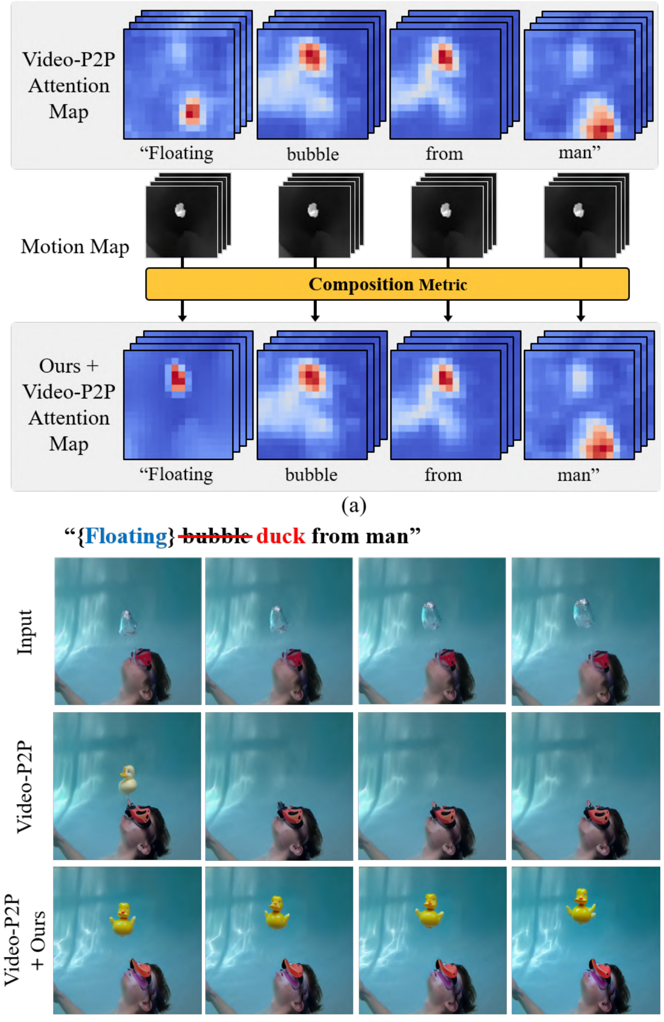
Motion-to-Attention: Attention Motion Composer using Optical Flow for Text-to-Video Editing

Seong-Hun Jeong*, Inhwan Jin*, Haesoo Choo*, Hyeonjun Na*, and Kyeongbo Kong

***Abstract*—Recent text-guided video editing research attempts to expand from image to video based on the text-guided image editing model. To this end, most researches focus on achieving temporal consistency between frames as a primary challenge in text-guided video editing. However, despite their efforts, the editability is still limited when there is a prompt indicating motion, such as 'floating'. In our experiment, we found that this phenomenon was due to the inaccurate attention map of the motion prompt. In this paper, we suggest the Motion-to-Attention (M2A) module to carry out precise video editing by explicitly taking motion into account. First, we convert the optical flow extracted from the video into a motion map. During the conversion, users can selectively apply directional information to extract the motion map. The proposed M2A module uses two methods: "Attention-Motion Swap," which directly replaces the motion map with the attention map, and "Attention-Motion Fusion," which uses the similarity between the motion map and attention map, measured by a Composition Metric, as a weight to enhance the attention map using the motion map. When comparing the results using various quality measurement metrics, the output of the video editing models that applied our proposed module were the best. Notably, our M2A module is effectively applied to the models performing video editing by controlling the attention map, enhancing editability. Extensive experimental results are https://currycurry915.github.io/Motion-to-Attention/**

***Index Terms*—Video editing, diffusion, optical flow, and vision language.**



(a)

(b)

Fig. 1. The existing T2V-model failed to estimate an accurate attention map for the motion prompt {Floating}, which resulted in a decrease in editability. (a) is a figure comparing the editability of Video-P2P and adding the proposed module to Video-P2P for input video. The proposed module improves the editability of existing video editing models through accurate estimation of attention maps. (b) briefly explains the method of enhancing the attention map by applying the proposed module to address the limitations that the existing T2V model cannot accurately generate.

I. Introduction

UNPRECEDENTED advancements in image generation and editing have recently been made thanks to research on text-guided diffusion models and large-scale language models. Unlike previous research [1], [2] that could only edit images globally using deep neural networks, recent research has been focusing extensively on editing images using only the prompts provided by the user [3]–[7], referencing the established investigations in this area. Among them, Prompt-to-Prompt [3] manipulates the attention map for each prompt, enabling semantic and local editing of images using only the prompt without external information such as a mask.

Along with these advancements, research on text-guided video editing is also increasing [8]–[11]. However, there is a lack of sufficient text-video pair training data to train large-scale video generation models. Therefore, most of the text-

This research was supported in part by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2020R1A6A3A01098940, 2021R1I1A1A01051225) and regional broadcasting development support project funded by the Foundation for Broadcast Culture.

Seong-Hun Jeong is with the Graduate School of Engineering, Department of Electrical and Electronic Engineering, Major of Communications, Electromagnetic Wave, Signal Processing, Pusan National University, Busan, South Korea (e-mail: tlqwkrk915@pusan.ac.kr).

Inhwan Jin, Haesoo Choo and Hyeonjun Na are with Pukyong National University, Busan, South Korea. Inhwan Jin and Hyeonjun Na are with the major of Journalism and Information (e-mail: bds06081@naver.com, wewe1st@naver.com). Haesoo Choo is with a major in Korean language and literature (e-mail: tndi28@naver.com).

Kyeongbo Kong, the corresponding author, is with the Department of Electrical and Electronic Engineering, Pusan National University, Busan, South Korea (e-mail: kbkong@pusan.ac.kr).

The symbol * implies that Seong-Hun Jeong, Inhwan Jin, and Haesoo Choo have contributed equally to this work.

guided video editing research extends image editing models to video editing models using the zero-shot approach [10], [12], [13] or by fine-tuning text-guided image editing models [8]. In contrast to image, video consists of multiple frames. Therefore, recent researches [10]–[12], [14]–[17] focus on achieving temporal consistency between frames as the primary challenge in text-guided video editing.

Despite the concerted efforts, as depicted in Fig. 1(a), the attempt to edit "bubble" into "duck" using existing methodologies in video editing was unsuccessful, resulting in the complete disappearance of the "bubble". To find the reason of this phenomenon, we delved the estimated results of the attention map. In this process, as can be observed in Fig. 1-(b), we found inaccurate estimation of the attention map for prompt representing motion, such as 'floating'. We speculate that this is due to the direct extension of the existing text-guided image model, which does not sufficiently learn about prompts indicating motion. Specifically, the inaccuracy of this attention map reduces the editability of not only the motion prompt but also the moving object.

In this paper, we introduce a method to enhance the attention map in text-guided video editing by extracting motion information from video. We essentially use optical flow [18], which is widely used in various video tasks [19], [20], to utilize motion information. Optical flow method can extract highly accurate motion information by estimating the change in pixels between video frames. We generate a motion map for each frame using the estimated optical flow. Additionally, the estimated optical flow can be used to specifically erase one of the various directions of motion, and the motion map generated through this can be used to perform video editing for a specific direction, providing improved editability.

Our key contribution is the effective incorporation of the generated motion map into the attention map. For this purpose, this paper proposes an **Motion-to-Attention** module consisting of "Attention-Motion Swap" and "Attention-Motion Fusion." First, Attention-Motion Swap is a method of correcting incorrectly estimated attention maps by swapping the motion map with the attention map of the motion prompt. Since the attention map of a motion prompt that is not accurately estimated can lower the editability of a text-guided video editing model, we directly swap motion information by swapping it with a motion map that contains accurate motion information.

Remarkably, enhancing the attention map for motion prompts alone significantly improved the overall video editability. In addition, to enhance the attention map for the entire prompt, we inject the motion map into the attention map with weights determined based on the Composition Metric between the motion map and the attention map. Since one attention map is semantically related to other attention maps, directly swapping without considering the interrelation between attention maps has limitations in enhancing editability.

Therefore, to enhance while considering the relationships between attention maps, we propose a "Attention-Motion Fusion" method that utilizes the similarity between the attention map and motion map as a weight. At this point, we compared a total of eight Composition Metric to calculate the similarity between the motion map and the attention map. In our experiments, Mutual Information (MI) [21] showed the highest performance. This process is observable in Fig. 1-(b), and the enhanced editing outcomes, in comparison to conventional text-guided video editing models, are evident in the last line of Fig. 1-(a).

The contributions of our paper are twofold. First, we find that inaccurately estimating the attention map for prompts indicating essential movements in text-guided video editing reduces video editability. This study raises the necessity of enhancing the attention map in video editing and is the first to introduce a method for enhancing the attention map of video through motion information estimated using existing optical flow. Second, when applied to the existing attention-based text-guided video editing models, we confirmed that all output results were improved. Our module will continue to be applied to future research in attention-based text-guided video editing models, providing enhanced editability for video editing.

II. Related Work

A. Text-Guided Editing

Unlike the traditional field of image processing research [22], [23], which relied on relatively simple deep neural networks to remove general noise or other artifacts from images, recent research of text-guided image editing [24] uses a variety of generative models. The diffusion model [25], [26], which is among the most prevalently used generative models in recent times, operates by either introducing noise into or eradicating noise from the input image. This process generates an output image with a probability distribution that closely resembles that of the input image. The results of recent studies on diffusion models are impressive, surpassing the achievements of previous research [27], [28] that used generative models such as Generative Adversarial Nets (GAN) [29]. Text-guided image editing models like DALL-E2 [30] and Imagen [31], and stable diffusion [32] show the outcomes of high-quality image editing based on this diffusion model. Research on video editing using other generative models [33]–[35] has been conducted, although it is in the minority, the main focus being research on video editing using the diffusion model. Subsequent video editing research diverges into two primary paths: 1) P2P based Model: Because P2P can perform semantically meaningful editing by controlling the attention map, numerous models have been developed that bring P2P's editing capability to the video domain. 2) Non-P2P based Model: These models do not control the attention map, a key internal feature of diffusion models, but instead perform editing by utilizing various decomposed elements.

Among P2P based Models, Video-P2P divides its framework into two branches—one for unchanged parts and another for edited parts—and incorporates each attention map to enable detailed editing. Alongside the aforementioned works, vid2vid-zero [10] achieves stable video editing and reconstruction by integrating cross-frame attention into the U-Net structure of an existing diffusion model. Additionally, FateZero [12] and InFusion [13] are zero-shot based and maintain the video's temporal consistency through the enforcement of

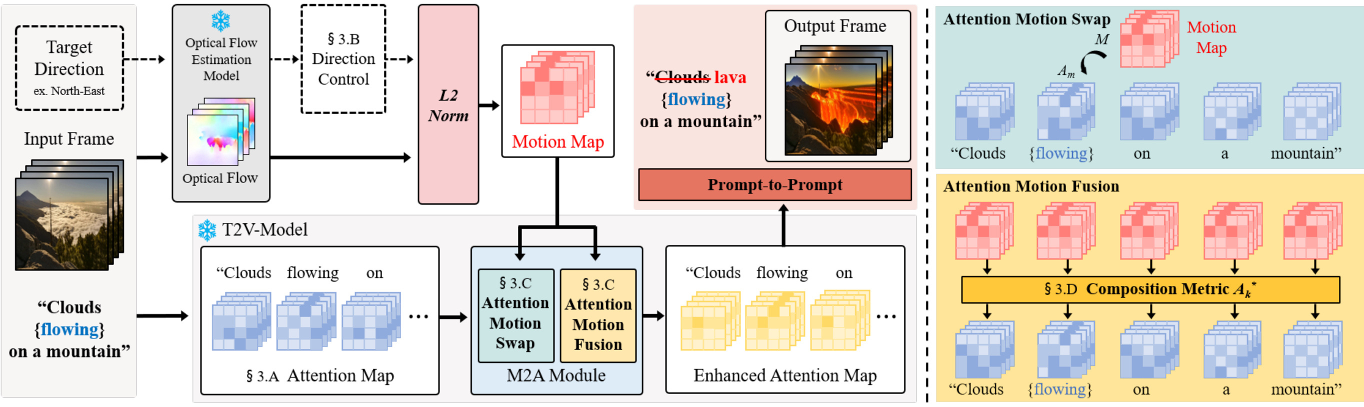


Fig. 2. Overall framework. First, the T2V-Model generates an attention map by receiving video and prompts as input. At the same time, it receives video frames and creates a motion map. In this process, a motion map can be created by estimating the entire motion for the frame or the motion for the target direction. Then, the motion map is injected into the attention map of the T2V-Model in two ways from the Motion-to-Attention (M2A) module: 'Attention Motion Swap' and 'Attention Motion Fusion'. After that, text-to-video editing is performed using the attention map that includes video motion information.

the attention map. Pix2video [36] focuses on identifying the structure and appearance of the frame through self-attention within the diffusion model, demonstrating how to incorporate a layer that can maintain temporal consistency, such as cross-frame attention.

Among Non-P2P based Models, ControlVideo [37] maintains temporal consistency in a manner similar to Pix2video [36], but it employs the editing method of the existing ControlNet [38] for performing edits. VideoControlNet [39], MeDM [40], and FLATTEN [41] maintain temporal consistency by leveraging frames generated from video motion information obtained via optical flow. TokenFlow [14] improves the temporal consistency of the video by enforcing the semantic correspondences of diffusion features, recognizing that the internal representation of the diffusion model exhibits similar properties across frames. Control-A-Video [15] integrates motion and content priors, introducing motion-adaptive noise initialization strategies to enhance the consistency and quality of the video. Rerender is a model that performs global editing rather than editing objects within the video and uses optical flow to maintain the shape of the edited objects.

In this study, we apply a proposed module to the P2P based Model and compare the results before and after its implementation. Furthermore, we compare the performance of the P2P based model with the proposed module to that of Non-P2P based models. In Sec IV, the results of our experiments indicate that the P2P based model with the proposed module applied shows the best performance.

B. Optical Flow Estimation

Estimating the motion of objects in a video sequence is the major goal of a computer vision task. The first fully convolutional neural network to estimate optical flow was called FlowNet [20]. Subsequently, a series of works, including SpyNet [42], PWC-Net [43], LiteFlowNet [44], RecSPy [45], and RAFT [46] were proposed to reduce computational costs by using a coarse-to-fine and iterative estimation method. Recently, GMFlow [47] was proposed, which performed global matching with a Transformer to produce highly accurate results without needing a lot of refinement. We generate the optical flow of the video using the most recent optical flow estimation model, UniMatch [18].

The proposed module is not dependent on UniMatch, can use various optical flow estimation models, and has stable performance. To demonstrate this, we conducted an experiment about UniMatch [18] and RAFT [46], another optical flow estimation models. Through this experiment, we confirm that there is no significant correlation between the accuracy of the optical flow and the performance of our module.

C. Similarity Measurement Methods

Template matching is a task that finds a matching area in an image by comparing it with a template image [48]. Early template matching was implemented using simple similarity metrics such as Cross-Correlation (CC) [49], [50]. Cross-Coefficient(CC), which is calculated by considering the mean and standard deviation, is also used for more detailed analysis because it is resistant to change or contrast. Squared Differences [51], which squares the differences in pixel values at each location and then calculates the sum of these values over all locations, is also commonly used in template matching.

Unlike template matching, which finds the location of a specific pattern within an image, there are also methods that focus on measuring similarity between images. Spectral Angle Mapper (SAM) [52] calculates the spectral angle between pixels to measure similarity by how similar the two spectra are oriented. Mutual Information (MI) [21], which measures the similarity between two images, considers the pixel values of the images as random variables and measures how much the values change together. In this study, we compared the performance using various metrics, including those mentioned, to determine how similar motion maps and attention maps are. We empirically confirm that MI shows the most stable results. Detailed information is provided in Sec. III-D

III. Proposed Method

In this paper, we propose the Motion-to-Attention (M2A) Module to improve the editability of the text-to-video (T2V) editing model by injecting motion information into the attenion map. Our method for enhancing the attention map through motion information proposes two modes that leverage the magnitude and direction information of optical flow. The first method enhances the attention map by infusing it with the entire motion information of the video, utilizing only the magnitude information of optical flow. The second mode suggests selectively using information about the direction specified by the user, by exploiting the direction information of optical flow.

Before delving into the specifics, we first provide a overview of our framework depicted in Fig. 2. The M2A module is built into our framework as an addition to the T2V model. Let the input video be $\mathcal{V}$, which consists of frames. As in the Prompt-to-Prompt (P2P) [3] setting, we define the source prompt as $\mathcal{P}$ and the target prompt as $\mathcal{P}^*$. In source prompt, the prompt containing motion information of the video is called motion prompt $\mathcal{P}_\mathcal{M}$ (e.g. 'running', 'moving'). $\mathcal{P}_\mathcal{M}$ within the $\mathcal{P}$ are indicated by the user using **{}**, such as "a {moving} car". And when the user provides direction information as input, the direction prompt is displayed as **[]**, such as [west].

Additionally, we propose that processes optical flow preprocessing based on the direction information provided by the user, to enable editing objects moving in a specific direction. This method is performed only when the user provides a direction. If no direction is provided, editing is performed on the entire area with motion. Our proposed M2A module enhances the attention map by utilizing optical flow, information external to the T2V-model.

In Sec III-A, we introduce the attention map as prior knowledge. Additionally, we present the editing method employed in P2P. In Sec III-B, we describe the method of extracting a motion map from the optical flow. Additionally, we explain how to specify the areas moving in the direction input by the user and extract motion maps only for those corresponding areas. In Sec III-C, we explain the two methods of the M2A module, namely "Attention-Motion Swap" and "Attention-Motion Fusion". In Sec III-D, we explain various similarity measurement methods for calculating the correlation score used in "attention-motion fusion".

A. Preliminary

Attention Map Prompt-to-Prompt (P2P) [3] is a text-guided image editing model that performs editing by manipulating the attention map in the image domain. The attention map estimated through P2P's cross-attention layer visually represents the correlation between a word and an image. The attention map A can be calculated using the following formula:

$$\mathcal{A} = \mathrm{Softmax}\left(\frac{\mathcal{Q}\mathcal{K}^\mathcal{T}}{\sqrt{d}}\right), \tag{1}$$

The spatial features of the image are transformed into the query matrix $\mathcal{Q}$, and text embeddings are transformed into the key matrix $\mathcal{K}$ and value matrix $\mathcal{V}$. To calculate the similarity between the spatial features $\mathcal{Q}$ of the image and the text embeddings $\mathcal{K}$, the two matrices are multiplied. To align the dimensions of the matrices, a transpose is performed on the $\mathcal{K}$ matrix. Here, d represents the dimensionality of the query and key. Afterwards, each pixel value is converted to a probability value by applying the softmax function.

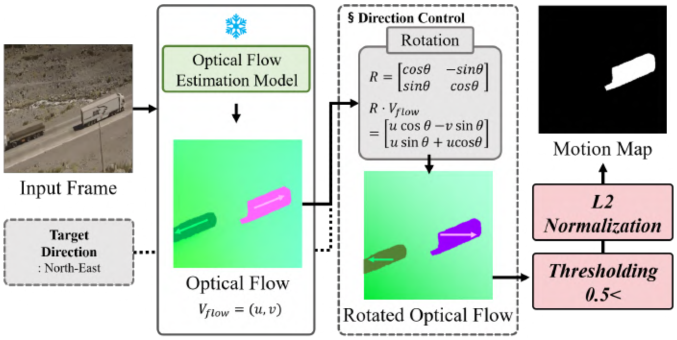

Input Frame
Target
Direction
: North-East
Optical Flow
Estimation Model
Optical Flow
$V_{flow} = (u, v)$
§ Direction Control
Rotation
$R = \begin{bmatrix} cos\theta & -sin\theta \\ sin\theta & cos\theta \end{bmatrix}$
$R \cdot V_{flow} = \begin{bmatrix} u\cos\theta - v\sin\theta \\ u\sin\theta + ucos\theta \end{bmatrix}$
Rotated Optical Flow
Motion Map
L2
Normalization
Thresholding
0.5<

Fig. 3. The process of Direction Control. A rotation matrix is applied to the optical flow to rotate vectors, and then only those vectors with positive values along the x-axis are selected for conversion into a motion map. This method allows for the editing of objects and pixels that move in a direction desired by the user.

Editing Method of Prompt-to-Prompt Recently, P2P [3] has been extended to the video domain, and research has been focused on various methods to maintain temporal consistency [8], [10]–[13], [36], [37]. However, the fundamental editing methods still utilize the approach of P2P. In this context, we describe the process of editing an image by **replacing** the words in the source prompt using the edit function $Edit(\cdot)$.

$$Edit(\mathcal{A}_t, \mathcal{A}_t^*, t) := \begin{cases} \mathcal{A}_t^* & \text{if } t < \tau \\ \mathcal{A}_t & \text{otherwise} \end{cases}, \tag{2}$$

where t refers to the time step used in the diffusion model within P2P, and τ is a parameter that determines when the replacing operation is applied. Through the above method, attention map $\mathcal{A}$ of source prompt $\mathcal{P}$ is replaced with attention map $\mathcal{A}^*$ of target prompt $\mathcal{P}^*$.

Otherwise, users want to change the style of the image or attribute of certain object. For example, $\mathcal{P}$ = "a car" to "a red car". In this case called **'prompt refinement'**, an alignment function $\mathcal{L}$ is used in order to preserve the common parts, which matches the index of between $\mathcal{P}$ and target prompt.

$$Edit\left(\mathcal{A}_t, \mathcal{A}_t^*, t\right)_{i,j} := \begin{cases} \left(\mathcal{A}_t^*\right)_{i,j} & \text{if } \mathcal{L}\left(j\right) = \text{None} \\ \left(\mathcal{A}_t\right)_{i,\mathcal{L}(j)} & \text{otherwise,} \end{cases} \tag{3}$$

where index i corresponds to pixel value, and j corresponds to word index.

B. Motion Map Extraction

We obtain the optical flow $\mathcal{V}_{flow} = (u, v)$ through the pre-trained optical flow estimation model [18] where u and v represent the movements in the x-axis and y-axis directions of the optical flow vector, respectively. In this paper, the goal is to enhance the attention map utilizing this optical flow. We extract a motion map with magnitude values by applying L2 norm to the estimated optical flow. The formula for obtaining the motion map $\mathcal{M}$ by applying L2 normalization to the optical flow $\mathcal{V}_{flow} = (u, v)$ is as follows:

$$\mathcal{M} = \sqrt{u^2 + v^2}. \tag{4}$$

Direction Control In this study, we propose a method for editing objects moving in the direction D provided by the user, prior to applying L2 normalization to the optical flow, as illustrated in Figure 3. The user can choose to edit the motion value in the desired direction because $\mathcal{V}_{flow}$ contains information on both the magnitude and direction of pixel movement between frames. By rotating the optical flow $\mathcal{V}_{flow}$ in accordance with the direction $\mathcal{D}$, our module enables the user to edit content in a particular direction. The rotated optical flow vector $\mathcal{V}'_{flow} = (u', v')$ is calculated as follows:

$$\begin{aligned} u' &= u\cos\theta - v\sin\theta \\ v' &= u\sin\theta + v\cos\theta, \end{aligned} \tag{5}$$

where θ is the rotate angle between the direction provided by the user and the x-axis.

Optical flow is thought to have any directions depending on movement, but in this study, only eight directions were used for $\mathcal{D}$. Our proposed model receives input from the user in eight directions, including the four cardinal directions of East, West, South, and North, as well as the intercardinal directions of Northeast (NE), Southeast (SE), Southwest (SW), and Northwest (NW), which are derived from combinations of the four directions in the 2D coordinate system.

After rotating the optical flow for the user-provided direction $\mathcal{D}$, only the region of pixels with positive directional motion of the x axis in the rotated coordinate system is specified. After applying L2 normalization to generate a motion map for the specific region and then applying it to the attention map, it was confirmed that only the designated area was edited.

C. Motion-to-Attention Module

In this section, we describe two methodologies the M2A module performs the attention map: "attention-motion swap" and "attention-motion fusion". The pseudo algorithm of our proposed M2A module is shown in Alg. 1.

Attention-Motion Swap In Fig. 1, we found that the existing T2V-model does not perform accurate estimation of the attention map $\mathcal{A}_{\mathcal{M}}$ of the motion prompt $\mathcal{P}_{\mathcal{M}}$. Moreover, experimental results confirmed that $\mathcal{A}_{\mathcal{M}}$ negatively influences the editability of the T2V model. The proposed M2A module directly swaps the inaccurately estimated attention map $\mathcal{A}_{\mathcal{M}}$ with the motion map $\mathcal{M}$ representing the motion information of the video. The formula is as follows:

$$\mathcal{A}_{\mathcal{M}} = \lambda \cdot \mathcal{M}, \quad \text{if } \mathcal{P} = \mathcal{P}_{\mathcal{M}}, \tag{6}$$

where λ is a hyperparameter for motion map injection rate. When employing this approach, experimental results confirmed the improvement in the editability of the existing T2V model, as illustrated in Fig. 6 "Only Attention-Motion Swap." However, the method of motion swapping, without considering the interrelation with other words, has limitations in enhancing editability. Therefore, a method is needed that considers the relationships between different attention maps while enhancing them.

Attention-Motion Fusion The proposed M2A module employs a weighted sum method using the correlation score $\mathcal{C}$ between attention maps $\mathcal{A}$ and motion map $\mathcal{M}$ to enhance while considering the relationships between different attention maps. First, the correlation score $\mathcal{C}_k$ between the attention map $\mathcal{A}_k$ corresponding to the k-th source prompt and the motion map $\mathcal{M}$ is computed using the similarity measurement method.

The calculated correlation score $\mathcal{C}_k$ is used as a weight to consider the relationships between attention maps of the entire prompt, and the enhancement method is as follows:

$$\mathcal{A}^*_k = \mathcal{A}_k + \lambda \cdot \frac{(\mathcal{C}_k \cdot \mathcal{M})}{t}, \tag{7}$$

where t denotes the denoising step used in the diffusion model of the T2V model. Through this method, attention maps that are similar to the motion map are further emphasized with motion information, while dissimilar attention maps retain their original values. As a result, the editability of the T2V-model can be improved.

Algorithm 1 M2A module

```
Input: attention maps of entire source prompt $\mathcal{A}$,
    motion map $\mathcal{M}$,
    number of source prompt $i$,
    index of motion prompt in source prompt $j$,
    correlation score $\mathcal{C}$,
    hyperparameter for motion map injection rate $\lambda$,
    denoising timestep in diffusion model $t$
Output: Enhanced attention of entire source prompt $\mathcal{A}^*$ # Attention-Motion Fusion
for k = 0,1,...,i do
    $\mathcal{A}^*_k = \mathcal{A}_k + \lambda \cdot \frac{\mathcal{C}_k \cdot \mathcal{M}}{t}$
end for
# Attention-Motion Swap
if $i = j$ then
    $\mathcal{A}^*_j = \lambda \cdot \mathcal{M}$
end if
return $\mathcal{A}^*$;
```

However, even when the Attention-Motion Fusion method using correlation score is used alone, the results are not so good. The incorrectly estimated attention map $\mathcal{A}_{\mathcal{M}}$ may persist, and even when enhanced, it may not be significantly influenced due to its low correlation score with the motion map $\mathcal{M}$ as in Alg. 1. Therefore, the proposed M2A module enhances the attention map by combining both the "Attention-Motion Swap" method and the "Attention-Motion Fusion" method based on the correlation score, thereby improving the editability of the T2V model.

D. Composition Metric

In this study, various methods are utilized to compare and measure the correlation score C between the attention map $\mathcal{A}$ and the motion map $\mathcal{M}$ in Attention-Motion Fusion.

1) ***Conventional Similarity Metric [48]:*** In this study, we utilize conventional metric of measuring the similarity between two images in the field of computer vision to assess the similarity between attention maps and motion maps. These methods quantitatively evaluate how similar two images are at the pixel-level.

a) ***Squared Differences [51]:*** Squared Differences involves calculating the square of the difference in pixel values between the base image and the target image, and the accuracy of the match is evaluated through the sum of these squares. The correlation score C calculated using the squared difference indicates that the lower the value, the more similar the two images are. The formula for calculating the correlation score C between the $\mathcal{A}$ and $\mathcal{M}$ through squared difference is as follows:

$$C = \sum_{x,y}(\mathcal{A}(x,y) - \mathcal{M}(x,y))^2. \tag{8}$$

where x, y represent the pixel coordinates of the two images.

In this study, due to the difference in the scale of values between the $\mathcal{A}$ and $\mathcal{M}$, the normalized squared difference is utilized for consistent comparison. The normalized squared difference method adjusts the range of values to [0, 1], enabling consistent comparison across different conditions. The formula is as follows:

$$C = \frac{\sum_{x,y}(\mathcal{A}(x,y) - \mathcal{M}(x,y))^2}{\sum_{x,y}(\mathcal{A}(x,y)^2 \cdot \sum_{x,y}(\mathcal{M}(x,y)^2}. \tag{9}$$

*b) **Cross Correlation [49]:*** This metric measures the similarity between two images by directly multiplying the pixel values of each image and then calculating the sum of these products to produce a correlation score. In this study, the formula for calculating the Cross Correlation to measure the similarity between the attention map $\mathcal{A}$ and the motion map $\mathcal{M}$ is as follows:

$$C = \sum_{x,y}(\mathcal{A}(x,y) \cdot \mathcal{M}(x,y)). \tag{10}$$

However, Cross Correlation can lead to significant variations in the score due to changes in brightness or contrast, even if the two images being compared are exactly matched. To address this issue, this study also employs the normalized cross correlation method, which is robust against changes in brightness and contrast. The formula is as follows:

$$C = \frac{\sum_{x,y}(\mathcal{A}(x,y) \cdot \mathcal{M}(x,y))}{\sqrt{\sum_{x,y}\mathcal{A}(x,y)^2 \cdot \sum_{x,y}\mathcal{M}(x,y)^2}}. \tag{11}$$

*c) **Correlation Coefficient [53]:*** This metric is utilized to assess the strength and direction of the linear relationship between two sets of data. In this study, it represents the similarity between the attention map $\mathcal{A}$ and the motion map $\mathcal{M}$. The formula is as follows:

$$C = \sum_{x,y}((\mathcal{A}(x,y) - \mu_{\mathcal{A}}) \cdot (\mathcal{M}(x,y) - \mu_{\mathcal{M}})). \tag{12}$$

where μ represents the average of the pixel values of the attention map and the motion map.

In this study, the normalized correlation coefficient is also employed to measure the correlation score C, in order to precisely match the scale of the attention map and the motion map. The formula is as follows:

$$C = \frac{\sum_{x,y}((\mathcal{A}(x,y) - \mu_{\mathcal{A}}) \cdot (\mathcal{M}(x,y) - \mu_{\mathcal{M}}))}{\sqrt{\sum_{x,y}((\mathcal{A}(x,y) - \mu_{\mathcal{A}})^2 \cdot \sum_{x,y}(\mathcal{M}(x,y) - \mu_{\mathcal{M}})^2)}}. \tag{13}$$

*2) **Spectral Angle Mapper [52]:*** This metric, used to measure the similarity between pixels of two images, evaluates the degree of similarity in the spectra of the pixels in the images. The spectrum of a pixel is represented as an n-dimensional vector, and the spectral angle between two pixels is calculated to evaluate their similarity. This angle indicates the extent to which the two spectra are oriented in similar directions, with a smaller angle signifying greater similarity between the spectra.

If a spectral set of pixels for an image is given as $X = \{x_1, x_2, \ldots, x_n\} \subseteq \mathbb{R}^q$, then the pixel spectral set of the image to be compared is $r = \{r_1, r_2, \ldots, r_c\} \subseteq \mathbb{R}^q$ where x and r are non-zero vectors. Here, q represents the number of spectral bands, n represents the number of pixels, and c represents the number of reference spectra. The angle $\theta_{c\times n} = \{\theta_{ki}\}(k = 1, \ldots, c,$ and $i = 1, \ldots, n)$ between x_i and r_k is defined by the following formula:

$$\begin{aligned}\theta_{ki} &= \cos^{-1}(\langle x_i \cdot r_k\rangle/(||x_i|| \cdot ||r_k||)) \\ &= \cos^{-1}\Bigg(\sum_{j=1}^{q}(x_{ij} \cdot r_{kj}/\Bigg(\sqrt{\sum_{j=1}^{q}x_{ij}^2} \cdot \sqrt{\sum_{j=1}^{q}r_{kj}^2}\Bigg)\Bigg),\end{aligned} \tag{14}$$

where θ_{ki} lies within the range $[0, \frac{\pi}{2}]$, and $\langle \mathbf{x}_i \cdot \mathbf{r}_k\rangle$ represents the dot product of $\mathbf{x}_i$ and $\mathbf{r}_k$. In this study, the spectral angle $\theta_{\mathcal{M}\mathcal{A}}$ between the extracted $\mathcal{A}$ and $\mathcal{M}$ is used as the correlation score C.

*3) **Mutual Information [21]:*** Mutual information is a measure of the statistical correlation between two random variables, and in the image domain, the pixel values of two images can be set as random variables. The mutual information between attention map $\mathcal{A}$ and motion map $\mathcal{M}$ can be defined as follows:

$$MI(\mathcal{A}, \mathcal{M}) = H(\mathcal{A}) + H(\mathcal{M}) - H(\mathcal{A}, \mathcal{M}), \tag{15}$$

where $H(\mathcal{A})$ and $H(\mathcal{M})$ are the entropies [54] of $\mathcal{A}$ and $\mathcal{M}$, respectively, and $H(\mathcal{A}, \mathcal{M})$ is the entropy for the joint probability distribution of the $\mathcal{A}$ and $\mathcal{M}$, defined as follows:

$$\begin{aligned}H(\mathcal{A}) &= -\sum_{a} p_{\mathcal{A}}(a)\log p_{\mathcal{A}}(a), \\ H(\mathcal{A}, \mathcal{M}) &= -\sum_{a,m} P_{\mathcal{A}\mathcal{M}}(a,m)\log p_{\mathcal{A}\mathcal{M}}(a,m).\end{aligned} \tag{16}$$

where $p_{\mathcal{A}}(a)$ is the probability distribution of $\mathcal{A}$, and $P_{\mathcal{A}\mathcal{M}}$ represents the joint distribution of $\mathcal{A}$ and $\mathcal{M}$. Entropy is also expressed as the amount of information, and mutual information signifies the common information between two random variables. When $\mathcal{A}$ and $\mathcal{M}$ match, the statistical correlation of their brightness values reaches its maximum, and thus, the mutual information, which is the amount of information common to both images, also becomes maximal. In this study, since the value ranges of the $\mathcal{A}$ and $\mathcal{M}$ being compared are different, mutual information is normalized using the joint probability distribution. The formula is as follows:

$$NMI(\mathcal{A}, \mathcal{M}) = \frac{H(\mathcal{A}) + H(\mathcal{M})}{H(\mathcal{A}, \mathcal{M})}. \tag{17}$$

In this study, the normalized mutual information

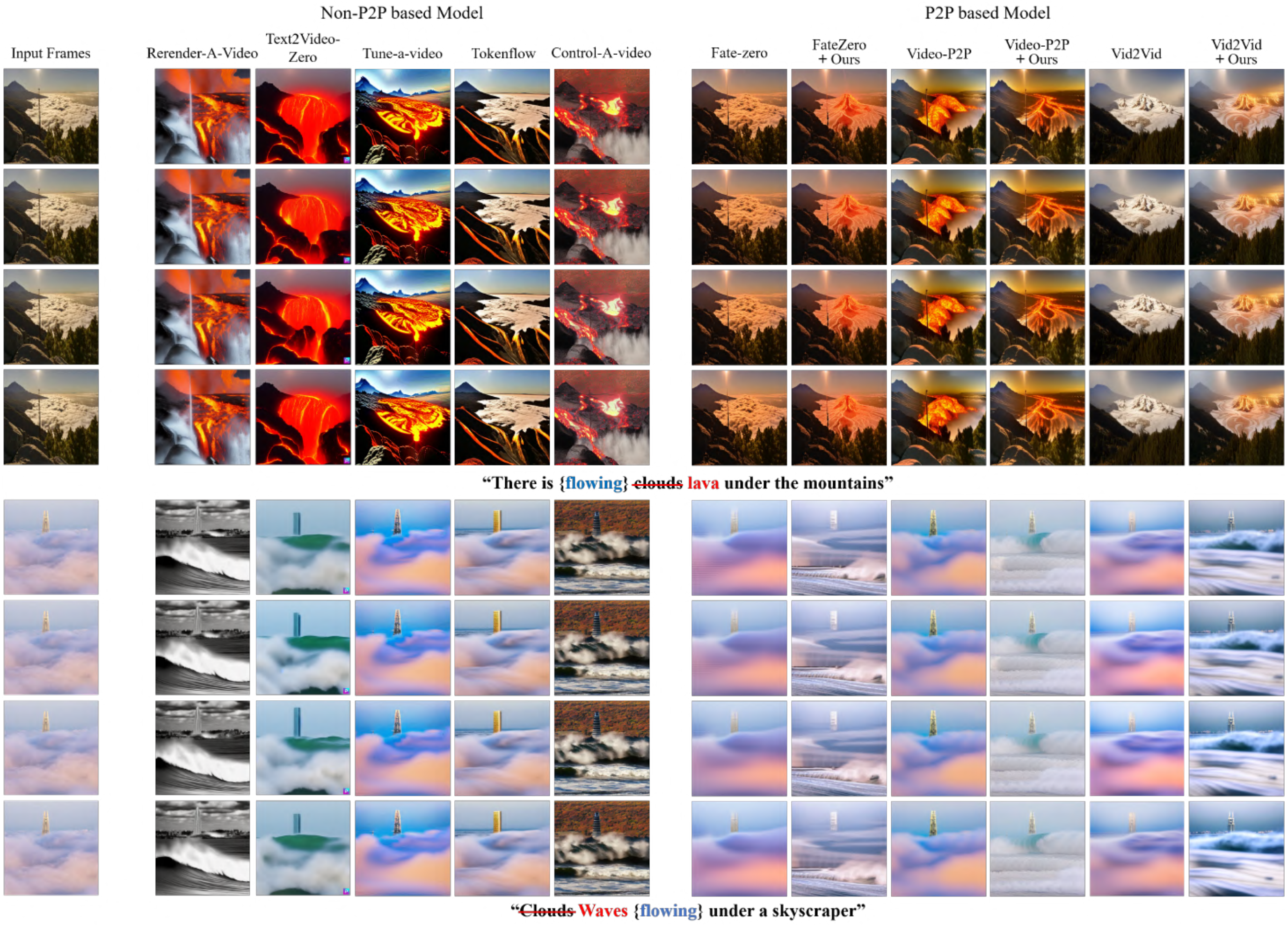


Fig. 4. Qualitative results of our study. It shows the result of two examples corresponding to each target prompt. The experimental results were distinctly divided into P2P based models and non-P2P based models. While editing was performed globally on the non-P2P based models, the application of the proposed M2A module to the P2P based models enabled precise targeting and editing of areas corresponding to the target prompt. It also shows good performance in editing not only fixed-shaped objects but also irregular objects (e.g. clouds).Additional results can be found in the supplementary materials.

$NMI(\mathcal{A}, \mathcal{M})$ between the $\mathcal{A}$ and $\mathcal{M}$ is used as the correlation score C.

IV. Experiments

A. Experimental Setup

Baseline Model We used the P2P-based editing method, which intuitively edits videos only with text by manipulating a cross-attention map, as our baseline in our module: FateZero [12], Video-P2P [11], vid2vid [10]. Additionally, we compared other text-guided video editing methods with the P2P-based model and our module: Tune-a-video [9], control-a-video [15], token-flow [14]. We compared six text-guided video editing models to demonstrate the applicability of our M2A module to P2P-based video editing models and to demonstrate its editability compared to state-of-the-art methods and concurrent works.

Dataset The Davis video dataset, which has been used for many video tasks, was used in the experiments. Additional experiments were conducted using collected YouTube videos.

Implementation details We used RTX 3090 GPUs in the experiment, and we set the image resolution to 512 × 512 as in the existing FateZero [12]. The number of video frames was set to 4 because this number is sufficient to demonstrate how well our method achieves our goal. The optical flow was extracted utilizing the UniMatch [18] model.

Evaluation Metrics Textual similarity and region preservation were assessed using trained CLIP model [55], as well as masked PSNR [11]. 'Frame-Acc' assesses the similarity between the edited frame and both the target prompt and the source prompt on a frame-wise basis, then converts into a probability score. A higher score indicates a higher similarity to the target prompt compared to the source prompt. The metrics used for assessing the level of structure preservation for the unintended region to be edited were the masked PSNR [11]. In this experiment, we regarded outside of the editing part as a mask. We used the No-Reference Image Quality Assessment methods, BRISQUE [56] to evaluate the quality of the edited videos.

TABLE I
PERFORMANCE OF THE PROPOSED MODULE MEASURED BY CLIP SCORE [55], MASKED PSNR [11], AND BRISQUE [56] METRICS. IN THIS STUDY, WE CONDUCTED EXPERIMENTS DISTINGUISHING BETWEEN P2P BASED MODELS AND NON-P2P BASED MODELS. FOR THE P2P BASED MODELS, WE USED FATEZERO [12], VIDEO-P2P [11], AND VID2VID-ZERO [10], AND COMPARED THEIR PERFORMANCE USING METRICS AFTER APPLYING OUR PROPOSED MODULE TO THE RESULTS OBTAINED FROM THESE MODELS. FOR THE NON-P2P BASED MODELS, WE UTILIZED TUNE-A-VIDEO [9], TOKENFLOW [14], CONTROL-A-VIDEO [15], TEXT2VIDEO-ZERO [16], AND RERENDER-A-VIDEO [17].

	P2P based Model						Non-P2P based Model				
	FateZero [12]		Video-P2P [11]		vid2vid-zero [10]		TAV [9]	TF [14]	CAV [15]	T2V-Zero [16]	RAV [17]
	w/o Ours	w/ Ours	w/o Ours	w/ Ours	w/o Ours	w/ Ours					
CLIP-Acc [55] ↑	36.21	59.86 +23.65	51.11	66.44 +15.33	51.66	71.72 +20.06	38.82	59.97	75.21	73.66	77.51
M.PSNR [11] ↑	25.29	25.35 +0.06	23.33	24.92 +1.59	19.81	20.15 +0.34	16.85	19.69	11.13	17.69	11.7
BRISQUE [56] ↓	40.06	37.94 -2.12	37.21	32.11 -5.10	15.99	15.09 -0.90	38.41	31.46	22.09	25.12	14.2

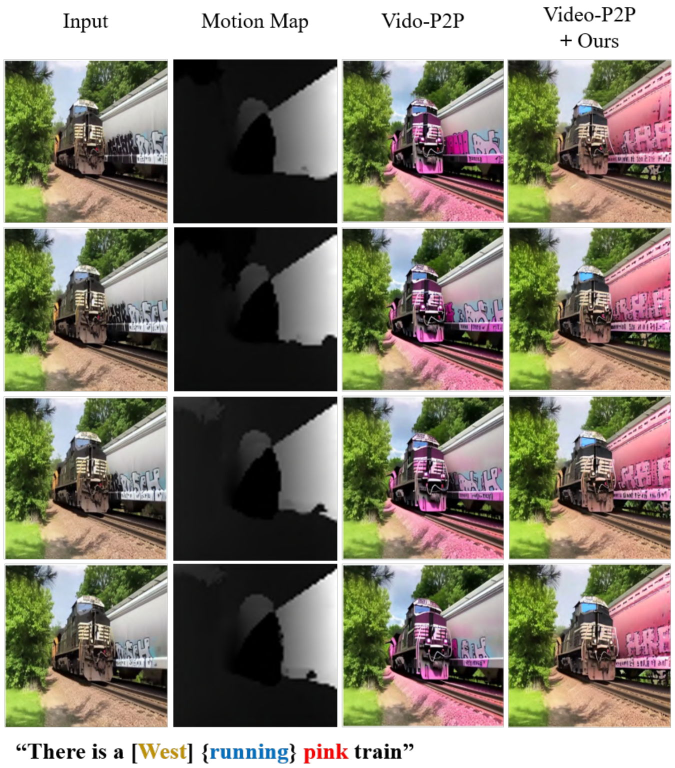


Fig. 5. Direction control method for objects moving in the direction specified by the user.

B. Qualitative Results

Fig. 4 shows a qualitative comparison at four frames between the six baseline models and our module. The non-P2P based video editing models were aligned with the target prompt but exhibited editing in areas other than the intended parts, resulting in a loss of overall structure in the edited video. P2P based video editing models, through manipulation of the cross-attention layer, achieve maintenance of the structure of the input video. However, they struggle to estimate precise motion words and overall attention maps, therefore not perfectly achieving the desired edits from the target prompt. By applying our module, motion information was injected into the attention maps of the entire prompt by estimating the motion of the frames. This enhancement in the attention map enabled editing that was previously impossible without our module. In addition, we observed that the structure of the input frames was preserved while only the desired areas were effectively edited. Therefore, it demonstrates the general applicability of P2P-based models and improvement in the performance of existing text-to-video editing models.

Direction control Prior to the editing, the user is required to specify the desired direction within the prompt. The proposed module then conducts the editing operation based on this specified direction. The Fig. 5 presents an example where only pixels displaying motion towards the West direction undergo editing. This approach allows for targeted directional control of object movements within the video content, leveraging optical flow information for precise manipulation.

C. Quantitative Results

We compared the results of 20 video data and six baseline models and our module with three metrics. As demonstrated in Table I left, the CLIP-Acc [55] shows evidence that all P2P-based models with our method understand target prompt better than those without our method. The non-P2P-based model control-a-video [15] exhibits high scores, but it demonstrates the lowest score in masked-PSNR [11]. Although there is high alignment with the target prompt, editing in unintended areas implies that text-guided editing was not effectively achieved for the input frames. Comparing our model with other baseline models, we observe that the masked PSNR [11] score is slightly higher. This is attributed to the enhanced attention maps, which preserve the unintended areas aligned to the text, thereby maintaining the structural information of the input frames. Lower BRISQUE [56] scores in all three models with M2A module demonstrate that it can improves the video quality. Especially, M2A module works best on Video-P2P [11]. It was confirmed that editing performance was improved in various aspects through a module that improved attention maps.

TABLE II
USER PREFERENCES FOR THE PROPOSED MODULE IN THE FATEZERO [12], VIDEO-P2P [11], AND VID2VID-ZERO [10] MODELS. HIGHER RATINGS WERE RECORDED FOR THE RESULTS WHEN THE PROPOSED MODULE WAS ADDED.

	Text Alignment ↑	Stucture Preserving ↑	Realism & Quality ↑	Temporal Consistency ↑
FateZero [12]	19.07	25.84	31.07	26.66
FateZero [12] + Ours	80.92 +61.85	74.15 +84.31	68.92 +37.85	73.33 +46.66
Video-P2P [11]	19.69	27.69	28.76	30.66
Video-P2P [11] + Ours	80.30 +60.61	72.30 +48.61	71.26 +42.50	69.33 +38.66
vid2vid-zero [10]	12.30	29.53	28.00	24.88
vid2vid-zero [10] + Ours	87.69 +75.39	70.46 +40.93	72.00 +44.00	75.11 +50.22

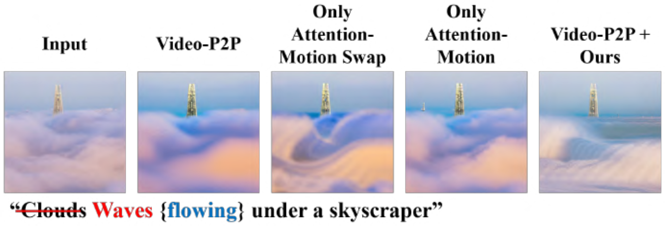

Input
Video-P2P
Only Attention-Motion Swap
Only Attention-Motion
Video-P2P + Ours
"~~Clouds~~ Waves {flowing} under a skyscraper"

Fig. 6. Our module's comparative results w/ w/o the two methods on Video-P2P [11]. Our module attains improved editability through the utilization of both methods.

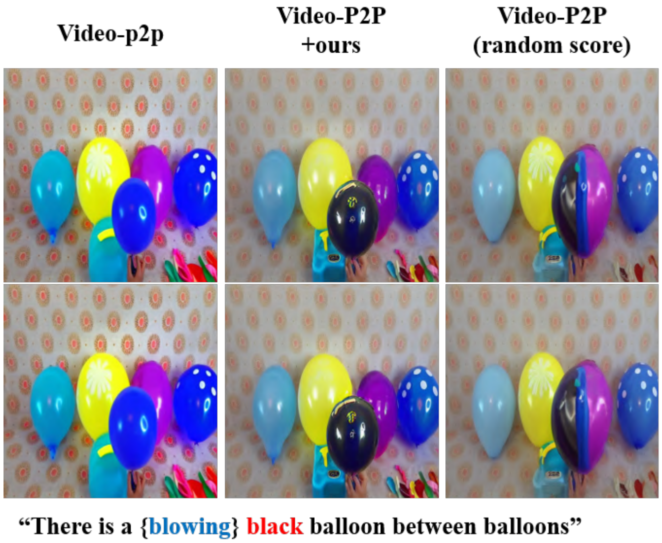


Fig. 7. The experiment shows the significance of computing correlation scores for motion-relevant prompts in the weighted sum. We utilized random scores to emphasize the importance of template matching within the M2A module to adjust the weights.

D. User Study

Since the utilized evaluation metrics may not fully represent human perception, we conducted a user study. To compare our proposed model with existing FateZero [12], Video-P2P [11], and vid2vid-zero [10], we prepared a total of 20 videos. We presented these videos to 60 participants, showing them the target prompt, input video, and each model's output video. Users were asked to make their selections based on the following aspects: (1) structure preservation, (2) text alignment, (3) quality, and (4) temporal consistency. The results in Table II right demonstrate that the output of our module is the best.

TABLE III
EVALUATION ON VARIOUS COMPOSITION-METRIC USING CLIP-ACC [12], MASKED-PSNR [11], BRISQUE [56].

Composition Metric	CLIP-Acc [55] ↑	M.PSNR [11] ↑	BRISQUE [56] ↓
Squared-Diff [51]	83.18	21.58	28.92
N.Squared-Diff [51]	62.56	21.26	29.46
Cross-Corr [49]	62.80	21.27	29.42
N.Cross-Corr [49]	82.56	21.98	25.42
Corr-Coeff [53]	62.70	21.26	29.52
N.Corr-Coeff [53]	82.90	22.06	26.04
SAM [52]	62.74	21.34	28.86
MI [21]	**83.65**	**22.31**	**23.85**

E. Ablation Study

The key part of our module is the attention-motion swap and attention-motion fusion using similarity score that incorporate the motion information from frames and attention maps.

*1) **Effect of M2A Module**:* In this study, we conducted various experiments to demonstrate the effect of the M2A module. For each of the "Attention-Motion Swap" and "Attention-Motion Fusion" components of the M2A module, we separately examined the impact of the motion map on enhancing the attention map.

a) Attention-Motion Swap: In Fig. 6, "Only Attention-Motion Swap", performance was improved by swapping the attention map of the motion prompt with motion map. When only the attention map of motion words was enhanced, it showed better editing than the Video-P2P [11], but there is a limitation as overall attention is not improved.

b) Attention-Motion Fusion: In "Only Attention-Motion Fusion" incorporates motion information by adjusting with the correlation score between attention map and motion map, it shows better editing results than Video-P2P [11]. However the attention map of the motion prompt was not accurately estimated, it was injected with a low score. The result confirm the inadequacy in enhancing the attention maps with only Attention-Motion Fusion. By using both methods, as indicated by the results of our module, we can efficiently inject motion map information to the attention map.

In this study, we utilized a total of eight Composition Metrics to calculate the correlation score between attention maps and motion maps for Attention-Motion Fusion. We utilized three metrics to measure the similarity between two images: "Squared Difference," [51] "Cross Correlation," [49] and "Correlation Coefficient" [53]. Additionally, we measured

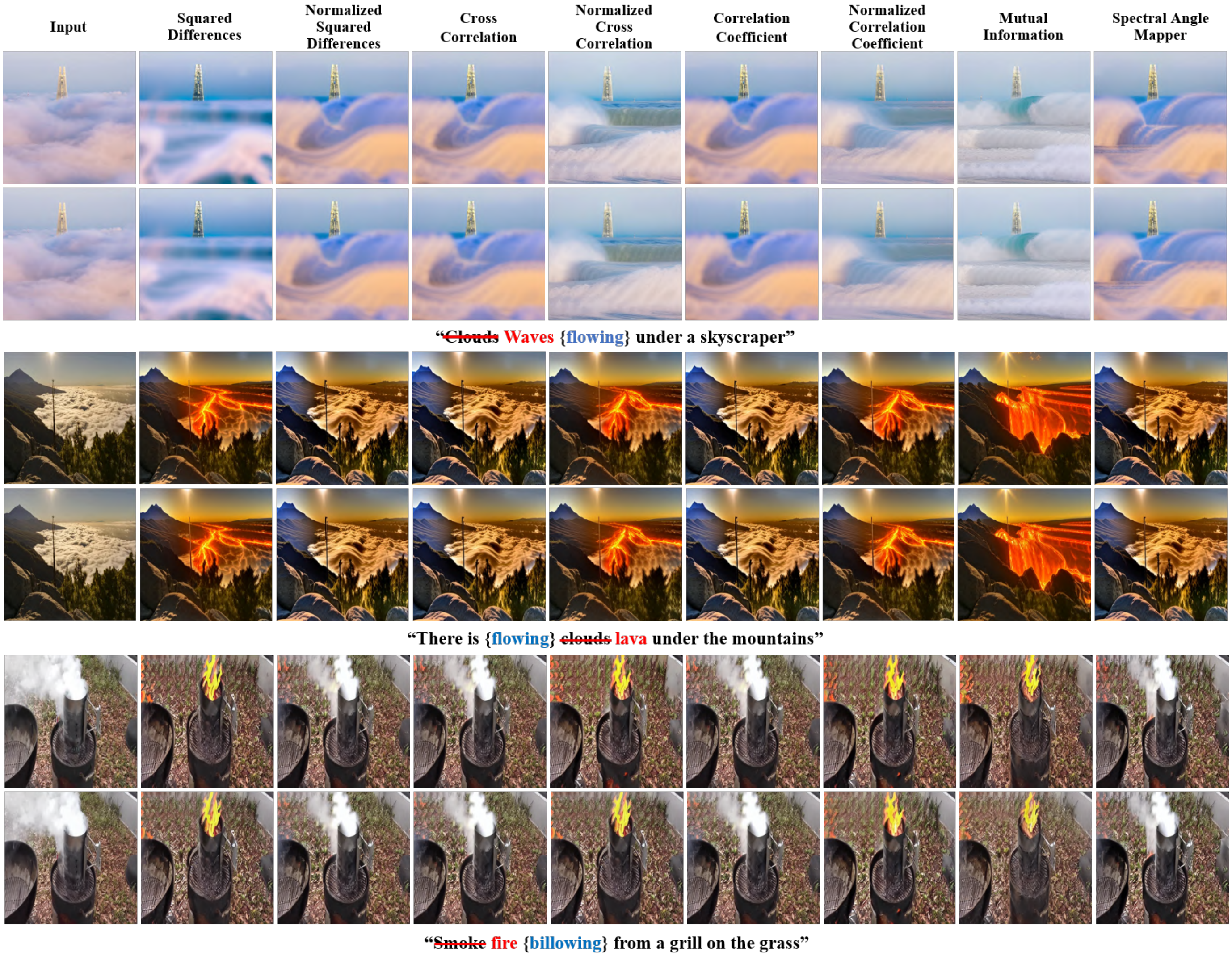

Input
Squared Differences
Normalized Squared Differences
Cross Correlation
Normalized Cross Correlation
Correlation Coefficient
Normalized Correlation Coefficient
Mutual Information
Spectral Angle Mapper
"~~Clouds~~ Waves {flowing} under a skyscraper"
"There is {flowing} ~~clouds~~ lava under the mountains"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

Fig. 8. Comparison on various metrics of template matching algorithms with Video-P2P [11]

the correlation score using normalized versions of each metric. We also measured the correlation score using the "Spectral Angle Mapper" [52] metric, which compares the angle between the spectra of pixels in two images, and "Mutual Information" [21], which assesses the degree of shared information between the two images.

In Fig. 8, the corresponding outcomes of different Composition Metric, which compute the correlation between two images, are illustrated. As depicted in Fig. 8, "Mutual Information" is the most suitable function for enhancing semantic editing. In Table III, Composition Metric were also measured through CLIP-Acc [12], Masked-PSNR [11], BRISQUE [56]. scores. "Mutual Information" [21] was also demonstrated to be the most appropriate Composition Metric by showing the highest quality in metrics. This can be interpreted as the correlation score, measured by comparing semantic information rather than the similarity of pixel values between images, more accurately assessing the Composition Metric between the attention map and motion map. This is because the attention map and motion map are not typical images but contain prompts, frames, and motion information, necessitating a comparison of meaningful information.

*2) **Random Correlation Score**:* The M2A module proposed in this study enhances the attention map through the Attention-Motion Swap and Attention-Motion Fusion processes. To effectively enhance the attention map, it is critical to adjust scores while infusing the motion map into the attention based on Composition Metrics. Motion map information must be adjusted based on correlation scores to improve the attention of each prompt. Fig. 7 shows the outcome of applying our module using random scores, without calculating correlation scores. This results in inaccurate editing by injecting motion information into "balloons" that do not represent the "blowing" motion. Therefore, this experiment underscores the importance of calculating correlation scores through Composition Metrics in the Attention-Motion Fusion process.

*3) **Comparison on Optical Flow Estimation Models**:* We compare the results of applying the optical flow estimation algorithms: Unimatch [18] and RAFT [46] to our M2A module. As can be seen in Fig. 9, Unimatch [18] allows for more precise optical flow estimation compared to RAFT [46]. However, applying both optical flow estimation to our

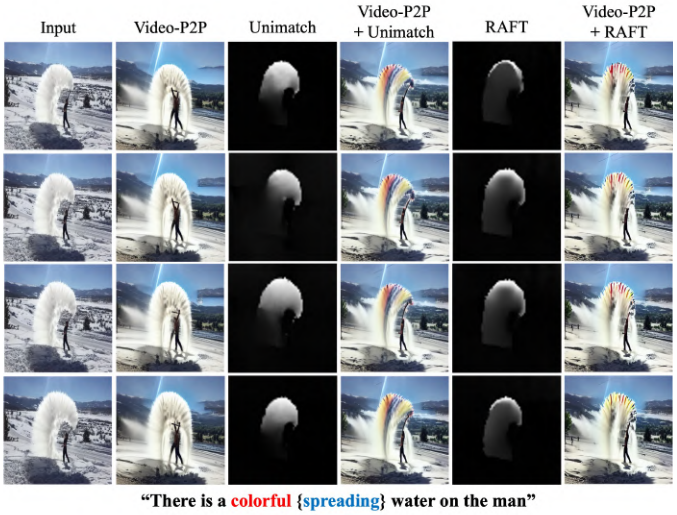


Fig. 9. The results of Video-P2P [11] and our M2A module using different optical flow algorithms: RAFT [46] and UniMatch [18]

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.

*4) **Edit Results for Static Area:*** In Fig. 10, the "forest" in the background should be edited, which is static area. However, without M2A module, the other part of the video was undermined, like the tail of the "boat". In T2V models calculate probability between an frame pixel and each prompt. This allows editing to be applied exclusively to the areas corresponding to those specific words. In this result, the pixel has a higher probability at "forest" than other words damage the tail of the boat. By employing our module, "boat" enhanced with high correlation scores and the directly injected "moving" have high weight in their motion areas, thereby reducing the impact of "forest". This outcome results in a better preservation of moving objects compared to editing static areas as original models.

F. Runtime Details

The runtime of the existing vid2vid-zero[10] is about 2m20s and the inference time is about 2m38s with our module. Also, the execution time of Video-p2p[11] increased from about 1m6s to 1m50s. Finally, FateZero[12] rose from 2m38s to 4m13s.

V. Conclusion

We propose a Motion-to-Attention (M2A) module to inject estimated motion map into the attention map of the image diffusion model. Injecting motion map into attention map with our proposed M2A module makes general video editing performance improved because attention map of motion prompt become apparent. This is proven by improved evaluation metrics.

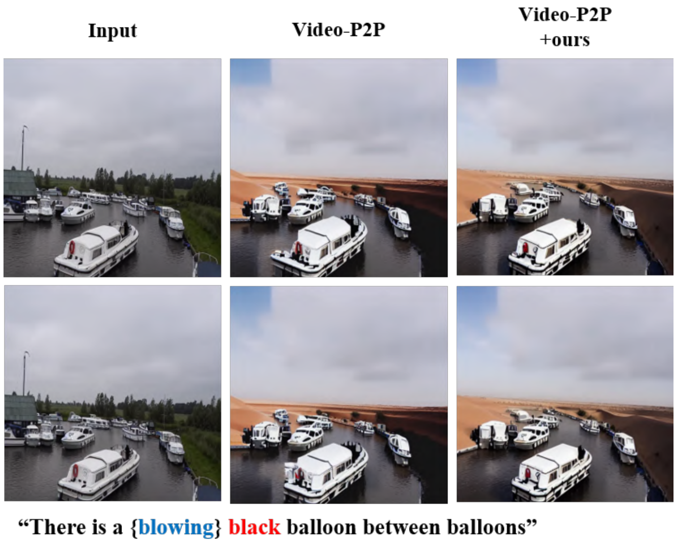


Fig. 10. Advantage of our module in editing static part. Without our module, we only tried to modify forest, but we can observe that the boat is distorted. However, with our module, the boat can be prevented from being distorted.

References

[1] M. Laavanya and V. Vijayaraghavan, "Residual learning of transfer-learned alexnet for image denoising," *IEIE Transactions on Smart Processing & Computing*, vol. 9, no. 2, pp. 135–141, 2020.

[2] J. Kim, S. Kim, C. Pyo, H. Kim, and C. Yim, "Progressive dehazing and depth estimation from a single hazy image," *IEIE Transactions on Smart Processing & Computing*, vol. 11, no. 5, pp. 343–350, 2022.

[3] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-or, "Prompt-to-prompt image editing with cross-attention control," in *The Eleventh International Conference on Learning Representations*, 2023. [Online]. Available: https://openreview.net/forum?id=_CDixzkzeyb.

[4] J. Choi, Y. Choi, Y. Kim, J. Kim, and S. Yoon, "Custom-edit: Text-guided image editing with customized diffusion models," *arXiv preprint arXiv:2305.15779*, 2023.

[5] Z. Zhang, L. Han, A. Ghosh, D. N. Metaxas, and J. Ren, "Sine: Single image editing with text-to-image diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6027–6037.

[6] T. Brooks, A. Holynski, and A. A. Efros, "Instructpix2pix: Learning to follow image editing instructions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18 392–18 402.

[7] J. Wang, P. Liu, J. Liu, and W. Xu, "Text-guided eyeglasses manipulation with spatial constraints," *IEEE Transactions on Multimedia*, vol. 26, pp. 4375–4388, 2024. DOI: 10.1109/TMM.2023.3322326.

[8] E. Molad, E. Horwitz, D. Valevski, *et al.*, "Dreamix: Video diffusion models are general video editors," *arXiv preprint arXiv:2302.01329*, 2023.

[9] J. Z. Wu, Y. Ge, X. Wang, *et al.*, "Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation," *arXiv preprint arXiv:2212.11565*, 2022.

[10] W. Wang, K. Xie, Z. Liu, *et al.*, "Zero-shot video editing using off-the-shelf image diffusion models," *arXiv preprint arXiv:2303.17599*, 2023.

[11] S. Liu, Y. Zhang, W. Li, Z. Lin, and J. Jia, "Video-p2p: Video editing with cross-attention control," *arXiv preprint arXiv:2303.04761*, 2023.

[12] C. Qi, X. Cun, Y. Zhang, *et al.*, "Fatezero: Fusing attentions for zero-shot text-based video editing," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 15 886–15 896.

[13] A. Khandelwal, "Infusion: Inject and attention fusion for multi concept zero-shot text-based video editing," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3017–3026.

[14] M. Geyer, O. Bar-Tal, S. Bagon, and T. Dekel, "Token-flow: Consistent diffusion features for consistent video editing," *arXiv preprint arXiv:2307.10373*, 2023.

[15] W. Chen, J. Wu, P. Xie, *et al.*, "Control-a-video: Controllable text-to-video generation with diffusion models," *arXiv preprint arXiv:2305.13840*, 2023.

[16] L. Khachatryan, A. Movsisyan, V. Tadevosyan, *et al.*, "Text2video-zero: Text-to-image diffusion models are zero-shot video generators," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 15 908–15 918.

[17] S. Yang, Y. Zhou, Z. Liu, and C. C. Loy, "Rerender a video: Zero-shot text-guided video-to-video translation," New York, NY, USA: Association for Computing Machinery, 2023, ISBN: 9798400703157. DOI: 10.1145/3610548.3618160. [Online]. Available: https://doi.org/10.1145/3610548.3618160.

[18] H. Xu, J. Zhang, J. Cai, *et al.*, "Unifying flow, stereo and depth estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[19] X. Shi, Z. Huang, D. Li, *et al.*, "Flowformer++: Masked cost volume autoencoding for pretraining optical flow estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 1599–1610.

[20] A. Dosovitskiy, P. Fischer, E. Ilg, *et al.*, "Flownet: Learning optical flow with convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2758–2766.

[21] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Physical review E*, vol. 69, no. 6, p. 066 138, 2004.

[22] S.-M. Woo, S.-E. Lee, and J.-O. Kim, "Deep texture-adaptive image denoising for practical application," *IEIE Transactions on Smart Processing & Computing*, vol. 11, no. 6, pp. 412–420, 2022.

[23] V.-G. Nguyen, "Digital radiography with a consumer camera: Image denoising and deblurring," *IEIE Transactions on Smart Processing & Computing*, vol. 10, no. 5, pp. 398–406, 2021.

[24] S. Cao, W. Chai, S. Hao, Y. Zhang, H. Chen, and G. Wang, "Difffashion: Reference-based fashion design with structure-aware transfer by diffusion models," *IEEE Transactions on Multimedia*, vol. 26, pp. 3962–3975, 2024. DOI: 10.1109/TMM.2023.3318297.

[25] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

[26] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020.

[27] H. Emami, M. M. Aliabadi, M. Dong, and R. B. Chinnam, "Spa-gan: Spatial attention gan for image-to-image translation," *IEEE Transactions on Multimedia*, vol. 23, pp. 391–401, 2021. DOI: 10.1109/TMM.2020.2975961.

[28] D. Saisanthiya and P. Supraja, "Neuro-facial fusion for emotion ai: Improved federated learning gan for collaborative multimodal emotion recognition," *IEIE Transactions on Smart Processing & Computing*, vol. 13, no. 1, pp. 61–68, 2024.

[29] I. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[30] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, 2022.

[31] C. Saharia, W. Chan, S. Saxena, *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," *Advances in Neural Information Processing Systems*, vol. 35, pp. 36 479–36 494, 2022.

[32] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 684–10 695.

[33] S. Li, S. Zhu, Y. Ge, *et al.*, "Depth-guided deep video inpainting," *IEEE Transactions on Multimedia*, vol. 26, pp. 5860–5871, 2024. DOI: 10 . 1109 / TMM . 2023 . 3340089.

[34] J. Zhu, H. Ma, J. Chen, and J. Yuan, "Motionvideogan: A novel video generator based on the motion space learned from image pairs," *IEEE Transactions on Multimedia*, vol. 25, pp. 9370–9382, 2023. DOI: 10.1109/TMM.2023.3251095.

[35] D. Chun, T. S. Kim, K. Lee, and H.-J. Lee, "Compressed video restoration using a generative adversarial network for subjective quality enhancement," *IEIE Transactions on Smart Processing & Computing*, vol. 9, no. 1, pp. 1–6, 2020.

[36] D. Ceylan, C.-H. P. Huang, and N. J. Mitra, "Pix2video: Video editing using image diffusion," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 23 206–23 217.

[37] M. Zhao, R. Wang, F. Bao, C. Li, and J. Zhu, "Controlvideo: Adding conditional control for one shot

text-to-video editing," *arXiv preprint arXiv:2305.17098*, 2023.

[38] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3836–3847.

[39] Z. Hu and D. Xu, "Videocontrolnet: A motion-guided video-to-video translation framework by using diffusion model with controlnet," *arXiv preprint arXiv:2307.14073*, 2023.

[40] E. Chu, T. Huang, S.-Y. Lin, and J.-C. Chen, "Medm: Mediating image diffusion models for video-to-video translation with temporal correspondence guidance," *arXiv preprint arXiv:2308.10079*, 2023.

[41] Y. Cong, M. Xu, C. Simon, *et al.*, "Flatten: Optical flow-guided attention for consistent text-to-video editing," *arXiv preprint arXiv:2310.05922*, 2023.

[42] A. Ranjan and M. J. Black, "Optical flow estimation using a spatial pyramid network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4161–4170.

[43] D. Sun, X. Yang, M.-Y. Liu, and J. 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*, 2018, pp. 8934–8943.

[44] T.-W. Hui, X. Tang, and C. C. Loy, "Liteflownet: A lightweight convolutional neural network for optical flow estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8981–8989.

[45] P. Hu, G. Wang, and Y.-P. Tan, "Recurrent spatial pyramid cnn for optical flow estimation," *IEEE Transactions on Multimedia*, vol. 20, no. 10, pp. 2814–2823, 2018. DOI: 10.1109/TMM.2018.2815784.

[46] Z. Teed and J. 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*, Springer, 2020, pp. 402–419.

[47] H. Xu, J. Zhang, J. Cai, H. Rezatofighi, and D. Tao, "Gmflow: Learning optical flow via global matching," in *Proceedings of the IEE E/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 8121–8130.

[48] R. Brunelli, *Template matching techniques in computer vision: theory and practice*. John Wiley & Sons, 2009.

[49] J. P. Lewis, "Fast template matching," in *Vision interface*, Quebec City, QC, Canada, vol. 95, 1995, pp. 15–19.

[50] R. Brunelli and T. Poggiot, "Template matching: Matched spatial filters and beyond," *Pattern recognition*, vol. 30, no. 5, pp. 751–768, 1997.

[51] M. Hisham, S. N. Yaakob, R. Raof, A. A. Nazren, and N. Wafi, "Template matching using sum of squared difference and normalized cross correlation," in *2015 IEEE student conference on research and development (SCOReD)*, IEEE, 2015, pp. 100–104.

[52] X. Liu and C. Yang, "A kernel spectral angle mapper algorithm for remote sensing image classification," in *2013 6th International Congress on Image and Signal Processing (CISP)*, IEEE, vol. 2, 2013, pp. 814–818.

[53] N. J. Napoli, L. E. Barnes, and K. Premaratne, "Correlation coefficient based template matching: Accounting for uncertainty in selecting the winner," in *2015 18th International Conference on Information Fusion (Fusion)*, IEEE, 2015, pp. 311–318.

[54] C. E. Shannon, "A mathematical theory of communication," *The Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.

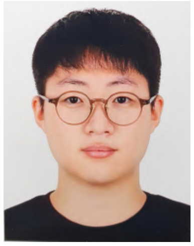
[55] J. Hessel, A. Holtzman, M. Forbes, R. L. Bras, and Y. Choi, "Clipscore: A reference-free evaluation metric for image captioning," *arXiv preprint arXiv:2104.08718*, 2021.

[56] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, 2012. DOI: 10.1109/TIP.2012.2214050.

Seong-Hun Jeong received the B.S. degree in journalism and broadcasting from Pukyong National University, Pusan, South Korea, in 2021, and the M.S. degree in media communication from Pukyong National University, Pusan, South Korea, in 2023. He is currently enrolled as a doctoral student in the department of electrical and electronic engineering at Pusan National University. His current research interests include image watermarking, computer vision, generative AI, and deep learning.

Inhwan Jin is currently enrolled in the School of Media and Communication at Pukyong National University, Pusan, South Korea. His current research interests include multi-modal, computer vision, generative AI, and deep learning.

Haesoo Choo is currently enrolled in the Department of Korean Language and Literature at Pukyong National University. Her current research interests include computer vision, generative AI, and deep learning.



Hyeonjun Na is currently enrolled in the School of Media and Communication at Pukyong National University, Pusan, South Korea. His current research interests include diffusion model, computer vision, image editing, and image generation.

Kyeongbo Kong received the B.S. degree in electronics engineering from Sogang University, Seoul, South Korea, in 2015, and the M.S. and Ph.D. degrees in electrical engineering from the Pohang University of Science and Technology (POSTECH), Pohang, South Korea, in 2017 and 2020, respectively. From 2020 to 2021, he worked as a Postdoctoral Fellow with the Department of Electrical Engineering, POSTECH, Pohang, South Korea. From 2021 to 2023, he was an Assistant Professor of Media School at Pukyong National University, Busan. He is currently an Assistant Professor of Electronics Engineering at Pusan National University. His current research interests include image processing, computer vision, machine learning, and deep learning.