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

Figure 1. Effectiveness of the proposed MMI module. The proposed method has improved editability in all text-to-video editing methods.

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 'flowing'. 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 Map Injection (MMI) module to carry out precise video editing by explicitly taking motion into account. The proposed MMI module employs two methodologies, "Motion Swapping" and "Weighted Sum using Correlation Score", to enhance the attention map using motion information. Notably, our MMI module is effectively applied to the attention-based text-guided video editing models, providing enhanced editability. Extensive experimental results are https://currycurry915.github.io/Attention-Flow/

1. 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. In particular, many researches have been conducted on image editing using only the prompt provided by the user [1–4]. Among them, Prompt-to-Prompt [1] 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 [5–8]. However, there is a lack of sufficient text-video pair training data to train large-scale video generation models. Therefore, most of the text-guided video editing research extends image editing models to video editing models using the zero-shot approach [7, 9, 10] or by fine-tuning text-guided image editing models [5]. In contrast to image, video consists of multiple

frames. Therefore, recent researches such as Video-P2P [8], FateZero [9], and vid2vid-zero [7] focuses on achieving temporal consistency between frames as a primary challenge in text-guided video editing.

However, despite their efforts, as shown in Fig. 1, 'cloud' is not edited to 'lava', or even if it is edited to 'lava', it is not 'flowing'. To find the reason of this phenomenon, we delved the estimated results of the attention map. In this process, we found inaccurate estimation of the attention map for prompt representing motion, such as 'flowing'. 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 methodology to enhance the attention map in text-guided video editing by extracting motion information from video. We essentially use optical flow [11], which is widely used in various video tasks [12, 13], 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.

Our key contribution is the effective incorporation of the generated motion map into the attention map. The most intuitive approach is to directly inject the generated motion map into the attention map for the corresponding 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 inject 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 correlation score between the motion map and the attention map. Since one attention map is semantically related to other attention maps, directly injecting 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 weighted sum method that utilizes the correlation score between the attention map and motion map as a weight. At this point, we compared a total of nine template matching [14] methods to calculate the similarity between the motion map and the attention map. In our experiments, Normalized Cross Correlation (NCC) [15, 16] showed the highest performance. Through this process, improved editing results compared to the existing text-guided video editing models are confirmed in Fig. 1.

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 methodology 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, as shown in Fig.1. Our module will continue to be applied to future research in attention-based text-guided video editing models, providing enhanced editability for video editing.

2. Related Work

2.1. Text-Guided Editing

The diffusion model [17, 18] is a generative model that, by introducing noise into or eradicating noise from the input image, generated an output image with a probability distribution that resembles that of the input image. The results of recent studies on the diffusion models are impressive, outperforming earlier methods like Generative Adversarial Nets (GAN) [19]. Text-guided image editing models like DALL-E2 [20] and Imagen [21], and stable diffusion [22] show the outcomes of high-quality image editing based on this diffusion model. Prompt-to-Prompt [1], in particular, offers text-guided image editing that utilizes attention maps to control the association between the prompt text token and the corresponding image pixels, which enables unprecedented semantic editing.

Research on text-guided video editing with generative models has recently been expanded as a result of the significant advancements made in text-guided image editing. The first diffusion-based method of text-guided motion and application editing of videos was presented by Dreamix [5] but there are problems with localized editing via word replacement. Video-P2P [8] divides their framework into two branches—one for unchanged parts and one for edited parts—and incorporates each attention map to enable detailed editing.

In parallel with the works mentioned above, vid2vid-zero [7] performs stable video editing and reconstruction by incorporating cross-frame attention into the U-Net structure of a pre-existing diffusion model. Furthermore, FateZero [9] and InFusion [10] are based on zero-shot and maintain the temporal consistency of the video through enforcing the attention map. Additionally, ControlVideo [23] and Pix2video [24] focus on determining the structure and appearance of the frame through self-attention within the diffusion model, and show how to add a layer that can maintain temporal consistency, such as cross-frame attention. Video-ControlNet [25], MeDM [26], and FLATTEN [27] maintain temporal consistency by enforcing frames generated from video motion information obtained through optical flow. As

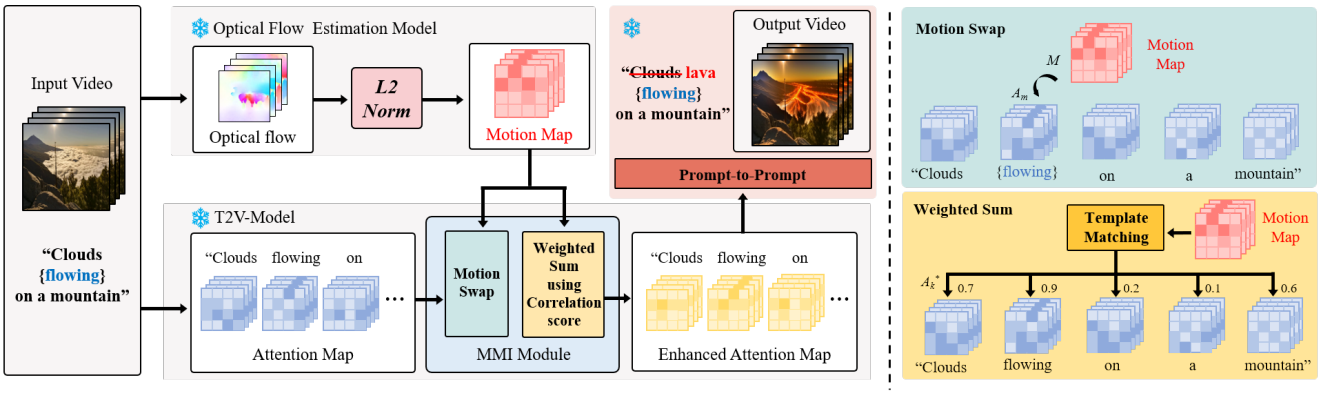

Figure 2. Overall framework. First, the T2V-Model generates an attention map by receiving video and prompts as input. At the same time, the Motion Map Injection (MMI) module receives the video frame and generates a motion map. Then motion map is injected into the attention map of the T2V-Model through two methodologies: "Motion Swapping" and "Weighted Sum using Correlation Score". After that, text-to-video editing is performed using the attention map that includes video motion information.

such, most of the recent text-guided video editing models are focusing on achieving temporal consistency in order to expand the text-guided image model into a successful video editing model. In contrast to prior studies, we enhanced the attention map using motion and effectively applied it to the existing text-guided video editing model to provide better editability of video editing.

2.2. Optical Flow Estimation

Estimating the motion of objects in a video sequence is the major goal of a computer vision tasks. The first fully convolutional neural network for estimating optical flow was called FlowNet [13]. Subsequently, a series of works, including SpyNet [28], PWC-Net [29], LiteFlowNet [30], and RAFT [31] were proposed to reduce computational costs by using a coarse-to-fine and iterative estimation methodology. Recently, GMFlow [32] 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 [11]. 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 [11] and RAFT [31], another optical flow estimation models. Through this experiment, we confirm that there is no significant correlation between the accuracy of optical flow and the performance of our module.

2.3. Template Matching

Template matching is a task that finds a matching area in an image by comparing it with a template image [14]. Because it enables the detection of specific patterns or objects within an image, it is frequently used in image processing. To complement high-cost algorithms, Yan et al. [33] proposed a faster single-object tracking algorithm that makes use of parallel strategies. Kim et al. [34] also attempted to accelerate template matching by reducing the time complexity of block matching, which requires a lot of computational time to obtain results. Early template matching was implemented using simple similarity metrics such as Cross-Correlation (CC) [35, 36]. Recent template matching, however, makes use of NCC [15, 16] because CC is sensitive to brightness variations and image size. For more robust and accurate template matching that is less sensitive to brightness variations and image size, NCC enables one to compute and normalize correlations between pixels. To determine the degree to which the motion map and the attention map are similar, template matching with NCC is used in this study.

3. Proposed Method

In this paper, we propose the Motion Map Injection (MMI) Module to improve the editability of the text-to-video (T2V) editing model by injecting motion information into the attenion map. Before delving into the specifics, we first provide a overview of our framework depicted in Fig. 2. The MMI 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) [1] 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". Our proposed MMI module enhances the attention map by utilizing optical flow, information external to the T2V-model.

In Sec. 3.1, we introduce the attention map as prior knowledge. Additionally, we present the editing method employed in P2P. In Sec. 3.2, through actual experiments, we first explain that existing text-guided video editing models may fail to perform editing due to inaccurate attention map estimation. Subsequently, we introduce the proposed

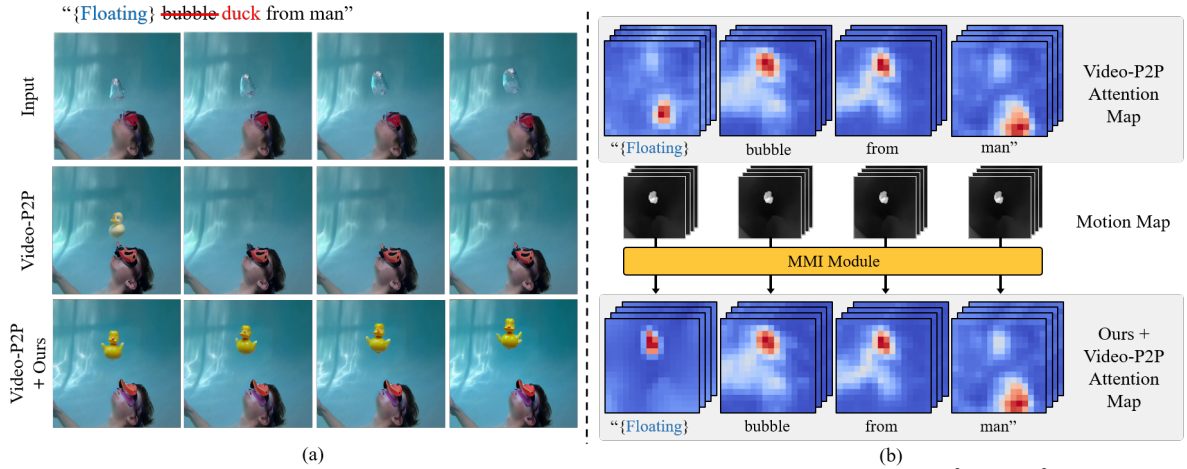

Figure 3. The existing T2V-model failed to estimate an accurate attention map for the motion prompt {Floating}, which resulted in a decrease in editability. The proposed MMI module enables the enhancement of the attention map, allowing for the accurate estimation of attention maps. Through this, it becomes possible to enhance editability. (a) is a comparison figure illustrating the output of Video-P2P for the input video, as well as the output when the MMI module is applied to Video-P2P. The result was more realistic editing when the MMI module was applied. (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.

MMI module, demonstrating its ability to enable accurate editing. In Sec. 3.3, we explain the two methods of the MMI module, namely "Motion Swapping" and "Weighted Sum using Correlation Score." In the final Section 3.4, we describe one of the methods for calculating the correlation score, which is the NCC, and provide an explanation of why we employ NCC. Note that the proposed MMI module can be applied to various models manipulating attention to perform video editing.

3.1. Preliminary

Attention Map Prompt-to-Prompt (P2P) [1] 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:

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

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.

Editing Method of Prompt-to-Prompt Recently, P2P [1] has been extended to the video domain, and research has been focused on various methods to maintain temporal consistency [5, 7–10, 23, 24]. 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.

Optical Flow We obtain the optical flow V_{flow} through the pre-trained optical flow estimation model [11]. 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 $\mathcal{V}_{flow}$ is as follows:

$$\mathcal{M} = \sqrt{\mathcal{V}_{\text{flow}}(:,:,0)^2 + \mathcal{V}_{\text{flow}}(:,:,1)^2}, \tag{4}$$

Algorithm 1: MMI 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 λ, denoising timestep in diffusion model t

Output: Enhanced attention of entire source prompt A^*

Def `MMI module`$(\mathcal{A}, \mathcal{M}, \mathcal{C}, i, j)$**:**
  # Weighted sum using correlation score
  for $k = 0,1,...,i$ **do**
    $\mathcal{A}_k^* = \mathcal{A}_k + \lambda \cdot \frac{\mathcal{C}_k \cdot \mathcal{M}}{t}$
  end
  # Motion swapping
  if $i = j$ **then** $\mathcal{A}_j^* = \lambda \cdot \mathcal{M}$
  return $\mathcal{A}^*$;

where $\mathcal{V}_{\text{flow}}(:,:,0)$ represents the component along the x-axis of $\mathcal{V}_{\text{flow}}$, and $\mathcal{V}_{\text{flow}}(:,:,1)$ represents the component along the y-axis of $\mathcal{V}_{\text{flow}}$.

3.2. Attention Map of Motion Prompt

In our observation of the attention map of the existing T2V-model, we captured that the attention map of the prompt representing the object was estimated relatively accurately, but the attention map of the prompt representing the movement was estimated inaccurately. In the upper part of Fig. 3 (b), it was confirmed that Video-P2P [8] was unable to accurately estimate the attention map of the motion prompt, "floating," leading to ineffective editing. With the MMI module, Video-P2P [8] captures accurate attention maps, as opposed to the previous results, which is evident in Fig. 3 (b). This results in accurate editing, as shown in the bottom line of Fig. 3 (a).

3.3. Motion Map Injection Module

In this section, we describe two methodologies the MMI module performs the attention map: "Motion Swapping" and "Weighted Sum using Correlation Score". The pseudo algorithm of our proposed MMI module is shown in Alg. 1.

Motion Swapping In Fig. 3, 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 MMI 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:

$$\mathcal{A}_{\mathcal{M}} = \lambda \cdot \mathcal{M}, \quad \text{if } \mathcal{P} = \mathcal{P}_{\mathcal{M}}, \tag{5}$$

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. 5 "Only 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.

Weighted Sum using Correlation Score The proposed MMI 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. Firstly, 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 template matching method. Template matching is a method of calculating similarity by comparing two images at the pixel-level. Among existing template matching methods, the MMI module uses NCC, which is a method of measuring the correlation score by normalizing two input images. A detailed description of NCC is provided in Sec. 3.4. 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:

$$\mathcal{A}_k^* = \mathcal{A}_k + \lambda \cdot \frac{(\mathcal{C}_k \cdot \mathcal{M})}{t}, \tag{6}$$

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.

However, even when the weighted sum 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 MMI module enhances the attention map by combining both the "Motion Swapping" method and the weighted sum method based on the correlation score, thereby improving the editability of the T2V model.

3.4. Normalized Cross Correlation for Correlation Score

When calculating the correlation score, the MMI module uses NCC. Template matching is a method that quantitatively measures how similar a template image is to different parts of a source image, aiming to find the region with the highest correlation. One of the metrics used in this method is NCC which normalizes two input images and then mea-

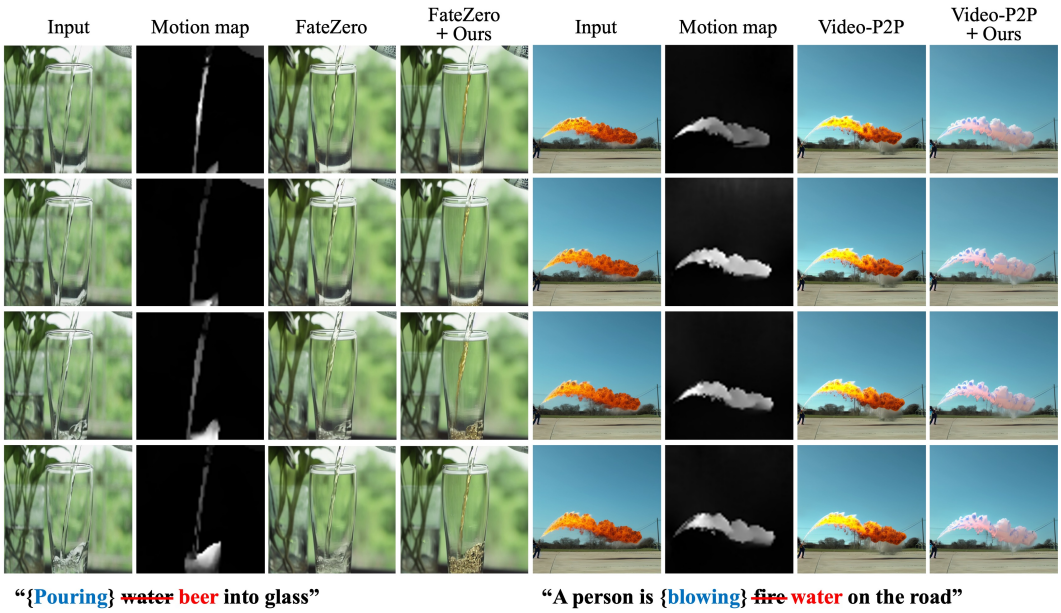

Figure 4. Qualitative results of our study. It shows the result of four examples corresponding to each target prompt. The first column of each example is the input video frames, the second column is the motion map extracted by the existing optical flow network [11], the third column is the output of each T2V model, the last colunm is output of each T2V model with MMI. It shows that the results of the model receiving the motion information of the video with MMI edited video better. It also shows good performance in editing not only fixed-shaped objects but also irregular objects (e.g. bubble, water, clouds).Additional results can be found in the supplementary materials.

sures the similarity pixel-wise. The formula is as follows:

$$\mathcal{C}_k = \frac{1}{n} \sum_{x,y} \frac{(\mathcal{M}(x,y) - \bar{\mathcal{M}})(\mathcal{A}_k(x,y) - \bar{\mathcal{A}}_k)}{\sigma_M \sigma_A}, \quad (7)$$

where n denotes the number of pixels, x, y denotes pixel of source image and target image, $\bar{\mathcal{M}}$ and $\bar{A}_k$ denote the average of pixels of $\mathcal{M}$ and A_k, σ_M and σ_A denote the standard deviation of pixels of $\mathcal{M}$ and A_k. k denotes the index of a particular word in the entire prompt.

4. Experiments

4.1. Experimental Setup

Baseline Model We used three text-guided video editing models, as a baseline: FateZero [9], Video-P2P [8], vid2vid [7]. These models can intuitively edit only with text without additional information by manipulating a cross-attention map that represents the relationship between text and video. We ran three text-guided video editing models to demonstrate the MMI module's applicability to other video editing models.

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 [9]. The number of video frames was set to 4 because this number is sufficient to demonstrate how well our method accomplishes our goal. (see supplementary material for 8 and 24 frames). The optical flow was extracted utilizing the UniMatch [11] model.(see supplementary material for further experiment on optical flow estimation model RAFT [31])

Evaluation Metrics Textual similarity and region preservation were assessed using the CLIP Score [37], as well as masked PSNR [8]. The CLIP Score [37] can be used to assess how well an image and a prompt correlate. It has been found to have a strong correlation with human judgment. The metrics used for assessing the level of structure preservation for the unintended region to be edited were the masked PSNR [8]. In this experiment, we regarded outside of the editing part as a mask. We used the No-Reference Image Quality Assessment methods, BRISQUE [38] and NIQE [39], to evaluate the quality of the edited videos.

4.2. Qualitative Results

Fig. 4 shows the qualitative results of the FateZero [9], Video-P2P [8], those with our proposed MMI module. Applying our module, motion information was injected into the attention maps of the entire prompt by estimating the motion of the frames. The attention map of "pouring" and "blowing" was not accurately estimated, so no editing was made for "water" and "fire". By extracting motion maps

Table 1. Left side represents the performance of the proposed module measured by CLIP Score, Masked PSNR, and BRISQUE metrics. Adding the proposed module to the FateZero [9], Video-P2P [8] and vid2vid-zero [7] models generally achieved higher scores on these metrics. The right side presents the results of a user preference survey. Users' preferences were recorded with higher ratings for the results when the proposed module was added.

	Metric			User Preference (%)		
	CLIP Score [37] ↑	Masked PSNR [8] ↑	BRISQUE [38] ↓	Text Alignment ↑	Structure Preserving ↑	Realism & Quality ↑
FateZero [9]	26.60	27.72	32.16	19.07	25.84	31.07
FateZero [9] + Ours	28.45 +1.85	27.30 -0.42	29.60 -2.56	80.92 +61.85	74.15 +48.31	68.92 +37.85
Video-P2P [8]	28.50	23.33	37.21	19.69	27.69	28.76
Video-P2P [8] + Ours	30.15 +1.65	24.92 +1.59	32.11 -5.10	80.30 +60.61	72.30 +48.61	71.26 +42.50
vid2vid-zero [7]	29.60	19.81	15.99	12.30	29.53	28.00
vid2vid-zero [7] + Ours	31.90 +2.30	20.15 +0.35	15.09 -0.90	87.69 +75.39	70.46 +40.93	72.00 +44.00

from the movement of frames and enhancing the overall attention map through our module, editing can be achieved using enhanced attention maps for the "beer" and "water". Therefore, It demonstrate that the general applicability and improved performance of the T2V editing model that controls the cross-attention layer.

4.3. Quantitative Results

We compared the results of 20 video data and three baseline models with three metrics. As demonstrated in Table 1 left, the CLIP Score [37] shows evidence that all three baseline models with our method understand target prompt better than those without our method. The performance of our framework was slightly higher for masked PSNR on Video-P2P [8] and vid2vid-zero [7]. This indicates that compared with the model applied MMI module, the existing model undertook more edits in areas that weren't intended. Indeed, the result suggests the capability to precisely editing the desired area by the user. Lower BRISQUE [38] scores in all three models with MMI module demonstrate that it can improves the video quality. Especially, MMI module works best on Video-P2P [8]. It was confirmed that editing performance was improved in various aspects through a module that improved attention maps.

4.4. 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 [9], Video-P2P [8], and vid2vid-zero [7], 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, and (3) quality. The results in Table 1 right demonstrate that the output of our module is the best. The supplementary material contains a detailed protocol for the user study.

Figure 5. Our module's comparative results w/ w/o the two methods on Video-P2P [8]. Our module attains improved editability through the utilization of both methods.

4.5. Ablation Study

The key part of our module is the motion swap and weighted sum using correlation score that incorporate the motion information from frames and attention maps.

Effect of MMI Module In Fig. 5, "Only 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 [8], but there is a limitation as overall attention is not improved. In "Only Weighted sum" incorporates motion information by adjusting with the correlation score between attention map and motion map, it shows better editing results than Video-P2P [8]. 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 weighted sum. By using both methods, as indicated by the results of our module, we can efficiently inject motion map information to the attention map.

Random Correlation Score Our module choose the weighted sum for the attention map from the motion map based on the template matching algorithm. To effectively enhance the attention map, it's crucial to adjust the scores while adding the motion map into the attention map, based on template matching. The motion map information should be adjusted based on similarity scores to enhance the attention of each prompt. Fig. 6 depicts the outcome obtained by applying our module with using random scores, without calculate similarity. This demonstrates how injecting motion information into "balloons" unrelated to "blowing" leads to

Figure 6. 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 MMI module to adjust the weights.

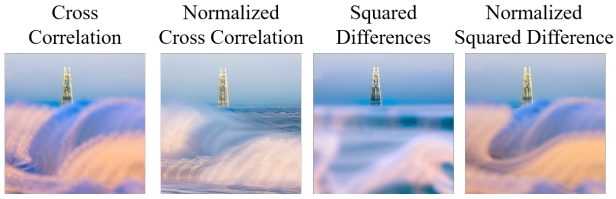

Figure 7. comparing the four template matching algorithms demonstrates that integrating normalized cross-correlation within our MMI module enables the most aligned editing with the prompt.

Table 2. Evaluation on 4 template matching algorithms using BRISQUE [38] and NIQE [39]

	BRISQUE [38]	NIQE [39]
Cross Correlation	47.87	14.92
Normalized Cross Correlation	**27.91**	**11.52**
Squared Differences	64.45	15.80
Normalized Squared Difference	52.48	13.91

a blending effect with the purple balloon. Therefore, this experiment emphasizes the significance of template matching in the weighted sum method.

Comparison of Various Template Matching Algorithm In Fig. 7, the corresponding outcomes of different template-matching algorithms, which compute the correlation between two images, are illustrated. As depicted in Fig. 7, "Normalized Cross Correlation", which aids in editing semantically by varying weights depending on the degree of association of the prompting word, is the most suitable function for enhancing semantic editing. In Table 2, template matcing algorithms were also measured through NIQE [39] and BRISQE [38] scores. "Normalized Cross Correlation" was demonstrated to be the most appropriate template matching algorithms by showing the highest quality

Figure 8. 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.

in both metrics. Additionally, we conducted experiments on two template matching algorithms and a total of nine metrics including SSIM, SAM, and MI. The results for these metrics are available for review in the supplementary materials.

Edit Results for Static Area In Fig. 8, the "forest" in the background should be edited, which is static area. However, without MMI 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.

5. Conclusion

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

References

[1] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-or. Prompt-to-prompt image editing with cross-attention control. In *The Eleventh International Conference on Learning Representations*, 2023. 1, 2, 3, 4

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

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

[4] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 1

[5] Eyal Molad, Eliahu Horwitz, Dani Valevski, Alex Rav Acha, Yossi Matias, Yael Pritch, Yaniv Leviathan, and Yedid Hoshen. Dreamix: Video diffusion models are general video editors. *arXiv preprint arXiv:2302.01329*, 2023. 1, 2, 4

[6] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Weixian Lei, Yuchao Gu, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. *arXiv preprint arXiv:2212.11565*, 2022.

[7] Wen Wang, Kangyang Xie, Zide Liu, Hao Chen, Yue Cao, Xinlong Wang, and Chunhua Shen. Zero-shot video editing using off-the-shelf image diffusion models. *arXiv preprint arXiv:2303.17599*, 2023. 1, 2, 4, 6, 7

[8] Shaoteng Liu, Yuechen Zhang, Wenbo Li, Zhe Lin, and Jiaya Jia. Video-p2p: Video editing with cross-attention control. *arXiv preprint arXiv:2303.04761*, 2023. 1, 2, 5, 6, 7

[9] Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. Fatezero: Fusing attentions for zero-shot text-based video editing. *arXiv preprint arXiv:2303.09535*, 2023. 1, 2, 6, 7

[10] Anant 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*, pages 3017–3026, 2023. 1, 2, 4

[11] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, Fisher Yu, Dacheng Tao, and Andreas Geiger. Unifying flow, stereo and depth estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2, 3, 4, 6

[12] Xiaoyu Shi, Zhaoyang Huang, Dasong Li, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, Jifeng Dai, and Hongsheng Li. Flowformer++: Masked cost volume autoencoding for pretraining optical flow estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1599–1610, 2023. 2

[13] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2758–2766, 2015. 2, 3

[14] Roberto Brunelli. *Template matching techniques in computer vision: theory and practice*. John Wiley & Sons, 2009. 2, 3

[15] MB Hisham, Shahrul Nizam Yaakob, RAA Raof, AB A Nazren, and NM Wafi. Template matching using sum of squared difference and normalized cross correlation. In *2015 IEEE student conference on research and development (SCOReD)*, pages 100–104. IEEE, 2015. 2, 3

[16] Kai Briechle and Uwe D Hanebeck. Template matching using fast normalized cross correlation. In *Optical Pattern Recognition XII*, volume 4387, pages 95–102. SPIE, 2001. 2, 3

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 2

[18] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2

[19] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 2

[20] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 2

[21] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 2

[22] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 2

[23] Min Zhao, Rongzhen Wang, Fan Bao, Chongxuan Li, and Jun Zhu. Controlvideo: Adding conditional control for one shot text-to-video editing. *arXiv preprint arXiv:2305.17098*, 2023. 2, 4

[24] Duygu Ceylan, Chun-Hao P Huang, and Niloy J Mitra. Pix2video: Video editing using image diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23206–23217, 2023. 2, 4

[25] Zhihao Hu and Dong Xu. Videocontrolnet: A motion-guided video-to-video translation framework by using diffusion model with controlnet. *arXiv preprint arXiv:2307.14073*, 2023. 2

[26] Ernie Chu, Tzuhsuan Huang, Shuo-Yen Lin, and Jun-Cheng Chen. Medm: Mediating image diffusion models for video-to-video translation with temporal correspondence guidance. *arXiv preprint arXiv:2308.10079*, 2023. 2

[27] Yuren Cong, Mengmeng Xu, Christian Simon, Shoufa Chen, Jiawei Ren, Yanping Xie, Juan-Manuel Perez-Rua, Bodo

Rosenhahn, Tao Xiang, and Sen He. Flatten: optical flow-guided attention for consistent text-to-video editing. *arXiv preprint arXiv:2310.05922*, 2023. 2

[28] Anurag Ranjan and Michael J Black. Optical flow estimation using a spatial pyramid network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4161–4170, 2017. 3

[29] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8934–8943, 2018. 3

[30] Tak-Wai Hui, Xiaoou Tang, and Chen Change Loy. Liteflownet: A lightweight convolutional neural network for optical flow estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8981–8989, 2018. 3

[31] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020. 3, 6

[32] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, and Dacheng Tao. Gmflow: Learning optical flow via global matching. In *Proceedings of the IEE E/CVF Conference on Computer Vision and Pattern Recognition*, pages 8121–8130, 2022. 3

[33] Baicheng Yan, Limin Xiao, Hang Zhang, Daliang Xu, Li Ruan, Zhaokai Wang, and Yiyang Zhang. An adaptive template matching-based single object tracking algorithm with parallel acceleration. *Journal of Visual Communication and Image Representation*, 64:102603, 2019. 3

[34] Seung-ho Kim, Sang-hyeob Song, Jong-hak Kim, Zhongyun Yuan, and Jun-dong Cho. Fast rotation-invariant template matching with candidate reduction using cuda. In *2015 International Symposium on Consumer Electronics (ISCE)*, pages 1–2, 2015. 3

[35] John P Lewis. Fast template matching. In *Vision interface*, volume 95, pages 15–19. Quebec City, QC, Canada, 1995. 3

[36] Roberto Brunelli and T Poggiot. Template matching: Matched spatial filters and beyond. *Pattern recognition*, 30(5):751–768, 1997. 3

[37] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 6, 7

[38] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708, 2012. 6, 7, 8

[39] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012. 6, 8