A. Method for Optical Flow Rotation

We propose a method to edit using information on the direction in which pixels move in optical flow representing information on pixel motion between previous frame F_{t-1} and present frame F_t. Our proposed model receives one of eight directions from the user, including Northeast (NE), Southeast (SE), Southwest (SW), and Northwest (NW), which are made from a combination of four directions in the 2D coordinate system.

$$\begin{bmatrix} X' \\ Y' \end{bmatrix} = \begin{bmatrix} X\cos\theta_D & -Y\sin\theta_D \\ X\sin\theta_D & Y\cos\theta_D \end{bmatrix}, \qquad (1)$$

where θ_D denotes the angel between axis in frame and user provided direction D. X, Y denotes each motion vector in axis X, and Y. X', Y' denote the motion vectors rotated by θ in each axis.

After rotating the optical flow for the user-provided direction D, only the region of pixels with positive directional motion of the x axis in the rotated coordinate system is specified. The value of the motion map is extracted from the specific region and video edit is performed on the corresponding area.

B. Detailed Description of Evaluation Metrics

CLIP Score The CLIP score is calculated in the CLIP model, generating embedding vectors for input images and prompts. The CLIP score is measured by computing the cosine similarity between image and caption embedding. We measured how close the target prompt and the edited video frames are semantically in the CLIP Score. We measured the CLIP Score between target prompt and each edited video frame, and quantitatively compared the performance of our model and other models by the average of the scores measured per each frame. The CLIP Score is calculated with the following equation.

$$extCLIPScore(F_t, \mathcal{P}^*) = max(100 * cos(E_{F_t}, E_{\mathcal{P}^*}), 0), \qquad (2)$$

where F_t denote the t th edited frame, and $\mathcal{P}^*$ denotes the target prompt. We use official ViT-Base-Patch16 CLIP model.

Masked PSNR To evaluate whether our proposed model performs undesired edit out of target region to be edited, we measured Masked PSNR (M.PSNR) proposed by Video-P2P [14]. On this purpose, we measure how much the external region of the target region has changed from the frame of the original video. In consideration of the averaged attention mask sequence M of the changed object, we measure Masked PSNR by computing the pixel distance in the out-of-target regions of the edited video V^* and the input video V,

$$M.PSNR(V^*, V) = PSNR(B(V^*, M), B(V, M)), \qquad (3)$$

according to Video-P2P [14], $B(V, M) = V_M$ is defined as a reversed mask binary function, so only regions not to be changed are involved in measuring Masked PSNR.

C. Code Descriptions

Our code is based on PyTorch version of Video-P2P. We use Video-P2P to edit videos. We set the parameters as follows: frame_size_h = 512, frame_size_w = 512, number of frames = 4,

Code is available at `https://anonymous.4open.science/r/AttentionFlow-197C/README.md`