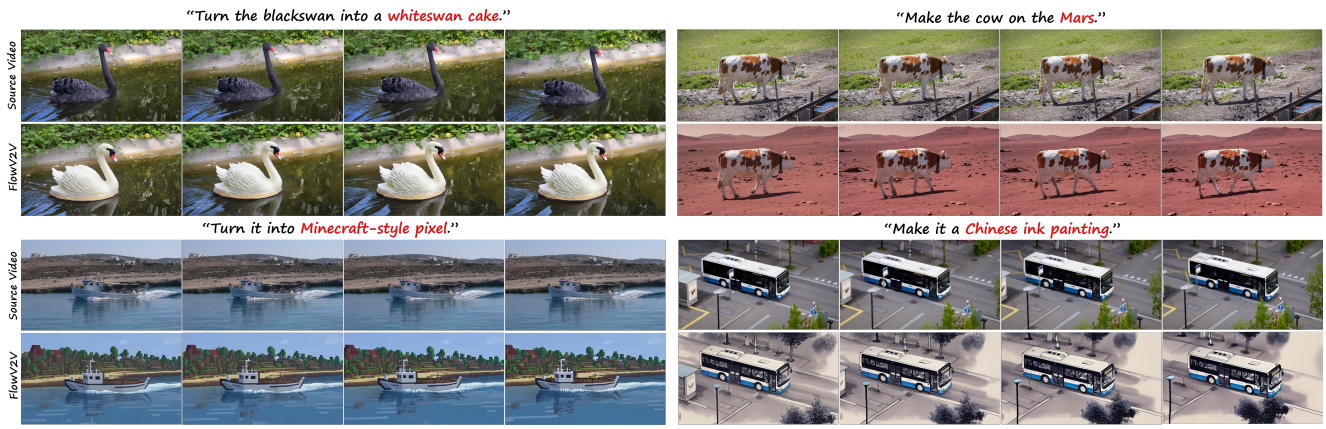

Figure 6. More results under the instruction-based editing scenario.

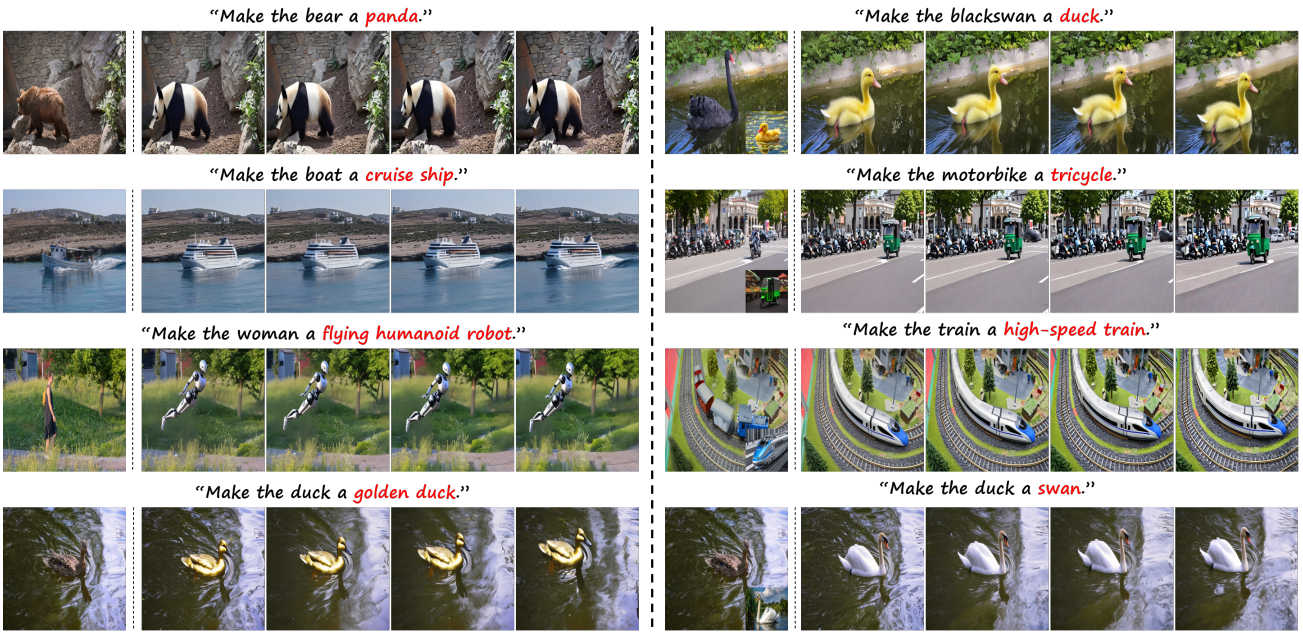

Figure 7. More results under the text-based (left) and image-based (right) editing scenarios.

Lighting Variation. RAFT is sensitive to significant variations in brightness or color between adjacent frames, which can substantially degrade feature matching accuracy. In particular, in face editing applications, lighting changes that occur as the subject's mouth closes often lead to unreliable optical flow estimation, resulting in noticeable temporal inconsistencies and reduced visual quality in the generated video.

D. Efficiency Study

In this section, we evaluate the average inference time of FlowV2V on the DAVIS-EDIT-C and DAVIS-EDIT-S datasets. The GPU used for these experiments was an H20-NVLink (96GB). As shown in Table 1, FlowV2V achieves efficient inference at a resolution of 512×512, completing a video editing example in only about 2.5 minutes, markedly outperforming other methods in terms of speed. However, when the resolution is increased to 854×480, corresponding to the original video size, the average inference time nearly doubles. This increase is primarily due to the higher computational cost associated with the larger image resolution.

E. Limitations

Despite the strong performance and versatile applications demonstrated by FlowV2V, its effectiveness in some editing cases remains limited, primarily due to inherent restrictions imposed by the optical flow estimation module and the generative capability of the employed I2V backbone.