Motion-to-Attention: Enhancing Attention Maps to Improve Performance of Text-Guided Video Editing Models

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

***Abstract*—Recent research in text-guided video editing aims to extend image-based editing models to video domains. A significant challenge in this transition is ensuring temporal consistency across frames. However, existing methods often exhibit limited editing accuracy when processing prompts associated with motion, such as *"floating"* or *"moving."* Our analysis indicates that this limitation arises from inaccurate attention maps corresponding to motion-related prompts. To address this, we introduce the Motion-to-Attention (M2A) module, explicitly integrating motion information for enhanced video editing precision. Specifically, we first convert optical flow extracted from the video into a comprehensive motion map. Optionally, users can specify directional information to refine motion map extraction further. The proposed M2A module incorporates two complementary techniques: *"Attention-Motion Swap,"* which directly substitutes the imprecise attention map of motion prompts with the extracted motion map, and *"Attention-Motion Fusion,"* which adaptively enhances attention maps based on the correlation with the motion map using a carefully selected Fusion metric. Experimental validation demonstrates that incorporating our M2A module into existing text-to-video editing frameworks significantly improves both quantitative performance metrics (CLIP-Acc, Masked PSNR, BRISQUE) and qualitative visual quality. Extensive experiments and comparative studies confirm the superior editability and robustness of our method over current state-of-the-art approaches. Comprehensive results are publicly available at https://currycurry915.github.io/Motion-to-Attention/.**

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

I. Introduction

UNPRECEDENTED advancements in image generation and editing have recently been achieved through text-guided diffusion models and large-scale language models. Unlike traditional methods [1], which primarily perform global edits using deep neural networks, contemporary research emphasizes precise and localized editing driven solely by user-provided textual prompts [2]–[4]. Among these approaches, Prompt-to-Prompt (P2P) [2] facilitates semantic and localized image editing by directly manipulating attention maps guided by textual prompts, eliminating the need for external inputs such as segmentation masks.

Parallel to image editing advancements, text-guided video editing research is also rapidly progressing [5]–[7]. However, video editing introduces unique challenges, particularly temporal consistency across frames. Due to limited availability of extensive text-video paired datasets, most existing methods extend image-based models to video scenarios through zero-shot approaches [8], [9] or fine-tuning techniques [5]. Nevertheless, applying image-centric editing frameworks to video frames independently often results in temporal inconsistencies, compromising overall video quality. Recent studies thus primarily focus on maintaining temporal coherence between edited frames [7]–[13].

Despite advancements in T2V editing techniques, as shown in Fig. 1, existing models struggle to accurately estimate attention maps for motion-related prompts such as “floating,” “moving,” “spreading,” and “parking.” These inaccuracies prevent the editing results from fully reflecting the intended meaning of the target prompt, often leading to incomplete edits or unintended disappearance of objects. This issue is attributed to the direct extension of text-guided image models to the video editing domain, which does not adequately capture motion information inherent to videos. This limitation indicates that such models are not trained enough to handle motion-representing prompts. Inaccurate attention maps not only restrict the editability of motion prompts but also undermine the effective editing of moving objects, ultimately limiting the performance of current T2V editing approaches.

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 [14], which is widely used in various video tasks to utilize motion information. The optical flow method can extract highly accurate motion information by estimating the change in pixels between video frames. We extract only the magnitude information from the optical flow to utilize it as a motion map. This motion map represents the regions indicated by the motion prompt. Furthermore, since the optical flow contains both magnitude and direction information, we can distinguish regions moving in a specified direction and represent them as a motion map. Based on this information, we propose an optional method for editing only the regions moving in a specific direction within a video by utilizing the motion map extracted with the

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 Electronics Engineering, Pusan National University, Busan, South Korea (e-mail: tlqwkrk915@pusan.ac.kr).

Inhwan Jin, Haesoo Choo and Hyeonjun Na are with the major of Human ICT at Pukyong National University, Busan, South Korea. (e-mail: bds06081@naver.com, tndi28@naver.com, wewe1st@naver.com).

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

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

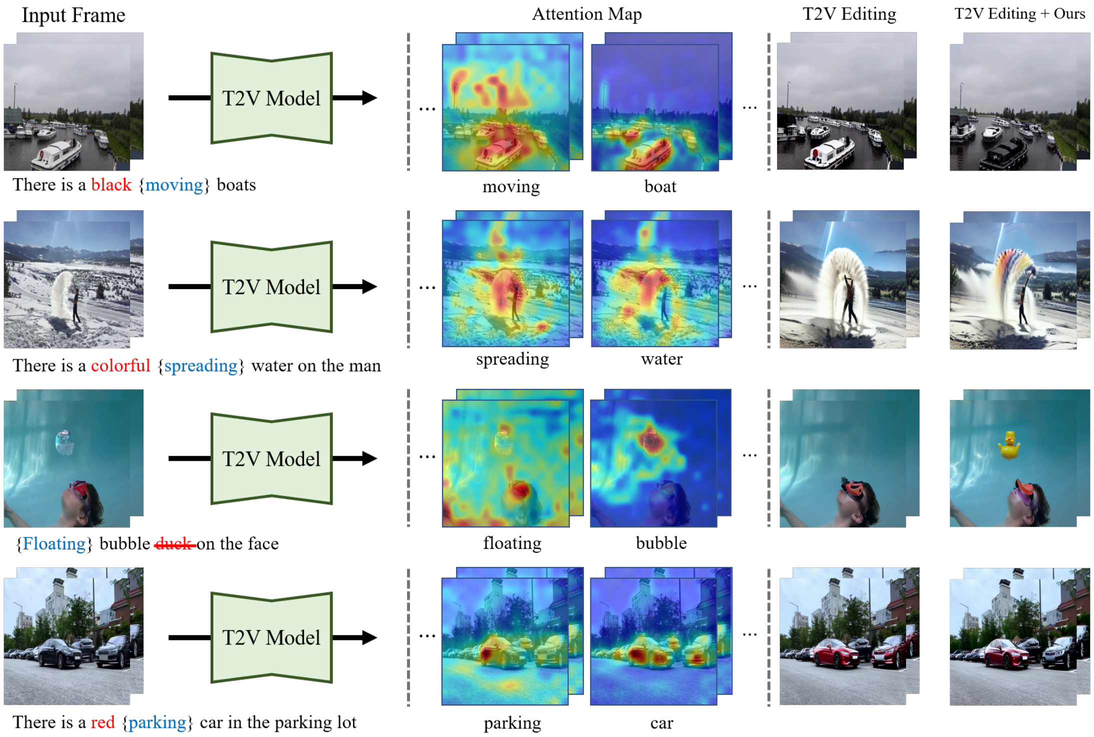
Input Frame
Attention Map
T2V Editing
T2V Editing + Ours
T2V Model
...
...
There is a black {moving} boats
moving
boat
T2V Model
...
...
There is a colorful {spreading} water on the man
spreading
water
T2V Model
...
...
{Floating} bubble ~~duck~~ on the face
floating
bubble
T2V Model
...
...
There is a red {parking} car in the parking lot
parking
car

Fig. 1. This figure visualizes the results of the T2V model for the input video and its corresponding attention map, confirming the inaccurate estimation of the motion prompt (e.g., floating, moving). The existing T2V model failed to accurately estimate the attention map for the motion prompt, resulting in restricted editability. The proposed Motion-to-Attention (M2A) module improves the attention map of the entire prompt, demonstrating enhanced editability for existing video editing models.

specified direction information. We call this method **Direction Guidance**, which is performed only when the user provides direction information. If no direction information is provided, only magnitude information is used.

Our primary contribution lies in effectively integrating the estimated motion map with the attention map to enhance video editing precision. To this end, we introduce the **Motion-to-Attention (M2A)** module, which comprises two complementary approaches: *Attention-Motion Swap* and *Attention-Motion Fusion*. Attention-Motion Swap directly substitutes the inaccurate attention map of the motion prompt with the precise motion map, immediately improving editing accuracy. However, while this method addresses inaccuracies specific to motion prompts, it does not fully leverage the correlation between the motion map and other attention maps within the prompt. To overcome this limitation, Attention-Motion Fusion optimizes the integration of the motion map by adaptively refining associations with other attention maps in the prompt. This adaptive integration is achieved by calculating a Fusion metric that assesses pixel-level, spectral, and informational correlations, enabling the selection of the most effective weighting mechanism. By combining these two methods within the M2A module, we significantly improve the editing accuracy and effectiveness of areas targeted by motion-related prompts compared to conventional approaches.

The contributions of our paper are as follows:

- We found that inaccurately estimating the attention map for prompts indicating essential movements in text-guided video editing reduces video editability. Our 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. By utilizing directional information present in optical flow, the proposed method enables the identification and editing of regions corresponding to movements in a user-specified direction. This approach allows for more precise and controlled editing, accommodating specific editing needs based on motion direction.
- We verified that applying the M2A module to existing attention-based text-guided video editing models improves the performance.

II. Related Work

A. Text-Guided Image/Video Editing

Unlike the traditional field of image processing research [15], which relied on relatively simple deep neural networks to remove general noise or other artifacts from images, recent research of text-guided image editing [16] uses a variety of generative models. The diffusion model [17], 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. Research on video editing using other generative models [18] is also being continuously conducted, but most recent studies are based on the diffusion model. Subsequent video editing research diverges into two primary paths: 1) P2P based Model, 2) Non-P2P based Model.

Among P2P based Models, Video-P2P [7] divides its framework into two branches—one for unchanged parts and another for edited parts—and incorporates each attention

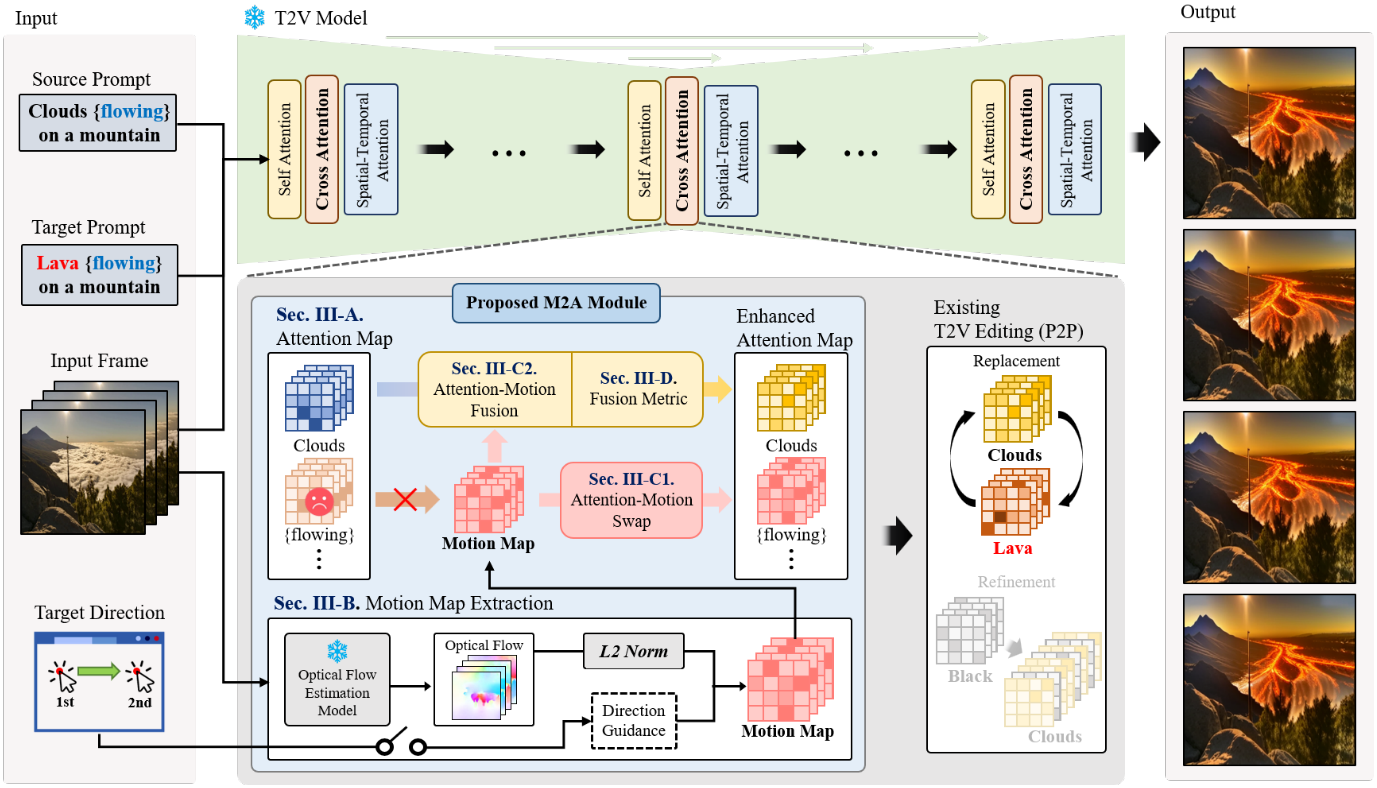
Input
T2V Model
Output
Source Prompt
Clouds {flowing} on a mountain
Target Prompt
Lava {flowing} on a mountain
Input Frame
Target Direction
1st
2nd
Self Attention
Cross Attention
Spatial-Temporal Attention
Proposed M2A Module
Sec. III-A. Attention Map
Sec. III-C2. Attention-Motion Fusion
Sec. III-D. Fusion Metric
Enhanced Attention Map
Clouds
{flowing}
Motion Map
Sec. III-C1. Attention-Motion Swap
Sec. III-B. Motion Map Extraction
Optical Flow Estimation Model
Optical Flow
L2 Norm
Direction Guidance
Motion Map
Existing T2V Editing (P2P)
Replacement
Clouds
Lava
Refinement
Black
Clouds

Fig. 2. The left side of the figure shows the overall framework of video editing by enhancing the attention map. First, the Text-to-Video (T2V) Model generates an attention map by receiving video and prompts as input. Simultaneously, the optical flow estimation model estimates the optical flow from the input video frames. The estimated optical flow is converted to a motion map by default using only magnitude information. Optionally, when direction information is provided by the user, the Direction Control converts the optical flow to a motion map that only shows movement in the user-specified direction. If the user indicates directional words with [], the model captures the direction information and performs Direction Control. Then, the motion map is injected into the attention map of the T2V-Model in two ways from the M2A module: Attention-Motion Fusion and Attention Motion Swap. After that, text-to-video editing is performed using the attention map enhanced by the motion map. The right side of the figure shows how the Attention-Motion Swap and Attention-Motion Fusion of the M2A module enhance the attention map with the motion map.

map to enable detailed editing. Alongside the aforementioned works, vid2vid-zero [8] achieves stable video editing and reconstruction by integrating cross-frame attention into the U-Net structure of an existing diffusion model. Additionally, FateZero [9] is zero-shot based and maintain the video's temporal consistency through the enforcement of the attention map.

Among Non-P2P based Models, ControlVideo [19] maintains temporal consistency in a manner similar to Pix2video [20], but it employs the editing method of the existing ControlNet [21] for performing edits. TokenFlow [10] 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 [11] integrates motion and content priors, introducing motion-adaptive noise initialization strategies to enhance the consistency and quality of the video. Object-aware Video Editing [22] reduces cross-frame attention costs by separating objects from the background and merging redundant background tokens. Diffusion Noise Injection (DNI) [23] improves editing accuracy by filtering initial latent noise with a spectral bandpass filter. SliceEdit [24] employs pre-trained T2I diffusion models to process spatial and spatiotemporal slices, leveraging the similarity between video slices and natural images. VidTome [25] enhances temporal consistency by aligning and compressing redundant tokens across frames.

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 [26]. Subsequently, a series of works, including RecSPy [27], and RAFT [28] were proposed to reduce computational costs by using a coarse-to-fine and iterative estimation method. We generate the optical flow of the video using the most recent optical flow estimation model, UniMatch [14]. 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 [14] and RAFT [28], another optical flow estimation models.

C. Fusion Metric

We aim to examine metrics that can quantitatively measure the association between the attention map and the motion map in the image domain, spectral domain, and information domain. Traditionally, various methods have been used to measure the association between two images in the image domain, and we utilize three methods: Squared Difference [29], Cross Correlation [30], and Correlation Coefficient [31]. Squared Difference [29], 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 to calculate association. Additionally, we use Cross Correlation [30], which calculates the association between two images by summing the product of their pixel values. We also use the Correlation Coefficient [31], which is calculated by considering the mean and standard deviation, to measure association. In the spectral domain, we use Spectral Angle Mapper [32], which calculates the association by comparing the angles between the spectral represented by each pixel in the two images. In the information domain, we use Mutual Information [33], which measures the association between two images by considering the pixel values of the images as random variables and assessing how much the values change together.

III. Proposed Method

Before delving into the specifics, we first provide an 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) [2] setting, we define the source prompt as $\mathcal{P}$ and the target prompt as $\mathcal{P}_T$. In the source prompt, the prompt containing motion information of the video is called the 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". Furthermore, users can specify a directional information $\mathcal{D}$ using the provided GUI. This allows the extraction of a direction vector and the angle between the vector and the origin. Utilizing this angle, the motion in the specified direction within the optical flow is identified. First, the T2V model receives the frames of the input video $\mathcal{V}$ and a source prompt $\mathcal{P}$ as input and generates an attention map $\mathcal{A}$ indicating the regions represented by the words in the prompt within a single frame. Simultaneously, the optical flow estimation model receives the frames of the input video and estimates the optical flow $\mathcal{V}_{flow}$. To utilize only the magnitude values of the $\mathcal{V}_{flow}$, we apply L2 normalization to convert the $\mathcal{V}_{flow}$ into a motion map $\mathcal{M}$. Additionally, if the user provides directional information $\mathcal{D}$, we propose an optional method called Direction Guidance, which extracts only the movement information in the user-specified direction $\mathcal{D}$ from the optical flow $\mathcal{V}_{flow}$ and converts it into a motion map $\mathcal{M}_D$. The proposed M2A module, which includes the Attention-Motion Swap and the Attention-Motion Fusion, enhances the attention map $\mathcal{A}^*$ by injecting the estimated motion map $\mathcal{M}$. Subsequently, in the Attention Control process of the P2P model used for editing in the T2V-Model, the enhanced attention map $\mathcal{A}^*$ estimated by the proposed method is utilized to perform video editing.

In Sec. III-A, we introduce the attention map $\mathcal{A}$ as prior knowledge. 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, including "Attention-Motion Swap" and "Attention-Motion Fusion". In Sec. III-D, we explain Fusion Metric for calculating the Fusion score $\mathcal{F}$ used in "Attention-Motion Fusion". In Sec. III-E, we briefly explain the process of editing a video through Attention Control in P2P, the editing model used by the T2V model, based on the enhanced attention map provided by the proposed M2A module.

A. Attention Map

P2P [2] is a text-guided image editing model that performs editing by manipulating the attention map $\mathcal{A}$ in the image domain. The attention map $\mathcal{A}$ estimated through P2P's cross-attention layer visually represents the correlation between a word and an image. The attention map $\mathcal{A}$ can be calculated as follows:

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

The spatial features of the image are transformed into the query matrix $\mathcal{Q}$, and the 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. d represents the dimensionality of the query and key. Afterwards, each pixel value is converted to a probability value by applying the softmax function.

The T2V model edits the video by replacing or refining the estimated attention map of the source prompt with the attention map of the target prompt. Before this process, we swap the attention map of the motion prompt in the source prompt with the motion map. Then, we enhance the attention map of the entire prompt with the motion map before proceeding with the subsequent process.

B. Motion Map Extraction

We obtain the optical flow $\mathcal{V}_{flow} = (u, v)$ through the pre-trained optical flow estimation model [14] 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 $\mathcal{A}$ utilizing this optical flow. We extract a motion map $\mathcal{M}$ 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:

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

Since optical flow $\mathcal{V}_{flow}$ is a vector with both magnitude and direction, we apply L2 normalization to the optical flow to extract only the magnitude information and use it as a motion map $\mathcal{M}$.

Additionally, we propose **Direction Guidance**, which allows for the editing of only the regions that exhibit movement

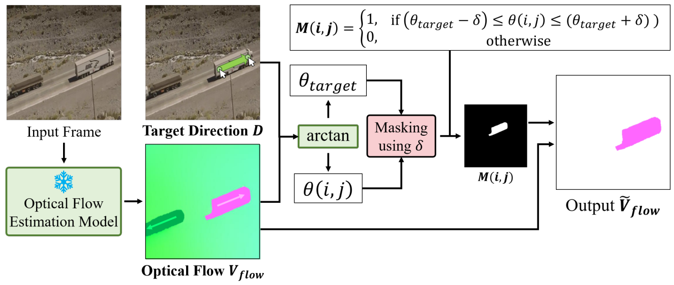
$M(i,j) = \begin{cases} 1, & \text{if } (\theta_{target} - \delta) \leq \theta(i,j) \leq (\theta_{target} + \delta)\,) \\ 0, & \text{otherwise} \end{cases}$
θ_{target}
arctan
Masking using δ
$\theta(i,j)$
$M(i,j)$
Output $\tilde{V}_{flow}$
Input Frame
Target Direction D
Optical Flow Estimation Model
Optical Flow V_{flow}

Fig. 3. The process of Direction Guidance. The user first specifies the target motion direction by selecting a region of interest in the video. An optical flow estimation model is then applied to extract optical flow across frames. The directional information is computed using the arctangent function, which derives the motion orientation for each pixel. Subsequently, a masking operation is performed based on a threshold δ to retain only the regions exhibiting movement in the user-defined direction. This process enables precise control over motion-aware editing by isolating directional motion components.

in a specific direction by utilizing not only the magnitude but also the directional information of the optical flow $\mathcal{V}_{flow}$. Direction Guidance is an optional method that is activated only when direction information is provided by the user. As shown in Fig. 3, when Direction Guidance is applied after obtaining the target directional information θ_{target}, we extract only the motion that aligns with this target direction for each pixel in the optical flow field (u, v). For each pixel, we first compute the angle θ of its optical flow vector as follows:

$$\theta = \arctan(u, v). \tag{3}$$

Given the target direction θ_{target} and an allowable tolerance δ, we then define a mask M as:

$$M(ij) = \begin{cases} 1, & \text{if } (\theta_{\text{target}} - \delta) \leq \theta(i,j) \leq (\theta_{\text{target}} + \delta), \\ 0, & \text{otherwise}, \end{cases} \tag{4}$$

where (i, j) denotes the pixel location within the image. This mask is applied element-wise to the optical flow components to suppress motion that does not fall within the specified directional range:

$$\tilde{u}(i,j) = M(i,j) \cdot u(i,j), \quad \tilde{v}(i,j) = M(i,j) \cdot v(i,j) \tag{5}$$

For the filtered optical flow $\tilde{\mathcal{V}}_{flow} = (\tilde{u}, \tilde{v})$, the motion map $\mathcal{M}$ is computed by applying the L2 normalization. Through this, we can specifically extract the regions in the optical flow that are moving in the user-provided direction.

The motion map $\mathcal{M}$ extracted using this method is injected into the attention map $\mathcal{A}$ through the proposed M2A module, thereby enhancing the attention map $\mathcal{A}$. The enhanced attention map $\mathcal{A}^*$ is then provided to the T2V model.

C. Motion-to-Attention Module

As discussed in the Introduction, existing video editing models rely on generative models trained on text-image pair datasets. As a result, they fail to accurately capture the attention map of motion prompts in videos, which represent movement. This inaccurate attention map limits the performance of video editing. To address this issue, we propose "Motion-to-Attention," which consists of two methods: 1) Attention-Motion Swap and 2) Attention-Motion Fusion. The pseudo code for the entire algorithm is provided in Alg 1. Additionally, Fig. 4 illustrates the overall process of the proposed M2A module.

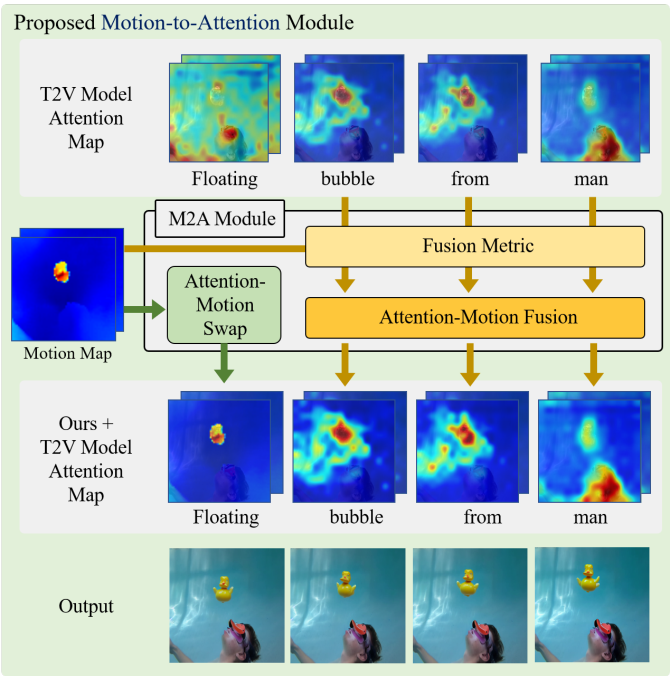

Fig. 4. The proposed Motion-to-Attention (M2A) module, which leverages the motion map to enhance the attention map. The M2A module replaces the inaccurate attention map of the motion prompt with the motion map (Attention-Motion Swap) and injects the motion map into other attention maps (Attention-Motion Fusion), thereby improving the performance of video editing.

Algorithm 1 Motion-to-Attention (M2A) Module

```
Input: Attention maps of entire source prompt $\mathcal{A}$,
    Motion map $\mathcal{M}$,
    Number of source prompts $N$,
    Index of motion prompt $j$,
    Fusion score $\mathcal{F}$,
    Hyperparameter for motion map injection rate $\lambda$,
    Denoising timestep in diffusion model $t$
Output: Enhanced attention maps of entire source prompt $\mathcal{A}^*$
for $k = 1, 2, ..., N$ do
    if $k = j$ then
        # Apply Attention-Motion Swap for motion prompt
        $\mathcal{A}^*_k = \lambda \cdot \mathcal{M}$
    else
        # Apply Attention-Motion Fusion for non-motion prompts
        $\mathcal{A}^*_k = \mathcal{A}_k + \lambda \cdot \frac{\mathcal{F}_k \cdot \mathcal{M}}{t}$
    end if
end for
return $\mathcal{A}^*$
```

1) Attention-Motion Swap: In Fig. 1, we observe that existing T2V models fail to accurately estimate the attention map $\mathcal{A}_M$ for the motion prompt $\mathcal{P}_M$, which negatively impacts their video editing capabilities. To address this issue, Attention-Motion Swap replaces the inaccurate attention map $\mathcal{A}_M$ with the motion map $\mathcal{M}$, representing the motion information in the video. The motion map $\mathcal{M}$ extracted in the previous step is resized to match the size of the attention map $\mathcal{A}$. Next, the values of $\mathcal{M}$ are normalized between 0 and 1 by dividing them by the maximum value of the motion map. This normalization ensures compatibility with the attention map $\mathcal{A}$, which is represented as a probability map with values ranging from 0 to 1. After aligning the dimensions and scales of the attention map and motion map, Attention-Motion Swap is processed according to the following equation:

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

where λ is a hyperparameter that controls the injection rate of the motion map and $\mathcal{A}_{\mathcal{P}_\mathcal{M}}$ is attention map of motion prompt $\mathcal{P}_\mathcal{M}$. By directly replacing the inaccurate attention map with the motion map, Attention-Motion Swap improves the editability of the T2V model. The input prompt and video frames first pass through a self-attention layer, followed by a cross-attention layer. Through self-attention, information from individual prompts is related, creating association between the motion prompt's attention map and the attention maps of other prompts. Although the Attention-Motion Swap method addresses inaccuracies in the motion prompt's attention map by replacing it with the motion map, this approach does not account for the inherent associations between the motion prompt and other prompts. Consequently, although some improvement in editing quality can be achieved, the inability to consider these associations limits the overall accuracy and consistency of the editing process. To overcome this limitation, a supplementary method capable of incorporating the relationships between the motion prompt and other prompts is essential.

2) Attention-Motion Fusion: Attention-Motion Fusion leverages the association between the motion map and the attention map to improve the attention maps across the prompt, excluding the motion prompt. The Fusion score $\mathcal{F}$, which quantifies the association between the attention map and the motion map, is computed using Fusion Metrics. Detailed descriptions of these Fusion Metrics are provided in Sec. III-D. The computed $\mathcal{F}_k$ is used as a weight to account for the association between the attention maps across the entire prompt. Based on this, the enhanced attention map $\mathcal{A}_k^*$ is calculated as follows:

$$\mathcal{A}_k^* = \mathcal{A}_k + \lambda \cdot \frac{(\mathcal{F}_k \cdot \mathcal{M})}{t}, \quad \text{if } \mathcal{A}_k \neq \mathcal{A}_{\mathcal{P}_\mathcal{M}}, \tag{7}$$

Where λ is a hyperparameter for adjusting the weight of the motion map during integration into the attention map. By sharing the same λ value across both Attention-Motion Swap and Attention-Motion Fusion, consistency in scaling is maintained for all motion maps. This ensures that motion maps are uniformly scaled, enabling their effective integration into the attention map without introducing inconsistencies.

The parameter t represents the denoising step in the diffusion process of the T2V model. In the early stages of the diffusion process, the t value is low, allowing the motion map to exert a strong influence on the overall structure of the attention map. This ensures that the foundational motion information is adequately incorporated. As the diffusion process progresses, the t value increases, which progressively reduces the influence of the motion map. This approach enables the refinement of details and ensures that the edited content aligns seamlessly with the target prompt.

Therefore, Attention-Motion Fusion and Attention-Motion Swap serve as complementary methods. By using both approaches together, more precise and effective editing can be achieved, significantly improving overall editing performance.

D. Fusion Metric

In Attention-Motion Fusion, to calculate the Fusion score $\mathcal{F}$ between the attention maps $\mathcal{A}$ of the entire prompt and the motion map $\mathcal{M}$, we utilize various Fusion Metrics. To calculate the Fusion score $\mathcal{F}$ between the two maps, we consider various metrics that quantitatively measure the association in the image domain, spectral domain, and information domain. The attention map, which represents the correlation between the words in the input prompt and the image as a pixel-wise probability value, and the motion map, derived from the magnitude of the optical flow divided by its maximum value, both have values ranging from 0 to 1. Since the motion map is resized to match the size of the attention map, the association between the two maps can be quantified by calculating the Fusion score $\mathcal{F}$.

1) Fusion Metric in Image Domain: In the image domain, the association between two images is measured using pixel values. Traditional methods such as **Squared Difference** [29], **Cross correlation** [30] and **Correlation coefficient** [31] are commonly used to calculate the Fusion score $\mathcal{F}$. The **Squared Difference** [29] metric computes the Fusion score as the sum of squared differences between pixel values, defined as $\mathcal{F} = \sum(\mathcal{A}(x,y) - \mathcal{M}(x,y))^2$, where (x,y) denotes the pixel coordinates. This method is useful for directly evaluating the absolute differences between two images. On the other hand, **Cross Correlation** [30] measures structural similarity by summing the product of pixel values, calculated as $\mathcal{F} = \sum \mathcal{A}(x,y)\mathcal{M}(x,y)$. This Cross Correlation [30] can measure the structural similarity between the two images, allowing for an accurate assessment of the similarity between the attention map $\mathcal{A}$ and the motion map $\mathcal{M}$. Lastly, the **Correlation Coefficient** [31] normalizes the values of the pixels using their mean μ and standard deviation σ, allowing the measurement of association in a scale-invariant manner. Unlike Squared Difference or Cross Correlation, this method ensures consistency by addressing scale variations between the attention map and the motion map. These metrics collectively allow for a precise calculation of the association between the two maps, facilitating accurate Fusion.

2) Fusion Metric in Spectral Domain: To calculate the Fusion score $\mathcal{F}$ between the attention map $\mathcal{A}$ and the motion map $\mathcal{M}$, we use the **Spectral Angle Mapper** (SAM) [32]. SAM compares the angles between the spectral vectors of pixels to measure the association between two images. This has the advantage of not being affected by the pixel value magnitudes of the attention map $\mathcal{A}$ and the motion map $\mathcal{M}$. The spectral angle θ between two pixels is defined as:

$$\theta = \cos^{-1}\left(\frac{\mathbf{a} \cdot \mathbf{m}}{\|\mathbf{a}\|\|\mathbf{m}\|}\right), \tag{8}$$

where $\mathbf{a}$ and $\mathbf{m}$ are the spectral vectors of pixels from the attention map and the motion map, respectively, and $\cdot$ denotes the dot product.

3) Fusion Metric in Information Domain: To measure the association between the attention map $\mathcal{A}$ and the motion map $\mathcal{M}$ in information domain, we use **Mutual Information** (MI) [33] in the information domain. Mutual Information evaluates the association based on probability distributions, accounting for complex and nonlinear relationships. It quantifies the amount of shared information between two images, enabling a more comprehensive analysis. Mutual Information is defined as:

$$\mathrm{MI}(\mathcal{A}, \mathcal{M}) = H(\mathcal{A}) + H(\mathcal{M}) - H(\mathcal{A}, \mathcal{M}), \tag{9}$$

where $H(\mathcal{A})$ and $H(\mathcal{M})$ represent the entropy of $\mathcal{A}$ and $\mathcal{M}$, respectively.

By using the previously described Fusion metrics, we measure the Fusion score $\mathcal{F}$ between the attention map $\mathcal{A}$ and the motion map $\mathcal{M}$ and use it as a weight. To enhance the attention map $\mathcal{A}$ with the motion map $\mathcal{M}$, we multiply the motion map $\mathcal{M}$ by the Fusion score $\mathcal{F}$ and then add it to the attention map $\mathcal{A}$.

The enhanced attention map $\mathcal{A}^*$ is then provided to the T2V model to perform video editing. The experimental results for various metrics can be found in Fig. 8.

E. Text Guided Video Editing Using Enhanced Attention Map

The enhanced attention map $\mathcal{A}^*$ through the proposed M2A module is provided to the T2V model and utilized for video editing. Several recent video editing models [7]–[9] extend the image editing model, P2P [2], to the video domain for editing. These video editing models mainly propose methods to ensure that frame-by-frame editing is performed consistently, but the method of editing the video is based on how P2P [2] edits images. P2P [2] controls the attention map of the source prompt $\mathcal{P}$ to the attention map $\mathcal{A}_T$ of the target prompt $\mathcal{P}_T$ to edit images. The main methods are Replacing and Refinement. The enhanced attention map $\mathcal{A}^*$ of the source prompt is replaced with the attention map $\mathcal{A}_T$ of the target prompt. This process can be expressed by the following equation using the edit function $Edit(\cdot)$:

$$Edit(\mathcal{A}^*_t, \mathcal{A}^*_{T,t}, t) := \begin{cases} \mathcal{A}^*_{T,t} & \text{if } t < \tau \\ \mathcal{A}^*_t & \text{otherwise} \end{cases}, \tag{10}$$

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 $\mathcal{A}_T$ of target prompt $\mathcal{P}_T$.

Otherwise, users want to change the style of the image or attribute of certain object. For example, $\mathcal{P}$ = "a car" to $\mathcal{P}_T$ = "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.

$$Edit\left(\mathcal{A}^*_t, \mathcal{A}^*_{T,t}, t\right)_{i,j} := \begin{cases} \left(\mathcal{A}^*_{T,t}\right)_{i,j} & \text{if } \mathcal{L}(j) = \text{None} \\ \left(\mathcal{A}^*_t\right)_{i,\mathcal{L}(j)} & \text{otherwise,} \end{cases} \tag{11}$$

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

IV. Experiments

A. Experimental Setup

1) 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 [9], Video-P2P [7], vid2vid-zero [8]. Additionally, we compared the results with those of non-P2P-based models among the latest studies with publicly released codes. These models include Text2Video-Zero (T2V-Zero) [12], Tune-a-video (TAV) [6], Control-A-Video (CAV) [11], Token-Flow (TF) [10], SliceEdit (SE) [24], and VidToMe [25].

2) Dataset: The experiments were conducted using the Davis video dataset [34], which is widely used in various video generation and editing studies due to its high-quality annotations and diverse video content. In addition to the Davis dataset, we expanded our evaluation by incorporating a collection of YouTube videos. These videos were selected to introduce more variability in terms of content, motion complexity, and environmental factors, ensuring that the proposed methods were tested on a broader range of real-world scenarios. This combination of datasets allowed for a more comprehensive assessment of model performance.

3) Implementation details: We used RTX 3090 GPUs in the experiment, and we set the image resolution to 512×512 as in the existing FateZero [9]. 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 [14] model.

4) Evaluation Metrics: The proposed M2A module enhances the inaccurate attention map of the motion prompt with a motion map, enabling both spatially and semantically accurate video editing. To evaluate the performance of M2A, we utilized three metrics: CLIP-Acc, Masked PSNR, and BRISQUE.

To quantitatively evaluate the semantic alignment between the edited video and the target textual prompt, we used the trained CLIP model [35], which calculates a similarity score between the textual prompt and the edited video. A higher score indicates better alignment with the target prompt while maintaining distinction from the source prompt. For structural evaluation, Masked PSNR [7] was employed to measure how well the unchanged regions in the video were preserved before and after editing. The visual quality of the edited videos was assessed using the No-Reference Image Quality Assessment method, BRISQUE [36]. This metric evaluates the perceptual quality of the edited videos without requiring reference data.

By combining CLIP-Acc for semantic accuracy, Masked PSNR for structural preservation, and BRISQUE for visual quality, this evaluation metrics comprehensively validates the effectiveness of the proposed M2A module in enhancing video editing.

B. Qualitative Results

Fig. 5 visually compares the results of the P2P based model, the results of the P2P based model with our module

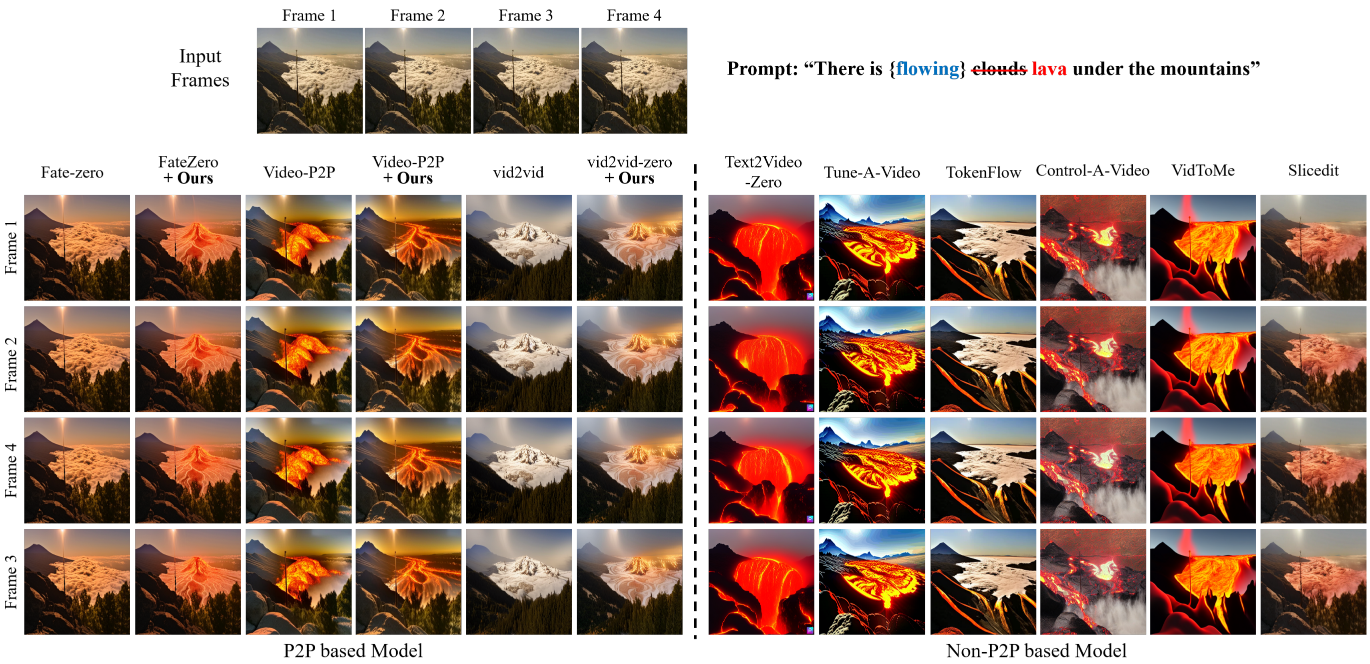

Fig. 5. Qualitative results of our study. The experimental results were distinctly divided into P2P based models and Non-P2P based models. While editing was performed globally on the Non-P2P based models, the application of the proposed M2A module to the P2P based models enabled precise targeting and editing of areas corresponding to the target prompt.

TABLE I

Performance of the proposed module measured by various metrics. In this study, we conducted experiments distinguishing between P2P based models and Non-P2P based models. For the P2P based models, we compared their performance using metrics after applying our proposed module to the results obtained from these models.

	P2P based Model						Non-P2P based Model					
	FateZero		Video-P2P		vid2vid-zero		TAV	TF	CAV	T2V-Zero	SE	VidToMe
	w/o Ours	w/ Ours	w/o Ours	w/ Ours	w/o Ours	w/ Ours						
CLIP-Acc ↑	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	35.64	56.96
M.PSNR ↑	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	24.40	16.01
BRISQUE ↓	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	24.20	36.28

TABLE II

User preferences for the proposed module in the FateZero [9], Video-P2P [7], and vid2vid-zero [8] models. Higher ratings were recorded for the results when the proposed module was added.

	Text Alignment ↑	Stucture Preserving ↑	Realism & Quality ↑	Temporal Consistency ↑
FateZero [9]	19.07	25.84	31.07	26.66
FateZero [9] + Ours	**80.92** +61.85	**74.15** +48.31	**68.92** +37.85	**73.33** +46.66
Video-P2P [7]	19.69	27.69	28.76	30.66
Video-P2P [7] + Ours	**80.30** +60.61	**72.30** +48.61	**71.26** +42.50	**69.33** +38.66
vid2vid-zero [8]	12.30	29.53	28.00	24.88
vid2vid-zero [8] + Ours	**87.69** +75.39	**70.46** +40.93	**72.00** +44.00	**75.11** +50.22

applied, and the results of the non-P2P based model. 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 attention map, 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. We confirmed that accurate editing was performed when using the enhanced attention map through the proposed method for video editing. 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 T2V models.

Before editing, the user specifies the desired direction through the provided GUI. Using this input, the direction vector and the angle between the vector and the origin are extracted. The proposed module rotates the optical flow based

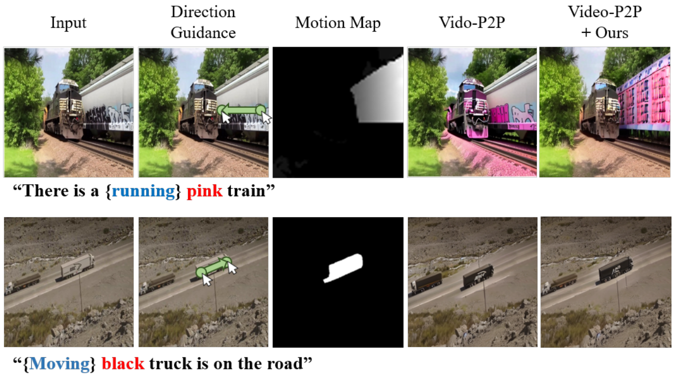
Input
Direction Guidance
Motion Map
Vido-P2P
Video-P2P + Ours
"There is a {running} pink train"
"{Moving} black truck is on the road"

Fig. 6. Direction Guidance method for objects moving in the direction specified by the user.

on the extracted angle and then generates the motion map. Fig. 6 illustrates an example where editing is performed using the motion representing movement in the user-specified direction. By applying the proposed module to existing editing models, the results confirm that edits are performed accurately in the direction specified by the user.

C. Quantitative Results

We quantitatively evaluated the results of 20 videos using three metrics: CLIP-Acc [35], Masked PSNR [7], and BRISQUE [36]. The evaluation compared three settings: (1) the original P2P-based models, (2) P2P-based models with the proposed M2A module applied, and (3) Non-P2P-based models.

Table I demonstrates that applying the proposed module to P2P-based models significantly improved CLIP-Acc scores. For example, FateZero's CLIP-Acc increased from 36.21 to 59.86, and the vid2vid-zero score increased from 51.66 to 71.72. These results indicate that the M2A module improves the semantic understanding of target prompts. Non-P2P-based models, such as TF [10], CAV [11], and T2V-Zero [12], achieved comparable or higher CLIP-Acc scores. However, this improvement can be attributed to unintended edits throughout theame, rather than accurate edits focused solely on the target regions.

The Masked PSNR results show a slight improvement when applying the M2A module to P2P-based models, indicating better structural integrity in non-edited regions. For example, FateZero's score increased marginally from 25.29 to 25.35, while Video-P2P showed a more notable improvement of +1.59. In contrast, Non-P2P models recorded lower Masked PSNR scores overall, likely due to unintended edits in non-target areas. An exception is SE [24], which exhibited higher Masked PSNR scores, but this is likely because the edits were incomplete, leaving most of the input video unchanged.

BRISQUE scores for P2P-based models improved after applying the M2A module, reflecting better visual quality in edited videos. Although Non-P2P-based models also showed improved BRISQUE scores, actual editing results often revealed unrealistic or incomplete edits, with unintended changes throughout the frame. This suggests that higher BRISQUE scores in Non-P2P models may not accurately reflect editing quality.

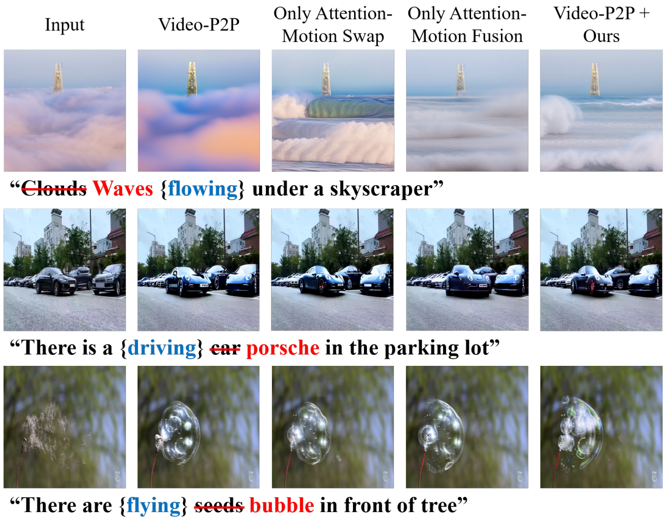
Input
Video-P2P
Only Attention-Motion Swap
Only Attention-Motion Fusion
Video-P2P + Ours
"~~Clouds~~ Waves {flowing} under a skyscraper"
"There is a {driving} ~~car~~ porsche in the parking lot"
"There are {flying} ~~seeds~~ bubble in front of tree"

Fig. 7. We conducted experiments to observe the results of applying the Attention-Motion Swap and Attention-Motion Fusion of the M2A module individually. The results showed that utilizing both models together achieved the best performance.

In summary, the M2A module demonstrates its effectiveness by significantly enhancing CLIP-Acc scores, maintaining structural integrity as evidenced by Masked PSNR, and improving visual quality measured by BRISQUE. These improvements are primarily attributed to the accurate motion map generated by the M2A module, which minimizes unintended edits in non-target regions and ensures precise editing aligned with the target prompt. The results, shown in Fig. 5 and Table I, underscore the superiority of the M2A module in achieving more accurate and meaningful video edits.

D. User Study

To address the possibility that evaluation metrics may not fully capture human perception, we conducted a user study. In this study, we compared the proposed module with existing models, including FateZero [9], Video-P2P [7], and vid2vid-zero [8], using a total of 20 videos. These videos were presented to 60 participants, along with the target prompt, input video, and the output video from each model. Participants were asked to evaluate their preferences based on four criteria: (1) Structure Preservation, (2) Text Alignment, (3) Quality, and (4) Temporal Consistency.

The results presented in Table II demonstrate that the proposed M2A module outperformed all models across all evaluation criteria. These results can be attributed to the comprehensive improvements introduced by the M2A module. Specifically, the M2A module effectively enhances the inaccurate attention map of motion prompts, improving alignment with the target prompt while maintaining the structure of non-target regions. These capabilities address common issues observed in existing models, such as unintended edits, poor prompt alignment, and low temporal consistency.

TABLE III
WE EVALUATED THE PERFORMANCE OF THE M2A MODULE BY MEASURING CLIP-ACC, MASKED PSNR, AND BRISQUE FOR THE RESULTS OF APPLYING ATTENTION-MOTION SWAP, ATTENTION-MOTION FUSION, AND THE INTEGRATED APPROACH COMBINING BOTH METHODS.

	Video-P2P	Only Swap	Only Fusion	Video-P2P+Ours
CLIP-Acc ↑	26.12	26.23	26.68	**27.29**
M.PSNR ↑	22.75	25.57	26.45	**27.31**
BRISQUE ↓	42.32	29.60	27.25	**23.68**

E. Ablation Study

We conducted various ablation studies to comprehensively evaluate the performance of the M2A module. Through these experiments, we confirmed that our proposed method significantly improves the performance of the existing video editing model.

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. 7, "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 [7], but there is a limitation as overall attention is not improved.

b) Attention-Motion Fusion: In "Only Attention-Motion Fusion", motion information is incorporated by adjusting with the Fusion score between the attention map and the motion map, showing better editing results than Video-P2P [7]. However, because the attention map of the motion prompt was not accurately estimated, it was injected with a low score. The results confirm the inadequacy of enhancing the attention maps with only Attention-Motion Fusion. As shown in Table III, our proposed module, which integrates the Attention Motion Swap and Attention Motion Fusion, effectively incorporates motion map information into the attention map. When each method is applied individually, for instance Only Swap or Only Fusion, there is a marked improvement in CLIP-Acc, Masked PSNR, and BRISQUE compared to the baseline called Video P2P. Notably, when both methods are employed together, referred to as Video P2P plus Ours, the approach achieves the highest performance across all three metrics and thus demonstrates the effectiveness of the proposed method.

c) Fusion Score: In this study, we utilized a total of eight Fusion Metrics to calculate the Fusion Score between the attention map and the motion map for Attention-Motion Fusion. Metrics suitable for the image domain, spectral domain, and information domain were employed to measure the associations. Overall, the experimental results demonstrate that the metrics in the spectral domain did not produce meaningful improvements either quantitatively or qualitatively compared to the image and information domains. This observation is reflected in Fig. 8 and Table IV.

TABLE IV
EVALUATION ON VARIOUS FUSION METRICS USING CLIP-ACC [9], MASKED-PSNR [7], BRISQUE [36].

Fusion Metrics		CLIP-Acc ↑	M.PSNR ↑	BRISQUE ↓
Image Domain	Squred-Diff	83.18	21.58	28.92
	N.Squred-Diff	62.56	21.26	29.46
	Cross-Corr	62.80	21.27	29.42
	N.Cross-Corr	82.56	21.98	25.42
	Corr-Coeff	62.70	21.26	29.52
	N.Corr-Coeff	82.90	22.06	26.04
Spectral Domain	SAM	62.74	21.34	28.86
Information Domain	MI	**83.65**	**22.31**	**23.85**

In the image domain, we calculated the Fusion Score using "Squared Difference" [29], "Cross Correlation" [30], "Correlation Coefficient" [31], and their normalized versions. As shown in image domain of Table IV and Fig. 8, the non-normalized version of Squared Difference consistently outperformed the normalized version across all quantitative evaluation metrics. This result suggests that Squared Difference, which is based on absolute pixel-level differences, is less sensitive to normalization and highlights that normalization may not always be the optimal choice in certain contexts. On the other hand, the normalized versions of Cross Correlation and Correlation Coefficient recorded higher evaluation scores compared to their non-normalized counterparts, producing semantically more precise and visually more realistic editing results. This indicates that normalization effectively addresses the scale mismatches between the attention map and the motion map, thereby enabling a more accurate comparison of their structural associations. These findings suggest that while normalization may not always be necessary for pixel-level evaluations, it is particularly beneficial when assessing structural relationships between maps, underscoring its importance in specific contexts of association measurement within the image domain.

Among all the metrics used across different domains, MI [33] demonstrated the most outstanding performance both quantitatively and qualitatively. MI achieved the highest scores not only in quantitative evaluations such as CLIP-Acc, Masked-PSNR, and BRISQUE but also in visual assessments of video quality. This is because MI effectively captures the semantic associations and nonlinear relationships between the attention map and the motion map, surpassing simple pixel-value comparisons. Since the attention map and motion map are not standard images but instead contain meaningful data such as prompts, frames, and motion details, MI was deemed the most suitable Fusion Metric for enhancing semantic editing.

2) Comparison on Optical Flow Estimation Models: We compare the results of applying the optical flow estimation algorithms: Unimatch [14] and RAFT [28] to our M2A module. As can be seen in Fig. 9, Unimatch [14] allows for more precise optical flow estimation compared to RAFT [28]. However, applying both optical flow estimation to our

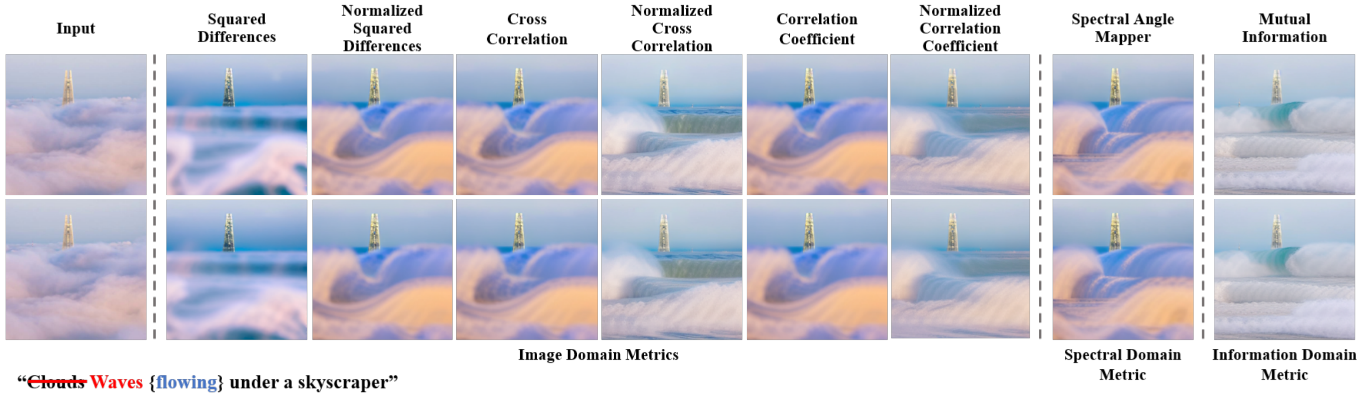
Input
Squared Differences
Normalized Squared Differences
Cross Correlation
Normalized Cross Correlation
Correlation Coefficient
Normalized Correlation Coefficient
Spectral Angle Mapper
Mutual Information
Image Domain Metrics
Spectral Domain Metric
Information Domain Metric
“~~Clouds~~ Waves {flowing} under a skyscraper”

Fig. 8. Results from using various Composition Metrics to measure the correlation score in the Attention-Motion Fusion within the M2A module. Among the different metrics, the results obtained using Mutual Information appeared to be the most realistic.

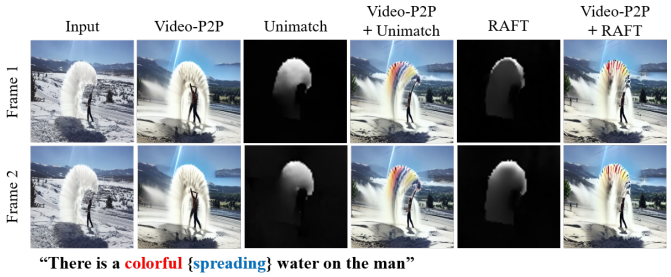

Fig. 9. The results of Video-P2P [7] and our M2A module using different optical flow algorithms: RAFT [28] and UniMatch [14]

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.

V. Conclusion

We propose an M2A module to inject an estimated motion map into the attention map of the image diffusion model. Injecting motion map into attention map with our proposed M2A module improves general video editing performance because the motion prompt attention map becomes apparent. This is shown by improving the evaluation metrics. In the future, we will conduct research in the direction of editing areas where complex optical flow is generated due to various camera movements.

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] 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.

[3] 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.

[4] 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.

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

[6] J. Z. Wu, Y. Ge, X. Wang, *et al.*, “Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 7623–7633.

[7] S. Liu, Y. Zhang, W. Li, Z. Lin, and J. Jia, “Video-p2p: Video editing with cross-attention control,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 8599–8608.

[8] 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.

[9] 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.

[10] M. Geyer, O. Bar-Tal, S. Bagon, and T. Dekel, “Tokenflow: Consistent diffusion features for consistent video editing,” in *The Twelfth International Conference on Learning Representations*.

[11] 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.

[12] L. Khachatryan, A. Movsisyan, V. Tadevosyan, *et al.*, “Text2video-zero: Text-to-image diffusion models are zero-shot video generators,” in *2023 IEEE/CVF Inter-*

national Conference on Computer Vision (ICCV), 2023, pp. 15 908–15 918.

[13] 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.

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

[15] 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.

[16] 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.

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

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] K. Kahatapitiya, A. Karjauv, D. Abati, F. Porikli, Y. M. Asano, and A. Habibian, "Object-centric diffusion for efficient video editing," in *European Conference on Computer Vision*, Springer, 2025, pp. 91–108.

[23] S. Yoon, G. Koo, J. W. Hong, and C. D. Yoo, "Dni: Dilutional noise initialization for diffusion video editing," in *European Conference on Computer Vision*, Springer, 2025, pp. 180–195.

[24] N. Cohen, V. Kulikov, M. Kleiner, I. Huberman-Spiegelglas, and T. Michaeli, "Slicedit: Zero-shot video editing with text-to-image diffusion models using spatio-temporal slices," in *Proceedings of the 41st International Conference on Machine Learning*, ser. ICML'24, Vienna, Austria: JMLR.org, 2025.

[25] X. Li, C. Ma, X. Yang, and M.-H. Yang, "Vidtome: Video token merging for zero-shot video editing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 7486–7495.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

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

[31] 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.

[32] 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.

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

[34] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung, "A benchmark dataset and evaluation methodology for video object segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 724–732.

[35] 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.

[36] 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.