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

Anonymous WACV Algorithms Track submission

Paper ID 920

Abstract

Text-to-image diffusion, which has been trained with a large amount of text-image pair dataset, shows remarkable performance in generating high-quality images. Recent research using diffusion model has been expanded for text-guided video editing tasks by using text-guided image diffusion models as baseline. Existing video editing studies have devised an implicit method of adding cross-frame attention to estimate frame-frame attention to attention maps, resulting in temporal consistent editing. However, because these methods use generative models trained on text-image pair data, they do not take into account one of the most important characteristics of video: motion. When editing a video with prompts, the attention map of the prompt implying the motion of the video, such as 'running' or 'moving', is not clearly estimated and accurate editing cannot be performed. In this paper, we propose the 'Motion Map Injection' (MMI) module to perform accurate video editing by considering movement information explicitly. The MMI module provides a simple but effective way to convey video motion information to T2V models by performing three steps: 1) extracting motion map, 2) calculating the similarity between the motion map and the attention map of each prompt, and 3) injecting motion map into the attention maps. Considering experimental results, input video can be edited accurately and effectively with MMI module. To the best of our knowledge, our study is the first method that utilizes the motion in video for text-to-video editing.

1. Introduction

Recent research using text guided diffusion models and large-scale language models has led to unprecedented advances in image generation and image editing. In the field of image editing, various studies have been conducted, including fine-tuning a diffusion model for a few images [1, 2], zero-shot image editing [3,4], guiding the region to be modified with mask [5, 6], or modifying an image using a text prompt [7–10]. Among them, Prompt-to-Prompt [7] offers users intuitive image editing by proposing several methods and applications to control the attention map that identifies the semantic relationship between prompt token and input image.

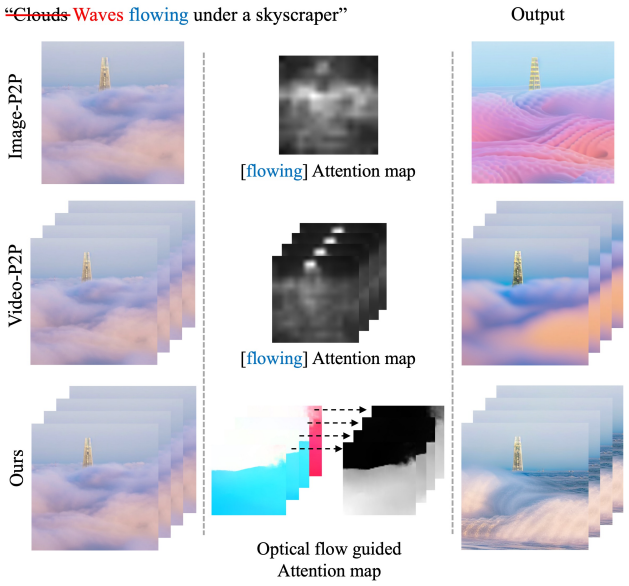


Figure 1. Comparison of video editing output and attention map compared to existing methods. Both Image-P2P and Video-P2P failed to accurately estimate the attention map, resulting in discrepancy with the prompt. Our study performed realistic video editing by enabling accurate attention maps through optical flow guided attention maps.

Research using diffusion model [11–14] has been expanded for text-guided video editing tasks. Most of these studies use text-guided image diffusion models as baseline. However, since a video consists of several frames unlike images, it should be edited in consideration of temporal information. To this end, existing video editing studies [13,15] have devised an implicit method of adding cross-frame attention to estimate frame-frame attention to attention maps, resulting in temporal consistent editing. Among them, video-P2P [14] is a study that enables video editing as

intuitive as image P2P [7] by applying image P2P to video editing.

However, a vast amount of text-video pair dataset is required to train the implicit structures that uses cross-frame attention to understand temporal information. However, there is no such amount of data enough to train those structure. That is why the most of video editing algorithms have chosen the method of fine-tuning image diffusion model for the input video or zero-shot approach to edit video in consideration of temporal information. Because the Image diffusion model was trained on the text-still image pair dataset, it is not good at estimating the attention map for prompts containing motion as shown in Figure 1. Therefore, it is difficult to perform appropriate attention control for the prompt, which reduces the capability of editing videos. Motion information needs to be estimated and reinforced explicitly through an additional network.

Optical flow estimation is the task estimating the information (i.e. motion) of pixels that have moved between frames of video, and is the algorithm that has been used in a wide range of video tasks. In recent years, deep learning has been applied to optical flow estimation, resulting in significant performance improvements. Starting with applying convolution neural network to optical flow, a series of following studies proposed a coarse-to-fine and iterative estimation methodology to reduce computational costs. Most recently, a methodology with Transformer improved accuracy and efficiency of estimating motion, while addressing the long-standing challenges of motion estimation on occluded pixels and large displacements. By exploiting optical flow estimation, which shows high performance and utility in motion extraction for video tasks, highly accurate motion can be estimated to be injected into the video editing model.

In this paper, we propose a method to complement attention maps with optical flow to make video editing more effective and accurate. Our proposed framework consists of the following three steps. Firstly, existing optical flow algorithm estimates motion in input video. Secondly, attention maps identified by off-the-shelf video diffusion model are complemented with motion estimated by optical flow. Lastly, modify the video by applying the complemented attention map.

The most important part of this framework is the Motion Map Injection (MMI) module, which adds optical flow information to the existing attention map. First, the optical flow of the input video is estimated, and the magnitude of this value is used as the motion map. After that, the similarity through template matching between the attention map and the motion map of all prompts is calculated. At this point, Normalized Cross Correlation (NCC) is used for similarity. Finally, the motion map is multiplied by the similarity weight, and added to the every attention map of each prompt. However, the image diffusion model does not accurately estimate the attention map of the motion prompt token such as 'flowing', so the motion map is directly inserted instead of template matching. Our proposed MMI module is a groundbreaking method that can supplement motion information in the prompt without learning a vast video dataset, which can dramatically improve the editing of the existing video editing framework.

Another benefit of our module is additional capability of video editing with motion control. Using motion, selective editing of objects moving in a specific direction becomes possible when there are objects moving in various directions. In this study, only contents moving in the direction specified by the user can be edited by rotating the motion map according to the direction provided by the user before injecting it. The motion can be estimated in various directions depending on movements, but in this study, directions are typically considered for areas in eight directions.

In essence, our main contributions include:

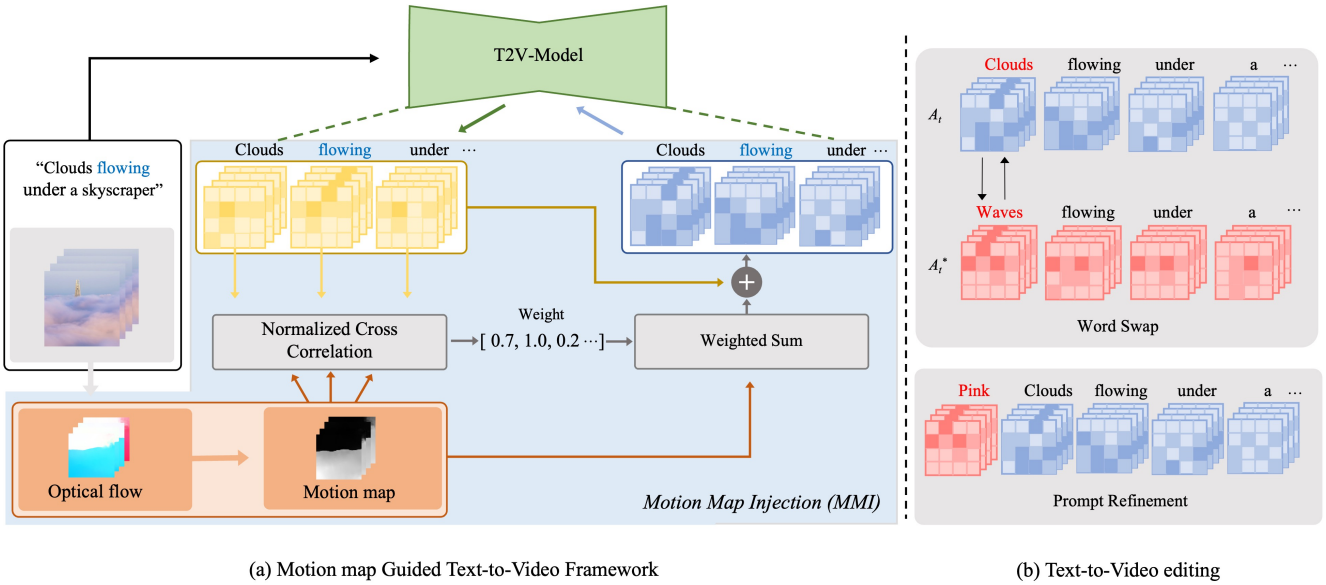
- To the best of our knowledge, our study is the first attempt to use optical flow information in the field of text-to-video editing.
- Our proposed MMI module gives a way to effectively inject motion extracted from video into the attention map of the prompt.
- We found that editing a video with the motion extracted from the video improves general editing performance and enables selective editing according to the direction in which the object moves.

2. Related Work

2.1. Text-Guided Editing

The diffusion model [16, 17], which has recently been actively studied, generates data from noise through the process of adding or removing noise. Based on this diffusion model, text-guided image editing models such as DALL-E2 [18], Imagen [19], and stable diffusion [20] show the results of high-quality image editing. In particular, Prompt-to-Prompt presents text-guided image editing that controls the relationship between the prompt text token and the corresponding image pixel with the attention maps, enabling unprecedented semantic editing. In addition, subsequent papers such as DreamBooth [1], EDICT [21], and Imagic [2] have been actively studied recently, showing impressive results for text-guided image editing.

Based on the significant progress of text-guided image editing, research has recently been expanded to text-guided video editing with the generative model. Dreamix [11] presents the first diffusion-based method of performing text-guided motion and application editing of videos through fine-tuning, but there are difficulties with localized editing by replacing a word. Video-P2P [14] divides



(a) Motion map Guided Text-to-Video Framework

(b) Text-to-Video editing

Figure 2. Overall framework of this study. First, the T2V-Model generates an attention map by receiving video and prompts as input. At the same time, the Motion Map Injection module receives the video frame, generates a motion map, and injects it into the attention map of the T2V-Model. After that, text-to-video editing is performed using the attention map that includes video motion information.

their framework into two branches for unchanged parts and edited parts, and incorporates each attention map to enable detailed editing.

Concurrent to above works, Vid2vid-zero [13] performs stable video reconstruction and editing by adding cross-frame attention to the U-Net structure of the existing diffusion model. In addition, FateZero [22], based on zero-shot, stores an attention map during the inversion process to maintain temporal consistency for structure and motion information. We attempt the first study to extract motion information directly from video and apply it to video editing.

2.2. Optical Flow Estimation

Optical flow estimation is a computer vision task that involves computing the motion of objects in a video sequence. Recently, this field is significantly advanced through the rise of deep neural networks. FlowNet [23] was the first fully convolutional neural network for estimating optical flow. Then, a series of works, represented by SpyNet [24], PWC-Net [25], LiteFlowNet [26], and RAFT [27] were proposed to reduce the computational costs through coarse-to-fine and iterative estimation methodology. Recently GMFlow [28] were proposed to achieve highly accurate results without relying on a large number of refinements by performing global matching with a Transformer. Among the various optical flow estimation models, we employ recent GMFlow in this work.

Optical flow estimation is used in various video tasks. First, video action recognition [29, 30] aims to automatically recognize the behavior of objects in video sequences, where optical flow is used as a useful motion representation in video motion representation. Using spatio-temporal information from surrounding scenes to fill in new content for damaged areas, video inpainting [31–33] enables spatio-temporally stable synthesis between frames of video through optical flow. Video super resolution [34–36] is the field of generating high-resolution video frames from low-resolution video frames, and generally maintains temporal consistency between video frames by using optical flow as motion compensation. Video frame interpolation (VFI) [37, 38], a technology that generates an intermediate frame between two consecutive frames, also effectively extracts motion and shape information between frames by utilizing optical flow to estimate motion information between frames. Our work is the first attempt to apply optical flow to text-guided video editing where motion information is important based on the proven validity of the optical flow estimation in various video fields.

3. Proposed Method

Let be input video $\mathcal{V}$, consists of n frames. We define the source prompt as $\mathcal{P}$ just like the Prompt-to-Prompt setting. In $\mathcal{P}$, the prompt containing motion information of the video (for example, running, moving, etc.) is called motion prompt $\mathcal{P}_{\mathcal{M}}$.

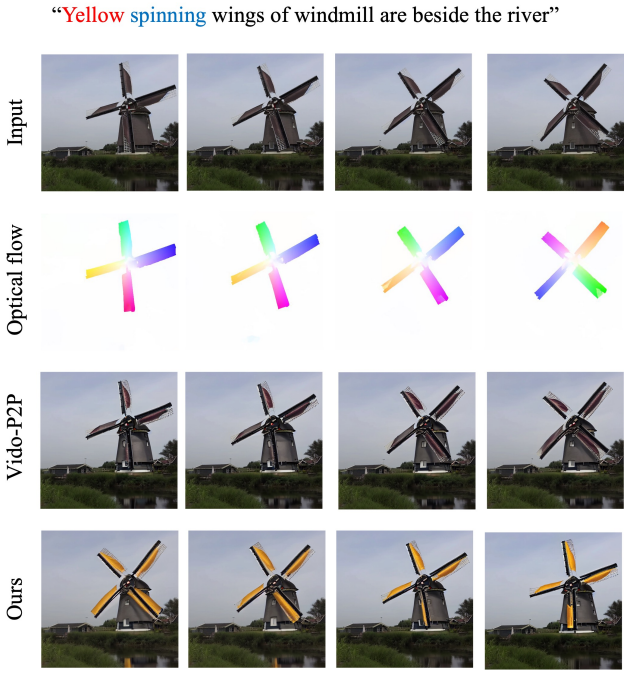


Figure 3. Examples that the existing video editing model could not edit object having motion. However, editing was performed by directly injecting optical flow into the object prompt.

In this section, we provide an overview of our framework, which is illustrated in Fig. 2. Our framework is structured to complement the text-to-video model through the Motion Map Injection (MMI) module. The attention map $\mathcal{A}_\mathcal{P}$ representing the correlation between the frame of video $\mathcal{V}$ and the prompt $\mathcal{P}$ is estimated through the cross-attention layer inside the network. The estimated motion map is injected into $\mathcal{A}_\mathcal{P}$ through the MMI Module, and then edited video $\mathcal{V}$ with n frames through the editing process of the Text-to-Video Editing network. Our proposed MMI Module consists of two steps for effective video editing. At the first step, pre-trained optical flow estimation network [28] estimates optical flow V_{flow} of input video $\mathcal{V}$. At the second step, a motion map is obtained by applying L2 norm to the optical flow. By calculating the correlation between the attention map of all prompts and the motion map, it complements the existing attention map according to the correlation. Lastly inject the motion map into the attention map of motion prompt. This framework is a method that enables effective video editing even for motion prompts.

In the following sections, we describe the Prompt-to-Prompt [7] method of editing an image using an attention map through Preliminary (Sec 3.1). In addition, we propose a new method of injecting motion maps into attention maps in MMI module (Sec 3.2.)

3.1. Preliminary

Prompt-to-Prompt P2P [7] based on text-guided diffusion model edits image by modifying source prompt $\mathcal{P}$. At each diffusion step t, diffusion model predicts noise ϵ from noise image z_t and text embedding $\psi(\mathcal{P})$ through U-Net. Cross-attention layer produces cross-attention map indicating spatial correlation between visual and textual feature. Spatial features of noise image $\phi(z_t)$ are projected on query matrix $Q = \ell_Q(\phi(z_t))$ and key matrix $K = \ell_K(\psi(\mathcal{P}))$ through learned linear projections ℓ_Q and ℓ_K. Attention map $\mathcal{A}$ can be written as

$$A = Softmax\left(\frac{QK^T}{\sqrt{d}}\right), \tag{1}$$

where d denotes latent projection dimension of key and query. Because the attention map A equals matrix product of spatial information of visual feature Q and transpose of textual embedding's key K^T, we can see where the textual feature affects the image. Because P2P can edit images by controlling these attention, it gives intuitive image editing by modifying only the prompt. Recent studies extend Prompt-to-Prompt to video domain from image domain.

3.2. Motion Map Injection Module

Motion Map Extraction In this paper, we use optical flow estimation algorithm to estimate motion of video. Optical flow refers to the amount of movement that appears between the previous frame F_{t-1} and the subsequent frame F_t. Among the various optical flow estimation algorithm, we exploit Unimatch [28]. 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.

$$D_{flow} = softmax(\frac{F_{t-1}F_t^T}{\sqrt{N}}) \in \mathbb{R}^{H\times W\times H\times W}, \tag{2}$$

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

$$V_{flow} = \hat{G}_{2D} - G_{2D} \in \mathbb{R}^{H\times W\times 2}. \tag{3}$$

We then apply L2 norm to the generated optical flow V_{flow} to obtain motion map V^*_{flow}. Before proposing our method, we would like to first show the feasibility of our method. Even if the prompt is written more specifically

or accurately to be given P2P-based video editing models, there are times when editing is not accurate. To solve this problem, we utilized the optical flow estimation task, which estimates motion information of the video. As a result of injecting optical flow directly into the attention map of the motion prompt, it was confirmed that accurate editing was performed, as shown in Fig 3. Inspired by the fact that simply injecting motion information can do an edit that existing video editing models cannot, we construct an algorithm to edit video by injecting the motion information of the video into the attention map. Specifically, motion information is injected into motion prompts (e.g. running, moving, etc.), which are essential to describe movement, but are not accurately estimated. We also proposes a method of propagating the motion information of the video to the entire prompt in consideration of the correlation between all prompts except the motion prompt.

Motion Map Injection using NCC It is necessary to calculate the similarity of each other to effectively supplement the existing attention map using the estimated motion map. To this end, we use the template matching algorithm. We set the source image to attention map and the target image to motion map and calculate the Normalized Cross Correlation (NCC) as below

$$C_k = \frac{1}{n}\sum_{x,y}\frac{(V^*_{flow}(x,y) - \bar{V}^*_{flow})(A_k(x,y) - \bar{A}_k)}{\sigma_V \sigma_A}, \quad (4)$$

where n denotes the number of pixels, x, y denotes pixel of source image and target image. $\bar{V}^*_{flow}$ and $\bar{A}_k$ denote the average of pixels of V^*_{flow} and A_k. σ_A denotes the standard deviation of pixels of V^*_{flow} and A_k. k denotes the index of a particular word in the entire prompt. We calculate the correlation score by performing template matching at every denoising step t of the diffusion model, making the sizes of V^*_{flow} and A_k alike for all k.

Finally, as shown in (5), a new attention map is identified by injecting $C_k \cdot V^*_{flow}$ to attention map of each prompt

$$A^*_k = A_k + \lambda \cdot \frac{(C_k * V^*_{flow})}{t} \quad (5)$$

where λ is a hyperparameter for motion map injection rate, t denotes denoising step. The affect of each hyperparameter is shown in ablation study of experimental result (Sec 4.5) for all k.

4. Experiments

4.1. Experimental Setup

Baseline Model We used the Video-P2P model as a baseline model, which edits video with text. Video-P2P is a model that can intuitively edits only with text without additional information by manipulating a cross attention map representing the relationship between text and video. We added the proposed MMI module to Video-P2P to verify the effectiveness of our framework.

Dataset Experiments were conducted using Davis video dataset, which is have been used for various video tasks, and YouTube videos were collected and used to make more experiments.

Implementation details We experimented with RTX 3090 GPUs and set the image resolution to 512×512 just like existing Video-P2P. The number of video frames was fixed to four, and the UniMatch [28] model was used to extract the optical flow.

Evaluation Metrics CLIP score [39] and masked PSNR [14] were used to evaluate textual similarity and region preservation. CLIP Score is a metric which can be used to evaluate the correlation between a prompt and image. It has been found to be highly correlated with human judgement. Masked PSNR is a metric which evaluates the quality of structure preservation only for the masked region. In this experiment, we exploited non-moving region as a mask.

Table 1. A user study evaluating the performance of Video-P2P and the output of our model conducted on three items: Structure Preserving, Text Alignment, Realism and Quality.

User Preference	Video-P2P (%)	Ours (%)
Structure Preserving	31.25	**68.75**
Text Alignment	21.87	**78.12**
Realism & Quality	43.75	**56.25**

Table 2. Quantitative comparison results. Higher performance was recorded for CLIP Score and Masked PSNR compared to Video-P2P.

CLIP Score		Masked PSNR	
Video-P2P	Ours	Video-P2P	Ours
28.98	**29.63**	22.09	**22.60**

4.2. User Study

To compare our proposed model with Video-P2P, we conducted user study on three criteria : structure preserving, text alignment, and realism & quality. We presented input video, target prompt, and videos edited with each model to users. Between two edited videos, the evaluation of outputs was conducted by making them choose more faithful videos for the presented criteria. The results filled in Table 1 show that the output of our module is more preferable.

4.3. Quantitative Results

The quantitative results of this study can be found in Table 2. Clip Score was measured by comparing target

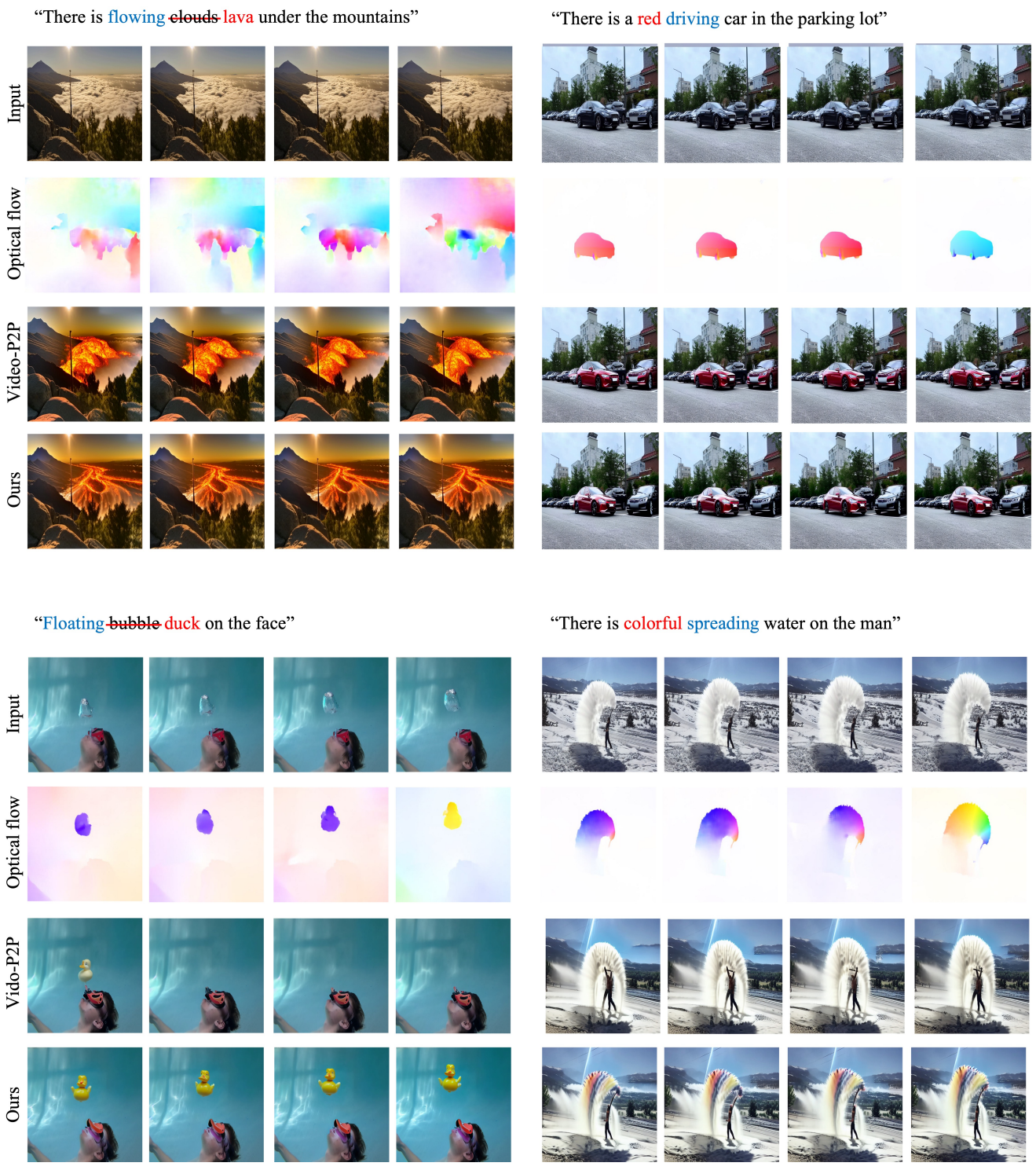


Figure 4. First experimental results of our study. It shows the result of four examples corresponding to each target prompt. The top row of each example is the input video frame, the second row is the optical flow extracted through the MMI module, the third row is the output of Video-P2P, and the last row is the result of our research. It can be seen that the result of the model receiving the motion information of the image through Optical Flow is better. It also shows good performance in editing not only fixed-shaped objects but also irregular objects (water droplets, snow, etc.).

Figure 5. Output of various methods to measure the correlation between the motion map and the attention map.

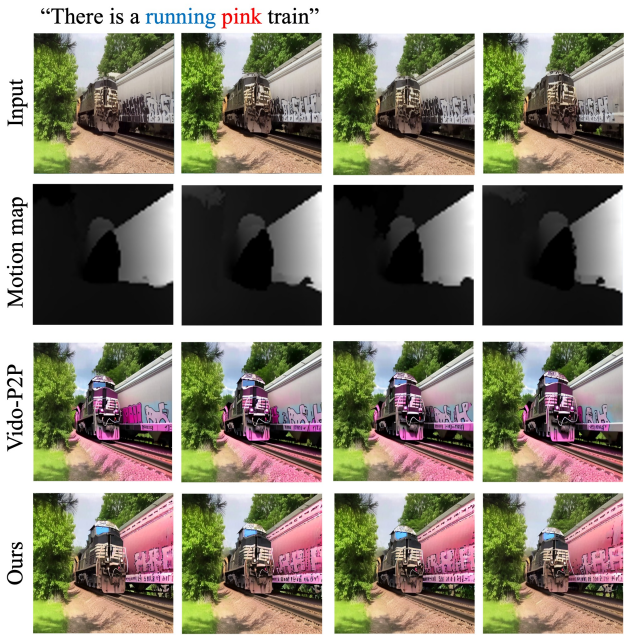


Figure 6. Editing method for objects moving in the direction specified by the user. Before editing, the user first selects one of the 8 directions.

prompt with edited image. It measures how well edit image representative the meaning of edit prompt. The results of this study confirmed that our method was better understood meaning of prompt compared to other comparison model. For Masked PSNR, the performance of our framework scored slightly higher. This means that the existing model did more edit in areas that were not intended than the model we propose. It means that it can accurately edit to the object that you want to edit, this is measured by applying a mask to the edited part.

4.4. Qualitative Results

Fig. 4 shows qualitative results of Video-P2P algorithm and our proposed MMI algorithm. Previous methods tend to produce structurally unnatural images, but our method produces structurally coherent and content-preserving frames. In 'Lava' video, only part of the cloud area of the Input frame is activated, and when edited with the video-P2P, it remains as a cloud without changing to lava. When applying ours, the corresponding area could be estimated through the optical flow of the cloud and edited the entire cloud to lava. In 'Car' video, when edited through the Video-P2P model, even though prompt contain "driving", the parked car is also caught in attention map of car, so it can be seen that some other car is edited red. Using our model, editing was applied only to moving cars that we want to refer to by estimating optical flow for moving cars through our model. In 'Bubble' video, the attention map of the bubble did not accurately represent the area because the information that "floating" was not apprehend. The editing was done only in first frame and the duck disappeared in the other frames. Through our model, the entire frame was edited through enhanced attention by estimating and injecting bubble's motion map. Editing was not made in Video-P2P due to inaccurate attention map of "spreading" not being grasped. In our model, it changed when we injected motion map.

4.5. Ablation Study

Fig. 5 is the results for different injection methods of motion map. "Directly Inject" injects the motion map directly into the attention map of the motion prompt. "correlation" injects motion map to entire prompt's attention maps through the correlation between the each word attention maps and the motion map. "Directly Inject & Correlation" is combination of above described two methods. Specifically, "Correlation" is firstly applied, and then, "Directly Injection" is applied to the motion prompt. To inject the motion map into all words, the correlation between the attention maps of the input prompt and motion map is calculated. We applied various functions of template matching that represent a correlation between the two images. The "CCOEF" applied to the first image used correlation coefficient, and the second image is the result of "CCOEF_N" that normalizes it. The third image used "SQDIFF", a sum of squared differences, and the fourth image used "SQDIFF-N", which was normalized. As you can see in Fig. 5, the most suitable function to reinforce semantically editable is seen as "CCOEFF_N", which helps to edit semantically by varying weights depending on the degree of association of the prompter word. Therefore, we choose "CCOEFF_N" in method which is same as NCC.

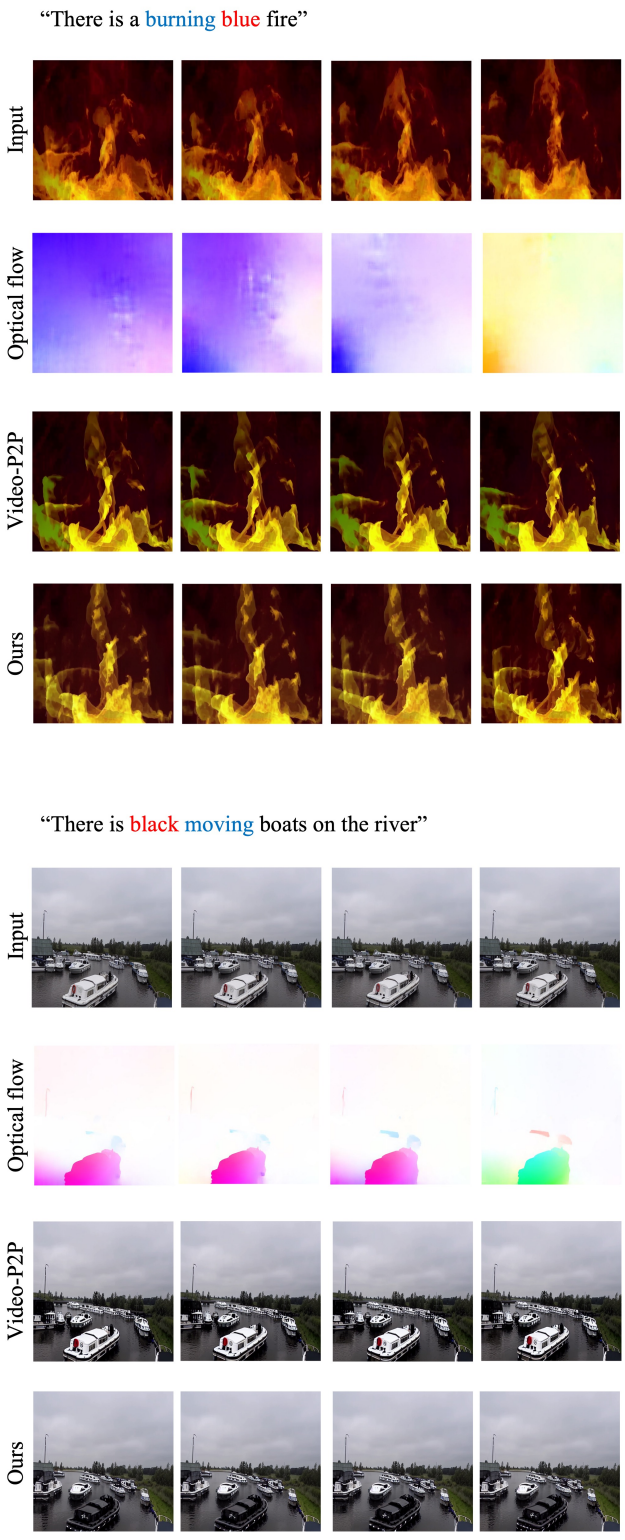


Figure 7. Video-p2p and our video editing results and estimated optical flow in inaccurate motion estimation

4.6. Application

Optical Flow Rotation Since V_{flow} has information on the magnitude of pixel movement between frames, as well as the direction in which pixel moved between frames, the user can select and edit the motion value in the desired direction. Our model allows the user to edit contents in a specific direction by rotating the optical flow V_{flow} according to the direction D provided by the user before injecting it. You can find the result in Fig. 6. Optical flow can be estimated to having various directions depending on movement, but in this study, D were dealt with only eight directions. The specific method would be explained written in the supplementary material.

4.7. Limitations

Accurate motion estimation of input video is essential for editing using optical flow. Therefore, even if optical flow is used, the bad results as shown in Fig. 7 may be obtained when it is difficult to estimate motion information from an image. The optical flow for the movement of fire could not be estimated, so there was no difference from the existing Video-P2P. In addition, in the example of boat, motion was estimated only for ships that occupy a large area of the image, and small ships were not estimated. If the estimated motion of the optical flow for input video is not accurate, it is confirmed that our model, like existing Video-P2P, is difficult to perform accurate editing.

5. Conclusion

There have been a problem that existing video editing algorithms applying an implicit method for the image diffusion model to see the relationship between frames do not understand the information between frames, 'motion'. We propose a Motion Map Injection (MMI) module to inject estimated motion map into the attention map of the image diffusion model.

Our proposed MMI module injects the motion extracted from input video by the additional network into the every attention map of each prompt estimated by the image diffusion model. To this end, the motion map is multiplied by the similarity (defined as NCC) between the attention map and the motion map. Only the attention map of the motion prompt, which is difficult to estimate due to the limitations of the image diffusion model, is replaced entirely by the estimated motion map.

Injecting motion map into attention map with our proposed MMI module makes general video editing performance improved because attention map of motion prompt become apparent. This is proven by improved evaluation metrics. In addition, selective editing according to direction of motion is enabled.

References

[1] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023.

[2] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6007–6017, 2023.

[3] Chen Henry Wu and Fernando De la Torre. Making text-to-image diffusion models zero-shot image-to-image editors by inferring" random seeds". In *NeurIPS 2022 Workshop on Score-Based Methods*.

[4] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. *arXiv preprint arXiv:2302.03027*, 2023.

[5] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18208–18218, 2022.

[6] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. *arXiv preprint arXiv:2210.11427*, 2022.

[7] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

[8] Jooyoung Choi, Yunjey Choi, Yunji Kim, Junho Kim, and Sungroh Yoon. Custom-edit: Text-guided image editing with customized diffusion models. *arXiv preprint arXiv:2305.15779*, 2023.

[9] Zhixing Zhang, Ligong Han, Arnab Ghosh, Dimitris N Metaxas, and Jian Ren. Sine: Single image editing with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6027–6037, 2023.

[10] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023.

[11] Eyal Molad, Eliahu Horwitz, Dani Valevski, Alex Rav Acha, Yossi Matias, Yael Pritch, Yaniv Leviathan, and Yedid Hoshen. Dreamix: Video diffusion models are general video editors. *arXiv preprint arXiv:2302.01329*, 2023.

[12] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Weixian Lei, Yuchao Gu, 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. *arXiv preprint arXiv:2212.11565*, 2022.

[13] 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.

[14] 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.

[15] 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.

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[17] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[18] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

[19] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.

[20] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[21] Bram Wallace, Akash Gokul, and Nikhil Naik. Edict: Exact diffusion inversion via coupled transformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22532–22541, 2023.

[22] 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.

[23] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2758–2766, 2015.

[24] Anurag Ranjan and Michael J Black. Optical flow estimation using a spatial pyramid network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4161–4170, 2017.

[25] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8934–8943, 2018.

[26] Tak-Wai Hui, Xiaoou Tang, and Chen Change Loy. Liteflownet: A lightweight convolutional neural network for optical flow estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8981–8989, 2018.

[27] 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.

[28] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, Fisher Yu, Dacheng Tao, and Andreas Geiger. Unifying flow, stereo and depth estimation. *arXiv preprint arXiv:2211.05783*, 2022.

[29] Shuyang Sun, Zhanghui Kuang, Lu Sheng, Wanli Ouyang, and Wei Zhang. Optical flow guided feature: A fast and robust motion representation for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1390–1399, 2018.

[30] Laura Sevilla-Lara, Yiyi Liao, Fatma Güney, Varun Jampani, Andreas Geiger, and Michael J Black. On the integration of optical flow and action recognition. In *Pattern Recognition: 40th German Conference, GCPR 2018, Stuttgart, Germany, October 9-12, 2018, Proceedings 40*, pages 281–297. Springer, 2019.

[31] Rui Xu, Xiaoxiao Li, Bolei Zhou, and Chen Change Loy. Deep flow-guided video inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3723–3732, 2019.

[32] Dahun Kim, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Deep video inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5792–5801, 2019.

[33] Kaidong Zhang, Jingjing Fu, and Dong Liu. Flow-guided transformer for video inpainting. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVIII*, pages 74–90. Springer, 2022.

[34] Zhigang Tu, Hongyan Li, Wei Xie, Yuanzhong Liu, Shifu Zhang, Baoxin Li, and Junsong Yuan. Optical flow for video super-resolution: a survey. *Artificial Intelligence Review*, 55(8):6505–6546, 2022.

[35] Mehdi SM Sajjadi, Raviteja Vemulapalli, and Matthew Brown. Frame-recurrent video super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6626–6634, 2018.

[36] Longguang Wang, Yulan Guo, Li Liu, Zaiping Lin, Xinpu Deng, and Wei An. Deep video super-resolution using hr optical flow estimation. *IEEE Transactions on Image Processing*, 29:4323–4336, 2020.

[37] Zhen Li, Zuo-Liang Zhu, Ling-Hao Han, Qibin Hou, Chun-Le Guo, and Ming-Ming Cheng. Amt: All-pairs multi-field transforms for efficient frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9801–9810, 2023.

[38] Wenbo Bao, Wei-Sheng Lai, Chao Ma, Xiaoyun Zhang, Zhiyong Gao, and Ming-Hsuan Yang. Depth-aware video frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3703–3712, 2019.

[39] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021.