AttentionFlow: Text-to-Video Editing Using Motion Map Injection Module

Supplementary Material

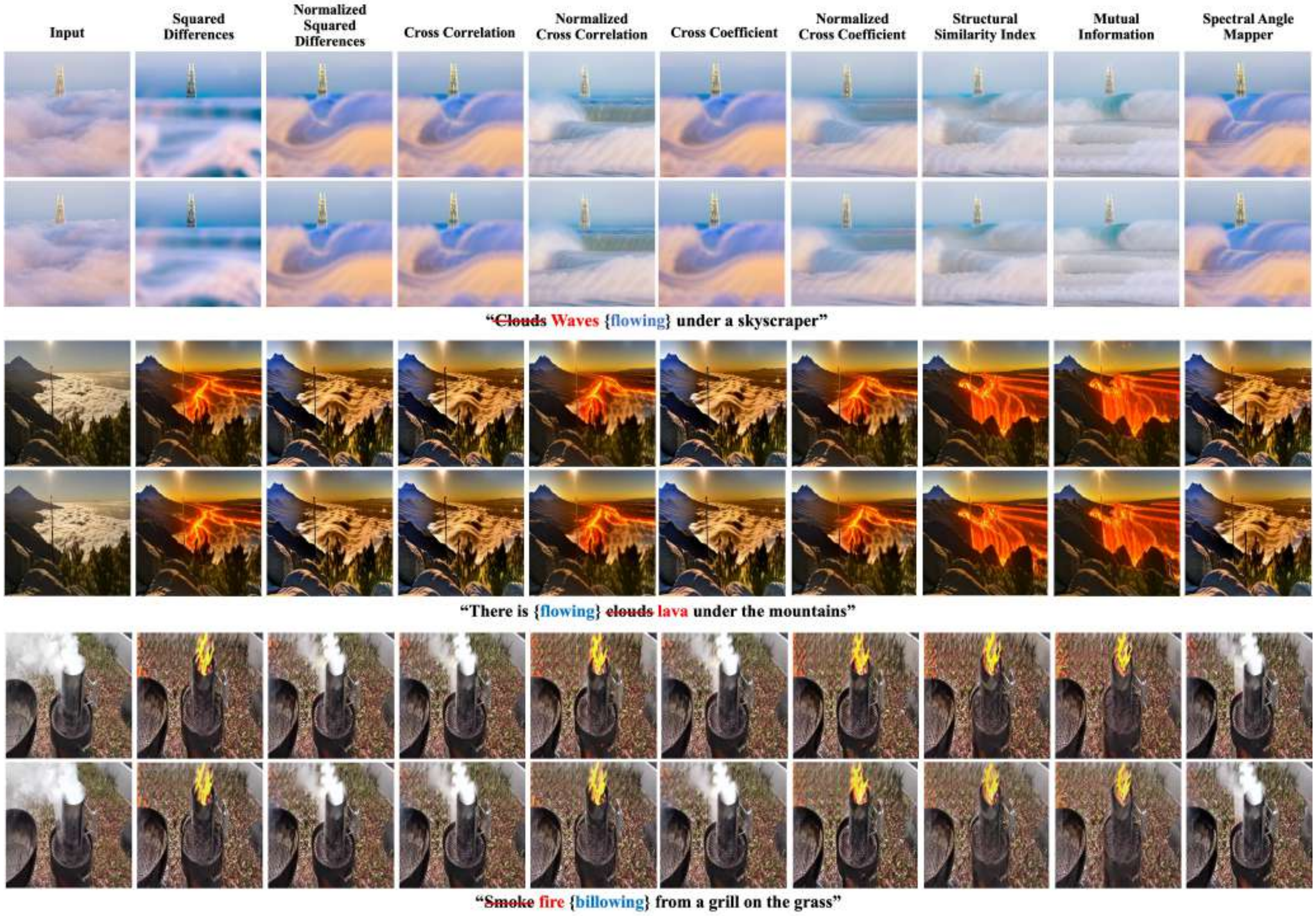

Input
Squared Differences
Normalized Squared Differences
Cross Correlation
Normalized Cross Correlation
Cross Coefficient
Normalized Cross Coefficient
Structural Similarity Index
Mutual Information
Spectral Angle Mapper
"~~Clouds~~ Waves {flowing} under a skyscraper"
"There is {flowing} ~~clouds~~ lava under the mountains"
"~~Smoke~~ fire {billowing} from a grill on the grass"

Figure 1. Comparison on various metrics of template matching algorithms with Video-P2P [1]

This supplementary materials are divided into the following sections.

- Sec. A.1 shows qualitative results from our module using various template matching algorithms and provides explanations for each methods.
- Sec. A.2 presents effect of the optical flow estimation models.
- Sec. A.3 presents runtime when integrating our model with existing models.
- Sec. B.1 provide a description of the code utilized in our study. Additionally, we introduce the project page where can find results not presented in the paper or supplementary materials.
- Sec. B.2 describes the method of extracting optical flow used in this study from Unimatch [2], providing equations and details alongside the explanation.
- Sec. B.3 provides an explanation of the metrics employed for quality evaluation in this study. It describes the CLIP Score, Masked PSNR, BRISQUE, and NIQE.
- Sec. B.4 provides detailed information on the User Study by presenting actual survey items and questionnaires.
- Sec. C introduces an alternative editing method that leverages optical flow information. Specifically, it suggests a technique for editing based on the directional information of optical flow, enabling edits on objects or pixels moving in a specific direction.
- Sec. D.1 presents additional experimental results for the 8-frame outcomes generated by existing video editing models set at default video length. It also shows results from extending the length beyond 8 frames, specifically editing at 24 frames.
- Sec. D.2 presents additional experimental results for 4 frame videos.
- Sec. E provides limitation of our proposed MMI module and our future work.

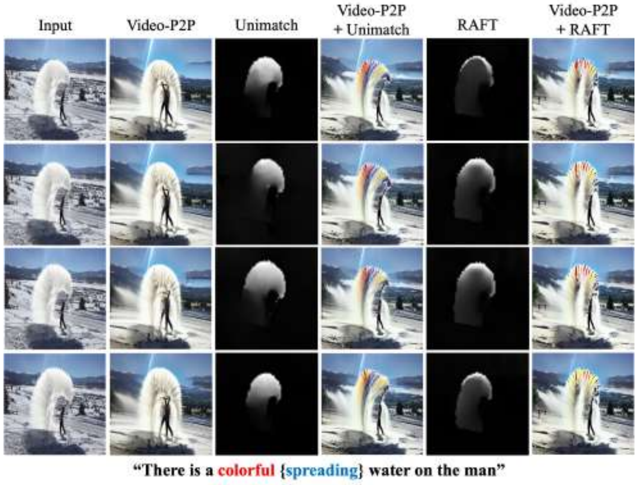


Figure 2. The results of Video-P2P [1] and our MMI module using different optical flow algorithms: RAFT [7] and UniMatch [2]

A. Ablation Study

A.1. Comparison on template matching algorithms

Our module employs Normalized Cross Correlation to calculate similarity, conducting a weighted sum of motion maps onto the entire attention maps, subsequently adjusting intensity levels. Therefore, we validate our module's editability by utilizing various metrics.

Within template matching, distance-based similarity metrics like Squared Difference, and correlation-based similarity metrics such as Cross Correlation [3] and Cross Correlation Coefficient were employed to measure the similarity. Additionally, we utilized Mutual Information [4], a concept from information theory quantifying the mutual dependence between random variables, Structural Similarity Index [5] to evaluate the structural similarity of images by modeling features of the human visual system, and Spectral Angle Mapper [6] to measure the spectral similarity between pixel spectra.

In Fig. 1, correlation-based normalized methods (Normalized Cross Correlation, Normalized Cross Coefficient) outperform other metrics including Spectral angle Mapper. In some case, both the Structural Similarity Index [5] and Mutual Information [4] shows novel results that were not observed within other template matching methods. Through this results, it is important to choose the template matching metrics well.

A.2. Comparison on Optical Flow Estimation Models

In this section, we compare the results of applying the optical flow estimation algorithms: Unimatch [2] and RAFT [7] to our MMI module. As can be seen in Fig. 2, Unimatch [2] allows for more precise optical flow estimation compared to RAFT [7]. However, applying both optical flow estimation to our module, it shows minor differences in editing results. This result demonstrated that identifying the overall motion of objects is more crucial than estimating detailed motion within specific areas. The failure to estimate the motion of the objects can be further verified in Sec. E.

A.3. Runtime Details

The runtime of the existing vid2vid-zero[8] is about 2m20s and the inference time is about 2m38s with our module. Also, the execution time of Video-p2p[1] increased from about 1m6s to 1m50s. Finally, FateZero[9] rose from 2m38s to 4m13s.

B. Detail Description

B.1. Code Descriptions

Our code is based on PyTorch version of Video-P2P [1]. We set the parameters as follows: frame_size_h = 512, frame_size_w = 512, number of frames = $4, 8, 24$ Code is available at https://anonymous.4open.science/r/AttentionFlow-197C/README.md Extensive experimental results are in https://currycurry915.github.io/Attention-Flow/.

B.2. Optical Flow Estimation

We obtain the optical flow V_{flow} through the pre-trained optical flow estimation model [2]. Firstly, we calculate the correlation for the pixels of the two frames by matrix product and then normalize for the last two dimensions using the softmax function. Then, matching distribution D_{flow} is obtained for each pixel location in F_{t-1} with respect to all pixel locations in F_t

$$D_{\text{flow}}(i,j,k,l) = \mathcal{S}\left(\frac{\sum_{d'=1}^{d} F_{t-1}(i,j,d') \cdot F_t(k,l,d')}{\sqrt{N}}\right), \quad (1)$$

where N denotes normalization coefficient to prevent the value from increasing after internal operation. We calculate the weighted average of the matching distribution D_{flow} on the 2D coordinates of the pixel grid G_{2D} to obtain correspondence $\hat{G}_{2D}$. Finally, the optical flow V_{flow} can be obtained by calculating the difference of corresponding pixel coordinates as follow

$$V_{flow} = \hat{G}_{2D} - G_{2D} \in \mathbb{R}^{H \times W \times 2}. \quad (2)$$

We then apply L2 norm to the generated optical flow V_{flow} to obtain motion map $\mathcal{M}$.

B.3. 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

$$CLIPScore(F_t, \mathcal{P}^*) = max(100 * cos(E_{F_t}, E_{\mathcal{P}^*}), 0), \tag{3}$$

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 [1]. 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,

$$M.PSNR(V^*, V) = PSNR(B(V^*, M), B(V, M)), \tag{4}$$

where $B(V, M) = V_M$ is defined as a reversed mask binary function, so only regions not to be changed are involved in measuring Masked PSNR.

BRISQUE (Blind Referenceless Image Spatial Quality Evaluator) One of the No-Reference image quality assessment used to measure the quality of the edited image. We evaluated the edited video without the original video.

NIQE (Natural Image Quality Evaluator) Another No-Reference metric used as an image quality assessment. It assesses the naturalness of an image by taking into account the deformation and distortion present in the image. NIQE analyzes features of an image to generate a quantitative quality score. Therefore, similar to BRISQUE, we also used NIQE to evaluate the quality of the edited video.

B.4. Details about User Study

In this section, we describe the details of the user study. A total of 60 participants were asked to choose the more preferable video for 20 videos. Among the 20 videos, Video-P2P [1] was included with 10, and vid2vid-zero [10] and FateZero [9] were included with 5 each. The users were asked to answer following questions: “Edited with consideration of the target prompt”, “Maintained the overall structure well after editing”, and “The editing result is realistic and of high quality”. The order of videos were randomized.

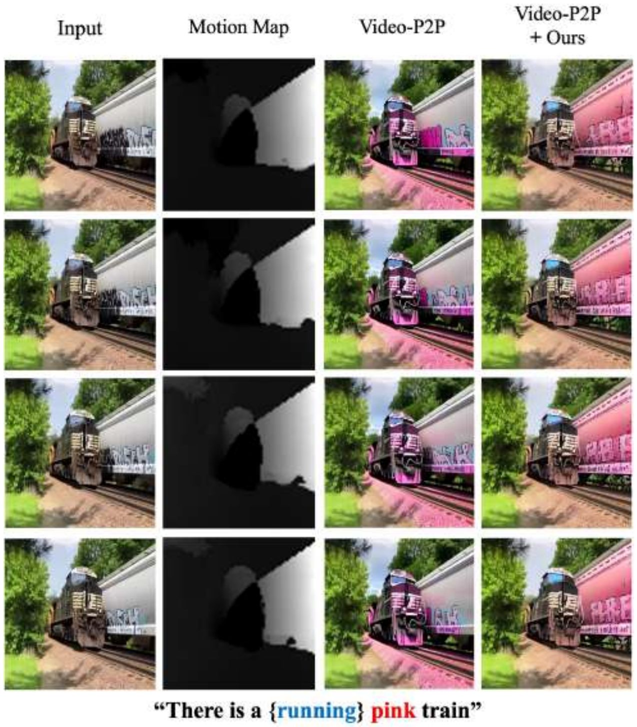


Figure 3. Direction control method for objects moving in the direction specified by the user. Before editing, the user first selects one of the 8 directions. Then, we can edit certain direction in video using optical flow information

The following Figs. 11, 12 are part of the questionnaire provided to participants.

C. Application: Direction Control

The user can choose to edit the motion value in the desired direction because V_{flow} contains information on both the magnitude and direction of pixel movement between frames. By rotating the optical flow V_{flow} in accordance with the direction D that the user specifies before injecting it, our model enables the user to edit content in a particular direction. Optical flow is thought to have any directions depending on movement, but in this study, only eight directions were used for D. 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 current 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

$$\begin{bmatrix} X' \\ Y' \end{bmatrix} = \begin{bmatrix} X\cos\theta_D & -Y\sin\theta_D \\ X\sin\theta_D & Y\cos\theta_D \end{bmatrix}, \tag{5}$$

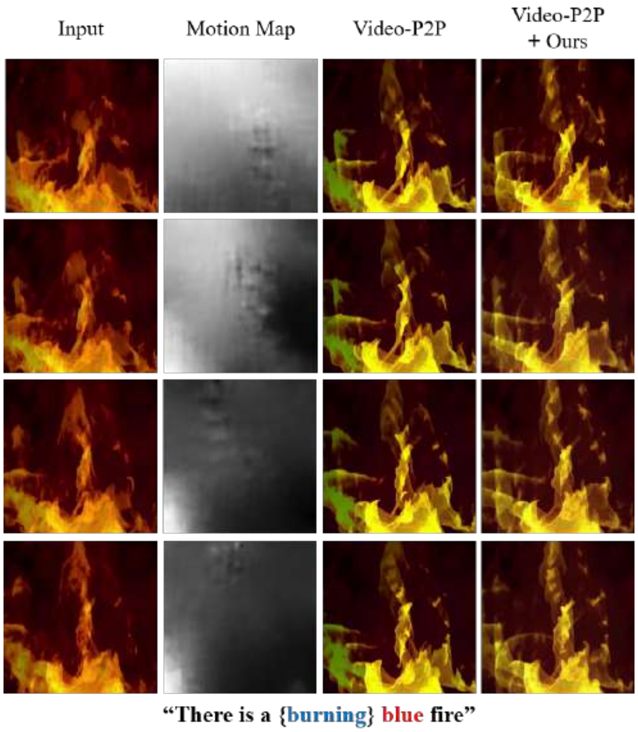


Figure 4. The results of Video-P2P [1] and our proposed framework when the motion of video is not estimated well, like the second column. Since the motion is obscure, the 'fire' is not colored into 'blue'.

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. Also, 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. The results can be seen in Fig. 3.

D. Additional Qualitative Results

We further conducted experiments to show the adaptability of our module to diverse number of frames and videos.

D.1. Additional Qualitative Results on 8, 24 Frames

In Figs. 5 and 6 present additional results of Video-P2P [1] with 8 frames which is widely used setting from existing video researches, including FateZero [9], Tune-A-video [11], vid2vid-zero [10], and Video-P2P [1]. Furthermore, Fig. 7 confirmed that our module could be extended to an 24 frame videos. It shows our module can enhance the editability regardless of the number of frames.

D.2. Additional Qualitative Results on 4 Frames

In Fig 8, 9 and 10, we additionally compared our module with baseline models such as Video-P2P [1], FateZero [9], and vid2vid-zero [10]. It confirm that the enhanced attention by our module can enhance editability in text-guided video editing models.

E. Limitations & Future Work

As shown in Fig. 4, since the optical flow for the movement of fire is not estimated well, there was no difference from the result of original Video-P2P [1]. Therefore, it was confirmed that our model has difficulty in enhancement of editing if the estimated motion of the input video is inaccurate.

As mentioned in the experiment of optical flow comparison, for our module, estimating the overall motion of objects is more critical than precise estimation of specific areas. Recent researches in optical flow estimation models focus on estimating the detail of optical flow. However, for our module, the model that better estimates the overall motion of objects is needed. Therefore, in the future work, we will aim to develop an optical flow model appropriate for our module.

References

[1] Shaoteng Liu, Yuechen Zhang, Wenbo Li, Zhe Lin, and Jiaya Jia. Video-p2p: Video editing with cross-attention control. *arXiv preprint arXiv:2303.04761*, 2023. 1, 2, 3, 4, 6, 7, 8, 9

[2] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, Fisher Yu, Dacheng Tao, and Andreas Geiger. Unifying flow, stereo and depth estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 1, 2

[3] Paul Bourke. Cross correlation. *Cross Correlation", Auto Correlation—2D Pattern Identification*, 1996. 2

[4] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical review E*, 69(6):066138, 2004. 2

[5] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 2

[6] Xiaofang Liu and Chun Yang. A kernel spectral angle mapper algorithm for remote sensing image classification. In *2013 6th International Congress on Image and Signal Processing (CISP)*, volume 2, pages 814–818. IEEE, 2013. 2

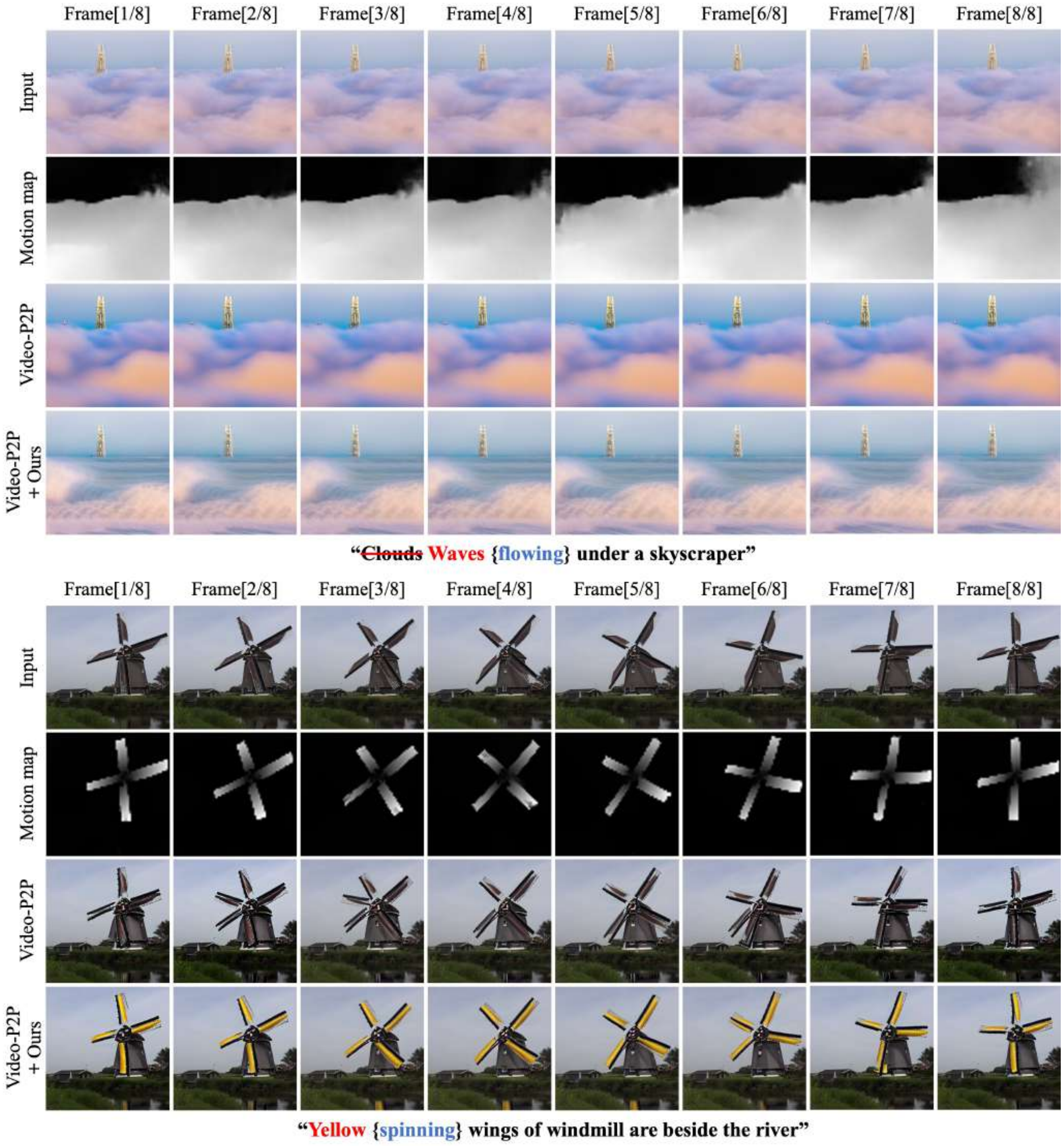
[7] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020. 2

[8] Shuai Yang, Yifan Zhou, Ziwei Liu, and Chen Change Loy. Rerender a video: Zero-shot text-guided video-to-video translation. *arXiv preprint arXiv:2306.07954*, 2023. 2

[9] Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. Fatezero: Fusing attentions for zero-shot text-based video editing. *arXiv preprint arXiv:2303.09535*, 2023. 2, 3, 4, 10

[10] Wen Wang, Kangyang Xie, Zide Liu, Hao Chen, Yue Cao, Xinlong Wang, and Chunhua Shen. Zero-shot video editing using off-the-shelf image diffusion models. *arXiv preprint arXiv:2303.17599*, 2023. 3, 4, 9

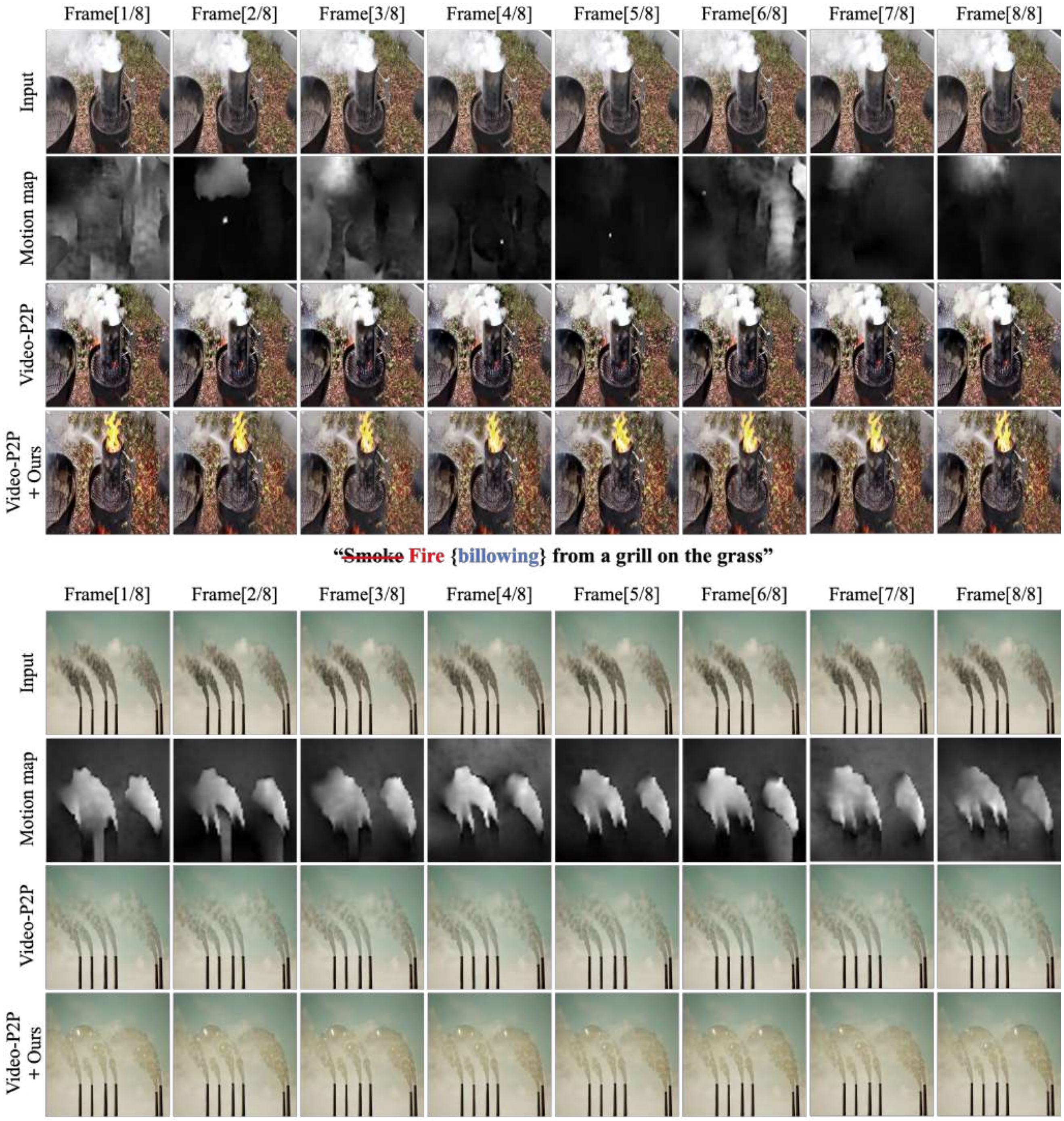
[11] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7623–7633, 2023. 4



“~~Clouds~~ Waves {flowing} under a skyscraper”

“Yellow {spinning} wings of windmill are beside the river”

Figure 5. Additional result of 8 frames editing with Video-P2P [1] (1/2)



"~~Smoke~~ Fire {billowing} from a grill on the grass"

"~~Smoke~~ Bubble {coming} out of the chimney"

Figure 6. Additional result of 8 frames editing with Video-P2P [1] (2/2)

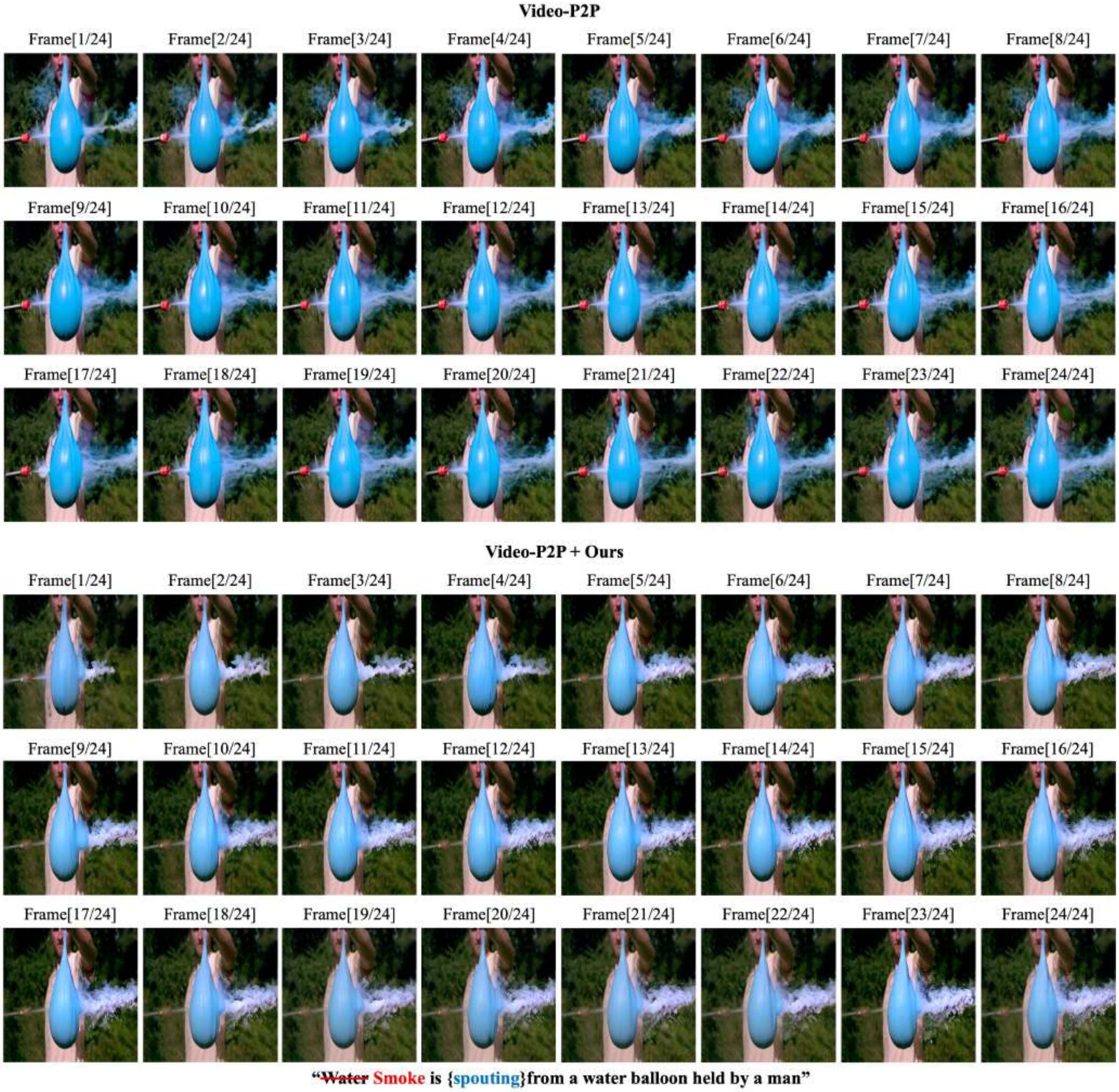


Figure 7. Additional result of 24 frames editing with Video-P2P [1].

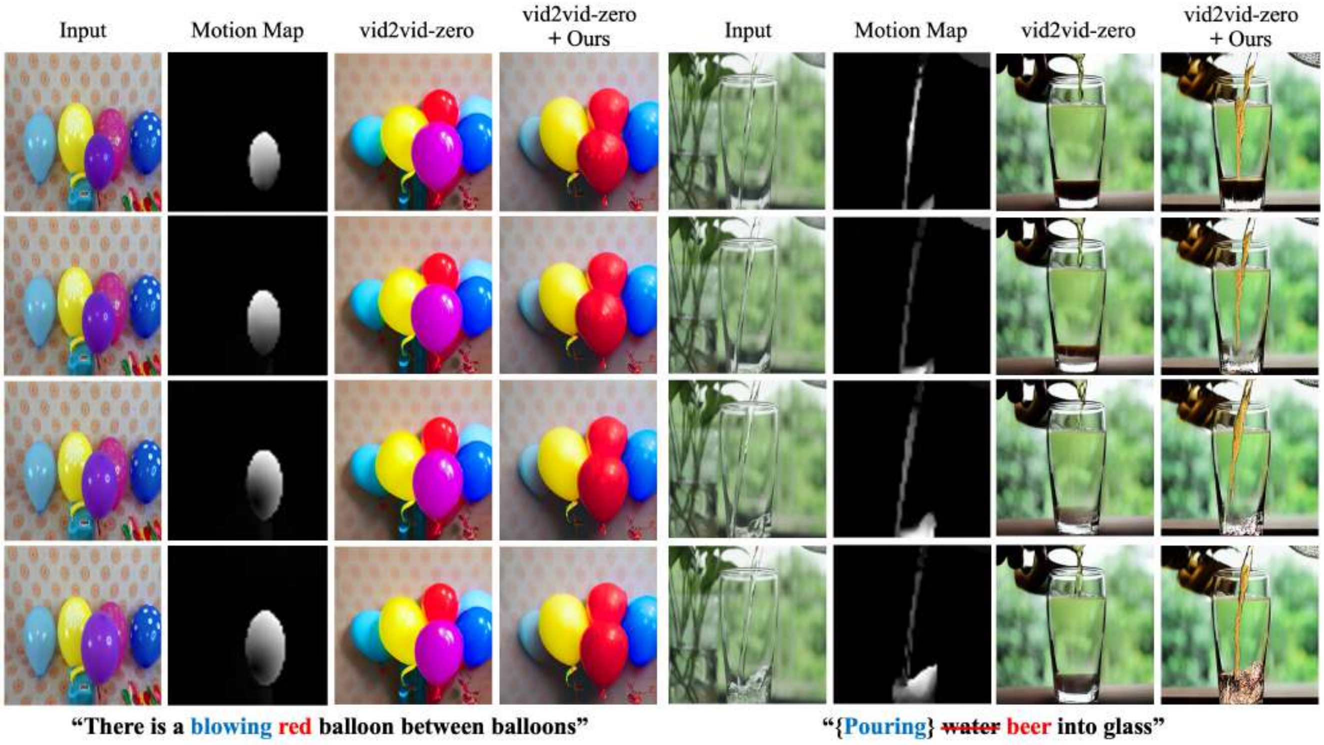


Figure 8. Additional results of 4 frames editing with vid2vide-zero [10]

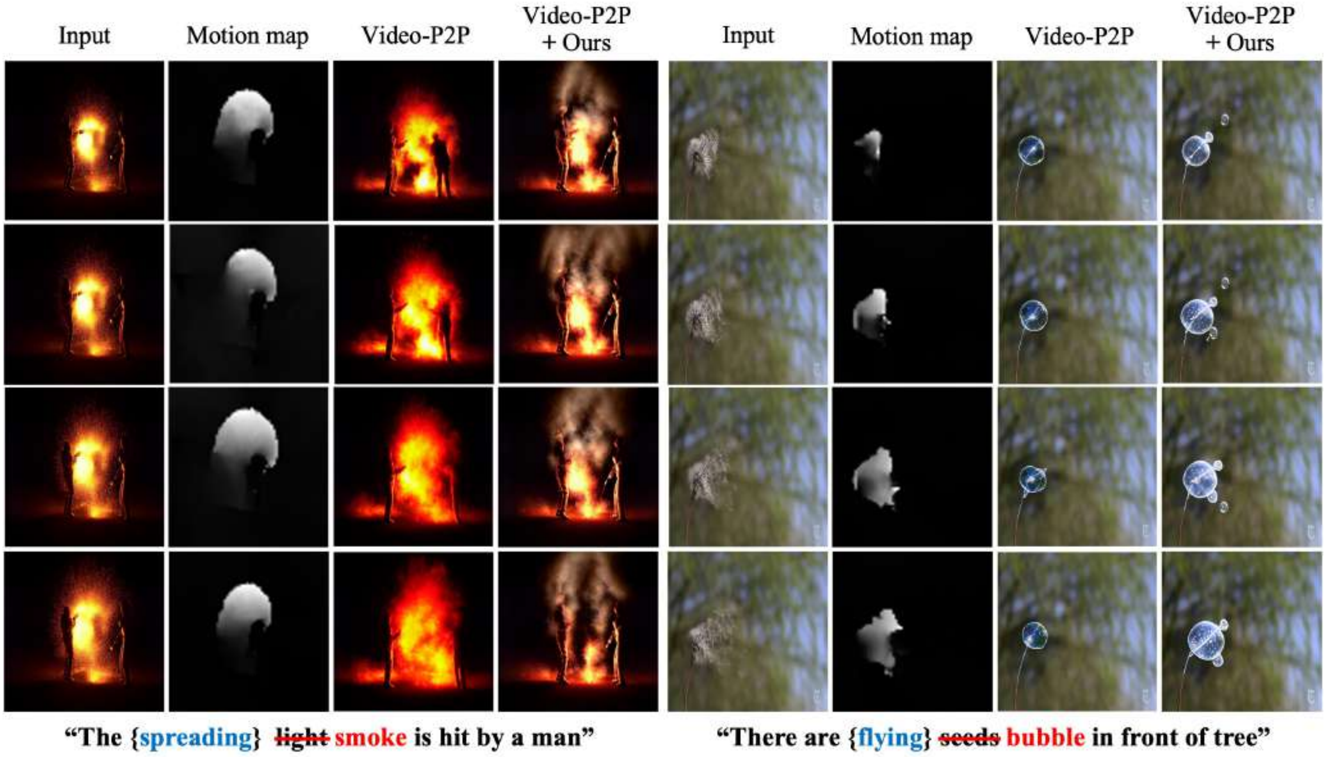

Input
Motion map
Video-P2P
Video-P2P + Ours
Input
Motion map
Video-P2P
Video-P2P + Ours
"The {spreading} ~~light~~ smoke is hit by a man"
"There are {flying} ~~seeds~~ bubble in front of tree"

Figure 9. Additional results of 4 frames editing with Video-P2P [1]

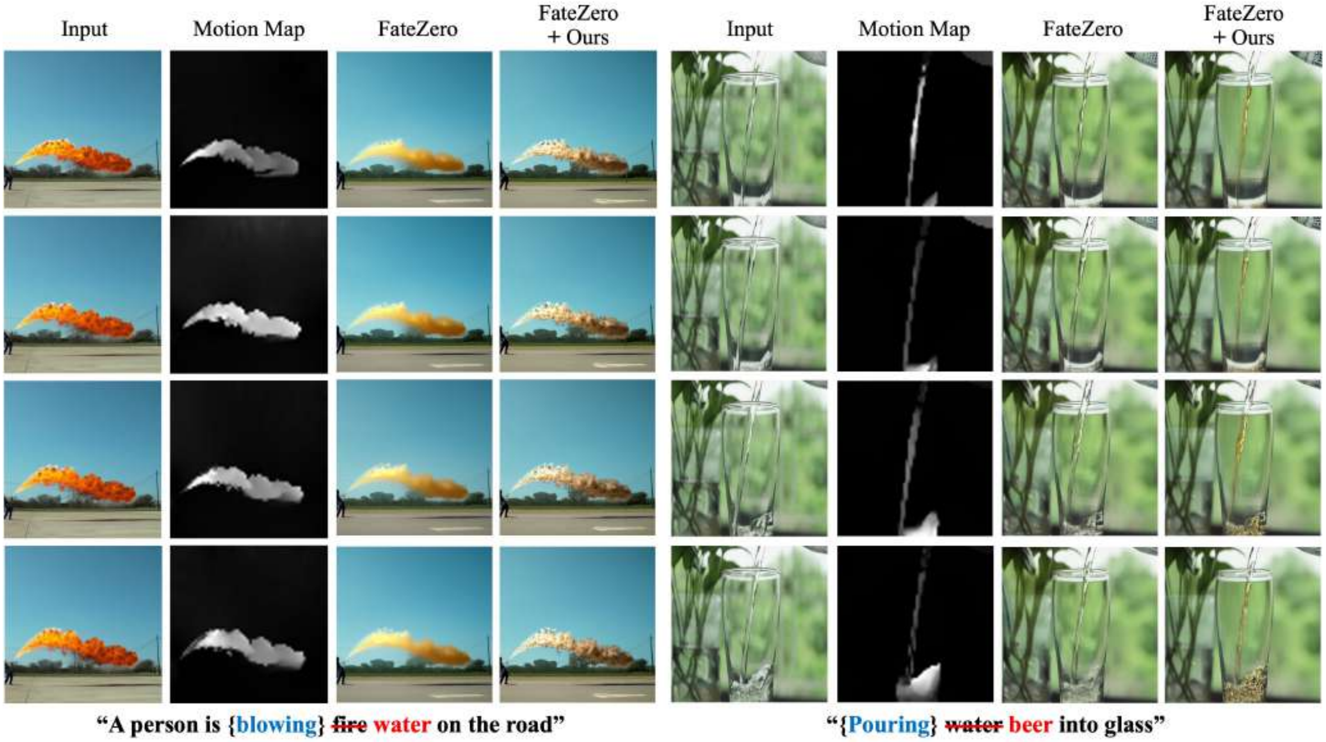
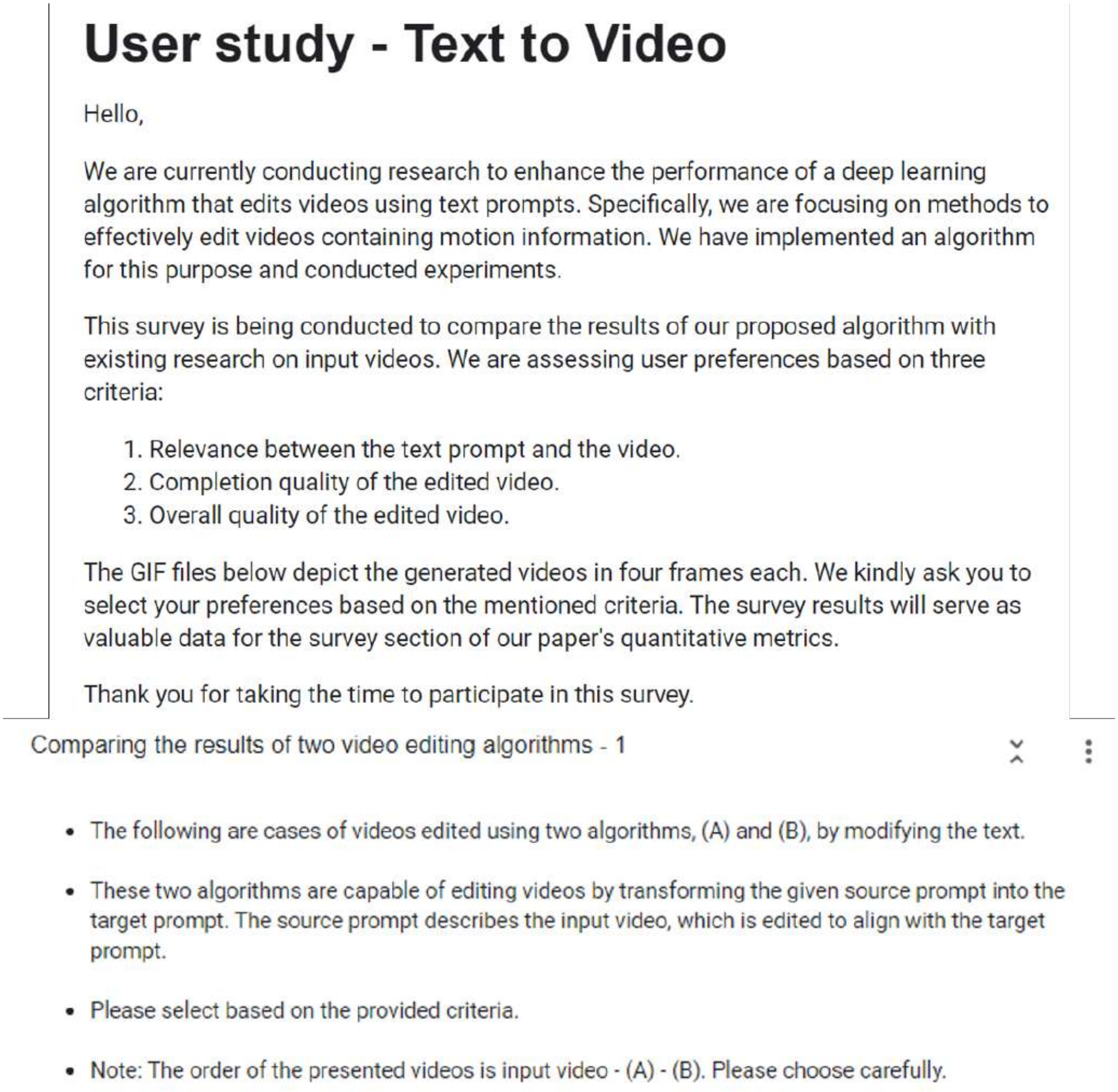

Input
Motion Map
FateZero
FateZero + Ours
Input
Motion Map
FateZero
FateZero + Ours
"A person is {blowing} ~~fire~~ water on the road"
"{Pouring} ~~water~~ beer into glass"

Figure 10. Additional results of 4 frames editing with FateZero [9]



User study - Text to Video

Hello,

We are currently conducting research to enhance the performance of a deep learning algorithm that edits videos using text prompts. Specifically, we are focusing on methods to effectively edit videos containing motion information. We have implemented an algorithm for this purpose and conducted experiments.

This survey is being conducted to compare the results of our proposed algorithm with existing research on input videos. We are assessing user preferences based on three criteria:

1. Relevance between the text prompt and the video.
2. Completion quality of the edited video.
3. Overall quality of the edited video.

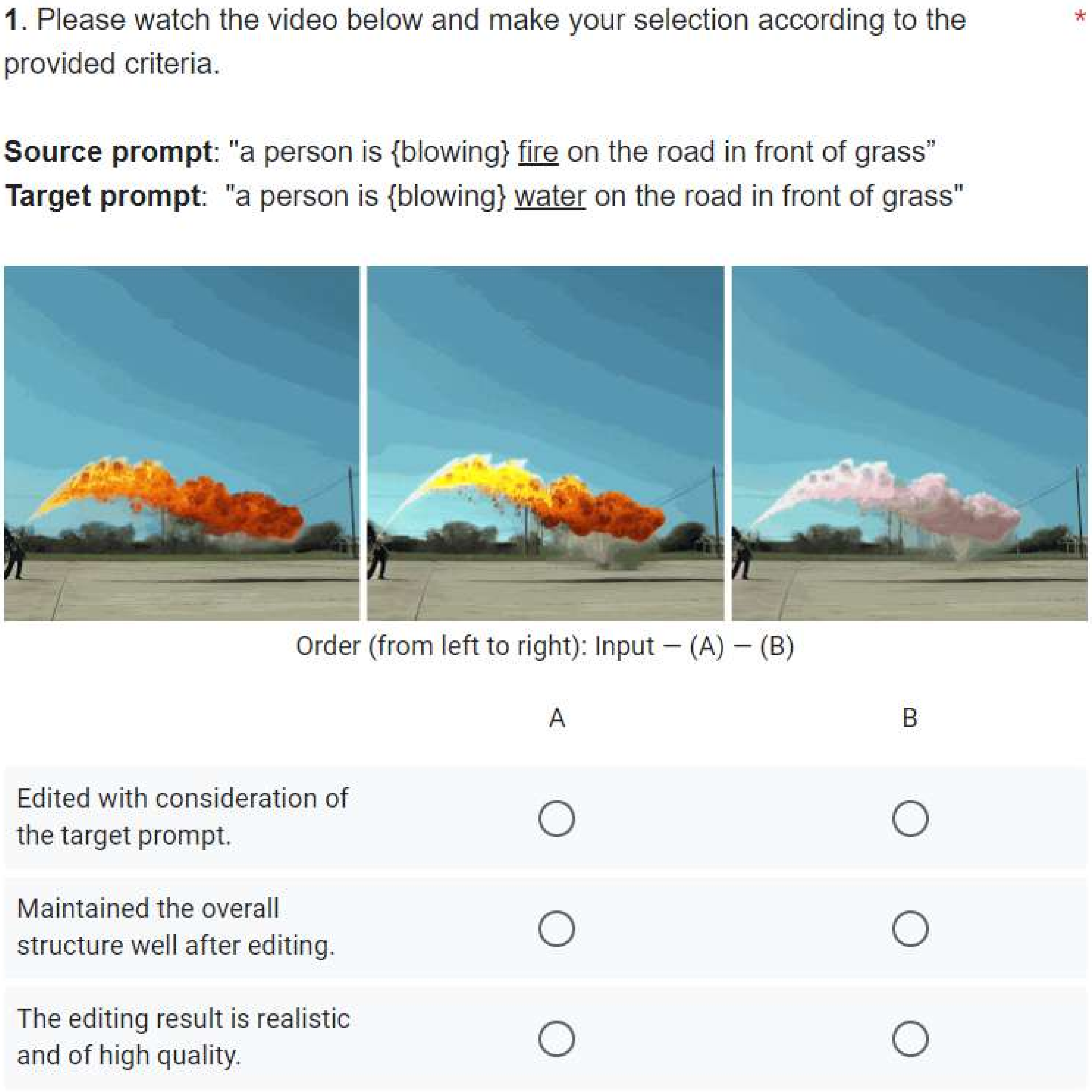
The GIF files below depict the generated videos in four frames each. We kindly ask you to select your preferences based on the mentioned criteria. The survey results will serve as valuable data for the survey section of our paper's quantitative metrics.

Thank you for taking the time to participate in this survey.

Comparing the results of two video editing algorithms - 1

- The following are cases of videos edited using two algorithms, (A) and (B), by modifying the text.
- These two algorithms are capable of editing videos by transforming the given source prompt into the target prompt. The source prompt describes the input video, which is edited to align with the target prompt.
- Please select based on the provided criteria.
- Note: The order of the presented videos is input video - (A) - (B). Please choose carefully.

Figure 11. General guidance on user study and information on video editing algorithm and precautions.



1. Please watch the video below and make your selection according to the provided criteria. *

Source prompt: "a person is {blowing} fire on the road in front of grass"

Target prompt: "a person is {blowing} water on the road in front of grass"

Order (from left to right): Input — (A) — (B)

	A	B
Edited with consideration of the target prompt.	○	○
Maintained the overall structure well after editing.	○	○
The editing result is realistic and of high quality.	○	○

Figure 12. Questionnaires viewed by actual participants.