SAVE: Protagonist Diversification with Structure Agnostic Video Editing

Yeji Song[1], Wonsik Shin[1], Junsoo Lee[2], Jeesoo Kim[2], and Nojun Kwak[1†]

[1] Seoul National University
[2] NAVER Webtoon AI

Abstract. Driven by the upsurge progress in text-to-image (T2I) generation models, text-to-video (T2V) generation has experienced a significant advance as well. Accordingly, tasks such as modifying the object or changing the style in a video have been possible. However, previous works usually work well on trivial and consistent shapes, and easily collapse on a difficult target that has a largely different body shape from the original one. In this paper, we spot the bias problem in the existing video editing method that restricts the range of choices for the new protagonist and attempt to address this issue using the conventional image-level personalization method. We adopt motion personalization that isolates the motion from a single source video and then modifies the protagonist accordingly. To deal with the natural discrepancy between image and video, we propose a motion word with an inflated textual embedding to properly represent the motion in a source video. We also regulate the motion word to attend to proper motion-related areas by introducing a novel pseudo optical flow, efficiently computed from the pre-calculated attention maps. Finally, we decouple the motion from the appearance of the source video with an additional pseudo word. Extensive experiments demonstrate the editing capability of our method, taking a step toward more diverse and extensive video editing. Our project page: https://ldynx.github.io/SAVE/

Keywords: T2V model · Video Editing · Motion Personalization

1 Introduction

The remarkable advancements in text-to-image (T2I) generation models [13, 27, 30, 31, 33, 36] have prompted an increasing demand for the generation of imaginative scenes featuring user-supplied personalized concepts [7, 19, 32, 38, 40]. With these personalization methods, one can compose novel scenes with the desirable objects contained in various contexts *e.g.*, pictures of *the same* own dog traveling around the world. Some approaches [15, 35] have expanded personalized concepts into a higher level beyond substantial objects: a relation between objects [15] or even an image's style [35].

† Corresponding author

Fig. 1: Protagonist Diversification. We present the video editing method that replaces the protagonist of a source video with the one described by the editing prompt while maintaining the motion. Different from previous works, our method is able to cope with diverse protagonists with substantial changes in their body structure. While other methods either fail to follow the editing prompt (second row) or generate a different motion (see Fig. 5), ours achieves success in both challenges (third row).

On the other hand, there have been many video generation methods [12, 14, 34, 41] that adopt the weights of T2I models leveraging the extensive T2I prior knowledge from large-scale image datasets, and inflating a model architecture to address temporal consistency. On top of this architecture, there are several researches [6, 22, 26, 42, 52] proposed to edit the appearance and semantics of a given source video while preserving its geometry and dynamics.

While they provide encouraging results in terms of frame consistency, the capability of understanding and reproducing a motion is confined within certain limitations. As shown in Fig. 1 and Fig. 5, current methods for video editing [22, 29, 42] have difficulty in generating a new protagonist whose body structure deviates significantly from that of the original protagonist while faithfully following the motion in a source video. We have found that in the existing methods, cross-attention maps of a motion-related word (*e.g.* 'sleeping' in Fig. 1) are dispersed to regions that are not related to the motion as training proceeds. Therefore, learning and generating the specific motion in a source video become heavily dependent on temporal self-attention layers in a network architecture. However, the temporal self-attention layers only consider a temporal change in one latent pixel by its nature and cannot fully understand the spatial relationship among pixels. Although features in the deeper layers could access the broader

pixel space, they also fall short of a full understanding of spatial relationships. Due to inductive bias, convolution layers are constrained to focus on the center of the receptive field, resulting in these features, even in the deeper layers, failing to grasp spatial relationships that involve distant pixels from the center. This leads to a limited video editing capability when a new protagonist has a different shape and arrangement.

In contrast, attention layers explicitly contemplate all the relationships between input tokens. We incorporate cross-attention layers to learn the motion along with accurate spatial information. More specifically, we relieve the burden of temporal self-attention and hand over the role to a more appropriate component, the word embedding vector, to capture the motion. It can also be viewed as reinterpreting the protagonist editing task as a motion inversion problem and establish the following two goals: (1) to broaden personalized concepts expanded to ***a motion*** in a source video and (2) to generate a conceptualized motion with various contexts *i.e.* protagonists. We introduce a new motion word (S_{mot}) that describes a specific motion performed by a protagonist in a source video. We have two advantages of utilizing S_{mot} in editing a protagonist across a wider spectrum. First, at training, features of this motion word are injected in the cross-attention based on a calculation of the attention map over a spatial axis. Therefore, the spatial characteristics of the learned motion can be fully explored during training. Second, at inference, an embedding vector of a motion word exchanges information with another embedding vector of the protagonist word via the text encoder layer in an early stage. Then, the overall structure of a new protagonist in the motion can be determined from the start, allowing a natural movement in the edited video. The temporal self-attention layers, meanwhile, can concentrate more on a temporal change in each latent pixel.

However, a pseudo-word [7,15,38,40] designed for image-level personalization leads to the discrepancy between image and video. Therefore, we expand the temporal axis of the textual embedding space that enables S_{mot} to find a proper cross-attention map on moving areas. Moreover, to let S_{mot} make an effect on a motion-related region and encourage effective motion learning, we introduce a novel *pseudo* optical flow. From pre-calculated spatio-temporal attention maps, we track the semantically same pixels across frames and estimate the flows in a source video without requiring an extra optical flow model that incurrs additional expenses. Using this pseudo optical flow, we specify a motion-related region and better involve S_{mot} in this area. During training, we also adopt an additional pseudo-word S_{pro} representing the protagonist in a source video. Training S_{mot} after registering S_{pro} into the model's dictionary alleviates the entanglement between the motion and the protagonist. As shown in the last row of Fig. 1 and Fig. 5, while faithfully following the specific motion in the given single source video, our method flexibly covers a broad range of the protagonist editing.

Our contributions can be summarized as follows:

- We are the first to identify the problem with a motion-related word and the following limitations in the existing video editing methods. We further enhance editing capabilities, effectively resolving the problem.

- We reformulate the video editing task as a motion inversion problem and introduce the novel concept of extending a pseudo word along the temporal axis. This approach has never been explored in this domain.
- We also propose pseudo optical flow and pre-registration of a protagonist pseudo-word S_{pro}. They are general methods dealing with critical topics in the overall video editing task *e.g.* estimating optical flow without a heavy computation burden and decoupling the motion from the protagonist.

2 Related Work

Video Diffusion Models. Leveraging the extensive prior knowledge of the image diffusion model has led to research endeavors [5, 9, 12, 14, 16, 23, 24, 34, 48, 49, 51] aimed at generating high-quality videos. In video generation, it is crucial to maintain consistency across generated frames–*temporal consistency*. As the image diffusion model has not learned information pertaining to temporality, these methods focus on injecting temporal information into a model architecture while retaining the existing wealth of spatial knowledge. Research also extends beyond video generation, delving into the realm of video editing and enabling straightforward modifications of videos provided by users [3,6,21,22,26,29,42,52]. Recently, zero-shot editing approaches have been proposed [2, 8, 17, 20, 50] that alter the overall style of the video. The outcomes of these studies are closely tied to the structure of the input video with minimal changes to the structure of the output rather than directly modeling the motion information. A significant distinction in our work lies in the potential for video editing even when there is a substantial alteration in the body structure of objects, setting it apart from the aforementioned studies. Concurrent to our work, there are methods [44, 53] to achieve generalization of the motion pattern from a given set of video clips. Our work focuses on customizing a unique, specific motion in a single source video.

Personalization. In the diverse research endeavors, methodologies for generating *specific concepts* beyond proficient creation have been studied in parallel. The most representative form is to generate a specific object a user gives. To this end, previous studies have investigated where and to what extent weights of the pre-trained model should be finetuned: a special token [7], text encoder [40], keys of attention layers [38], and the whole model [32]. There is research [1, 45] exploring the combination of various concepts. Moreover, research extends beyond the scope of objects such as style [35] or relationship [15]. In this paper, we aim to integrate personalization with the video domain by endowing the capability of editing the protagonist while preserving the original motion information within a source video.

3 Preliminaries

Latent Diffusion Models (LDMs). LDMs are diffusion models that recursively denoise an image latent code z_t (backward process) into the previous

timestep image latent code z_{t-1} which is generated by repetitively adding noise to z_0 (forward process). If needed, we can add a condition $\mathcal{C}$ during the backward process and the objective of LDMs can be defined as minimizing the following LDM loss:

$$\mathcal{L}_{\text{ldm}} = \mathbb{E}_{z_0,\epsilon\sim\mathcal{N}(0,I),t\sim\text{Uniform}(1,T)} \left\|\epsilon - \epsilon_\theta(z_t, t, \mathcal{C})\right\|_2^2, \tag{1}$$

where z_t is the latent code of timestep t and ϵ is a random noise that the model $\epsilon_\theta(\cdot)$ should predict.

Text-to-Video Diffusion Models (T2V). To generate videos, the Video Diffusion model ϵ_θ is structured with 3D UNet architecture [4, 9, 22, 42, 50]. Within the 3D UNet, there are layers dedicated to handling spatial and temporal information. The spatial layers are initialized from the weights of the image diffusion model, leveraging its extensive knowledge. We employed *spatio-temporal* attention (ST-Attn) in the first layer of UNet block, a method that considers the first frame along with the preceding frame when generating each frame. Meanwhile, *temporal self-attention* (T-Attn) layers in the last layer of UNet block aim to align frames by processing videos in a temporal dimension. These approaches aid in maintaining temporal consistency.

4 Method

Our goal is to introduce a way of enabling a broader range of choices for a new protagonist in video editing task. We first discover the underlying causes of why the existing methods struggle to change a protagonist into a new object that is dissimilar in shape and arrangement (Sec. 4.1). Then, we introduce our method that utilizes a new motion word to transcend the limitation of the existing works (Sec. 4.2) as well as an additional regularization term and a training strategy that guides the motion word to effectively learn the motion in a source video (Sec. 4.3 and Sec. 4.4). The overall pipeline is illustrated in Fig. 3.

4.1 Location Bias of Motion-Related Words

As shown in Fig. 1 and Fig. 5, the existing methods fail to produce a new, differently structured protagonist appropriately following the motion in the source video. To identify the causes, we start by investigating a bias in cross-attention maps of motion-related words. We have found that those words often highlight specific parts of the protagonist unrelated to the motion. As shown in the first row of Fig. 2, cross-attention maps of motion-related words (*e.g.* 'roaring') scatter over the neck of a cat. We define this phenomenon as *location bias*.

Why does this bias appear especially in motion-related words? The pretrained text encoder is usually trained on large-scale text-image datasets and this makes the encoder to embed motion-related words based on how that motion is commonly depicted in an image. (*e.g.* 'roaring' is related to images with a mouth open wide to the object's neck). However, in a video, motion proceeds in the

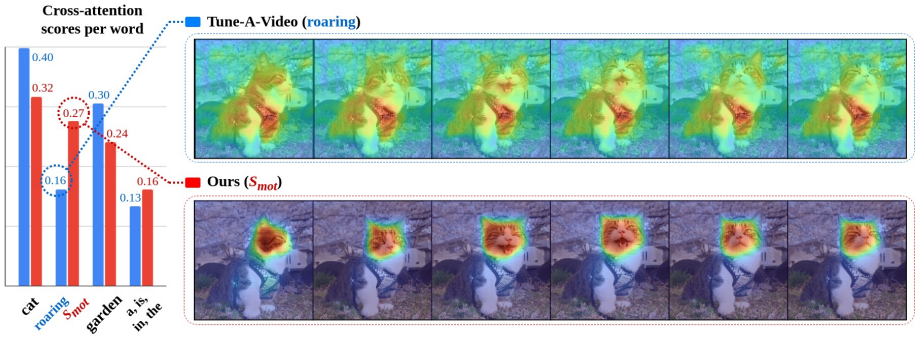

Fig. 2: Cross-attention scores and the attention maps of the motion-related word. Using the existing method [42] and ours, we compute the cross-attention scores of each semantic token and visualize regions to which the motion-related word ('roaring' and S_{mot}) attends. As in the left graph, little information about 'roaring' is used in [42] compared to the other tokens *e.g.* nouns (cat and garden). This is because 'roaring' attends to an inaccurate region due to location bias. Meanwhile, our method actively utilizes S_{mot} which provides more accurate information about the motion attending to the proper facial regions.

stream of time where motion-location relation exists only in several specific frames (*e.g.* 'roaring' video also includes frames where the mouth is closed). Naturally, the cross-attention map of those words lies on unnatural areas.

As shown in the upper row of Fig. 2, location biases produce inaccurate features injecting textual information in an incorrect position. The model, consequently, finds an alternative to represent the motion instead of utilizing the motion-related word, and heavily relies on T-Attn layers. The blue bars in left graph of Fig. 2 shows that textual information participates little in the motion generation process (only 16% of total attention is on the verb term). In Fig. 7, we also show that the model largely depends on T-Attn layers where the motion cannot be reconstructed without training T-Attn layers.

Meanwhile, T-Attn layers remain tangential to the spatial axis by its nature. When the input has B batch size, N sequence length, and $H \times W$ spatial dimension, they treat $B \cdot H \cdot W$ encoded vectors independently exchanging information only among N features in the same latent pixel. For the k-th pixel on a latent code from the i-th frame $z^{i,k}$ as the query, T-Attn is calculated with $z^{j,k}, j \in \{1, \cdots, N\}/i$ as key.

$$Q = W^Q z^{i,k}, K = W^K z^{j,k}, V = W^V z^{j,k} \tag{2}$$

where W^Q, W^K and W^V are projection matrices. Therefore, when T-Attn layers encounter a protagonist with a new body structure, they struggle to reproduce a proper motion on the protagonist. In 2–4th row of Fig. 5-left, for example, a newly generated dog cannot achieve a learned motion around the head and the mouth since the original protagonist (a cat) and edited protagonist (a dog) have a dissimilar facial arrangement and mouth shape.

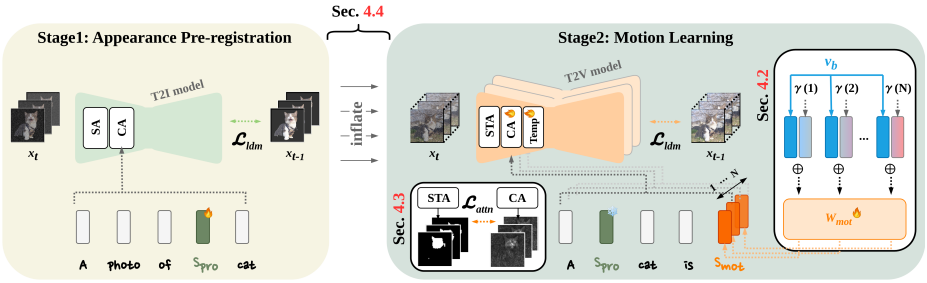

Fig. 3: The overall training pipeline of our method. Using expanded text embeddings of S_{mot}, we optimize W_{mot} that maps embeddings of the original motion word ('roaring') to a specific motion in a source video. Under cross-attention regularization, S_{mot} is optimized to primarily focus on the moving area while pre-registered S_{pro} disentangles the appearance from the motion. This dual approach facilitates S_{mot} in effectively learning the motion.

4.2 Expanded Text Embeddings with Time

As we aim to put a specific motion on various types of protagonists, we focus on the approach to revitalizing a motion-related word *by reducing location biases*. To this end, we expand the textual embedding space of a motion word to represent a time flow in videos rather than a frozen moment in images: we add a temporal axis to an embedding space of our new motion word (S_{mot}) and let S_{mot} inject its information into a proper region in *each frame*.

To formulate embedding vectors of S_{mot}, we have two separate components to take on different roles. The first component utilizes prior knowledge of the T2I model and represents the common aspect of the motion across frames, conveying its information to the second component. The second component learns the residual motion for each frame and encodes the overall motion in the video. The embedding vectors of S_{mot} for N video frames can be gained as follows:

$$v^i_{mot} = W_{mot}\ (v_b \oplus \gamma(i)), \quad i \in \{1, \cdots, N\} \tag{3}$$

where v_b is a textual embedding of the original motion-related word in a source prompt and W_{mot} is learnable linear layers. We denote concatenation operation and positional encoding as $\oplus$ and $\gamma(\cdot)$ respectively. v^i_{mot} is then treated like other embedding vectors and passed through a text encoder. Contrary to conditioning a single text embedding to N latent codes in the existing text-to-video methods, we provide N text embeddings, $v^i_{mot},\ i \in \{1, \cdots, N\}$ to each latent code of the corresponding frame.

As shown in the second row of Fig. 2, with consideration for a time flow, our S_{mot} renders more accurate features via cross-attention whose attention maps accurately attend to moving areas of each frame. Note that despite an increasing number of video frames, the number of learnable parameters stays fixed, enabling video editing in various lengths.

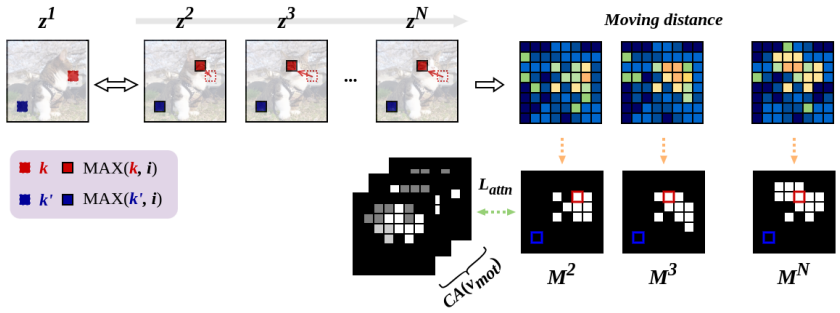

Fig. 4: Cross-attention regularization on S_{mot}. We assume that a pixel $z^{1,k}$ in the moving areas becomes distant from a MAX(k,i)-th pixel in z^i that has the maximum attention score while a pixel $z^{1,k'}$ in the static areas stays close to MAX(k',i)-th pixel in z^i. Therefore, we calculate the distance between k and MAX(k,i) for all k and generate motion masks $M_2, \cdots M_N$.

4.3 Motion Aware Cross-attention Loss

A cross-attention map of a newly learned concept is likely to pervade the whole scene, easily overfitting to the background [38]. While a video is composed of various dynamics *e.g.* camera moving or background movement, we want S_{mot} to learn specifically the motion of the protagonist. Inspired by [1,45], we constrain a cross-attention map of S_{mot} to focus exclusively on the protagonist. However, forcing S_{mot} to pay attention on the entire protagonist obscures what S_{mot} needs to learn *i.e.* the motion of the protagonist.

To resolve this issue, we introduce a motion-aware cross attention loss enabling S_{mot} to focus on the movement of the protagonist. Specifically, we define motion area as the union of pixels whose optical flow have a positive magnitude. However, existing optical flow estimation models [37,46] either require additional memory usage or involve iterative refine processes, which are unsuitable for training already resource-intensive video diffusion models. Therefore, we introduce a novel pseudo optical flow to better represent the moving area without using the optical flow models. We first collect pre-calculated ST-Attn maps from specific decoder layers. In ST-Attn mechanism, the i-th attention map is computed by using the i-th frame as the query and the first frame as the key. When we denote a latent code from the i-th frame as z^i, attention scores $SA(z^1, z^i)$ represent a similarity between z^1 and z^i. Our intuition lies in that if the k-th pixel of the first frame $z^{1,k}$ and the l-th pixel of the i-th frame $z^{i,l}$ have a high attention score, then they tend to be the same semantic point at different frames, *i.e.* a point that has been in k-th location at the first frame moves to the l-th location at the i-th frame. By tracking down the spatial locations of these similar points across frames, we can estimate the temporal flow of each pixel in the video.

To find the points that are likely to be the same, we store spatial locations MAX$(k,i) \in [0, h-1] \times [0, w-1]$, indicating that the k-th pixel in z^1 has the

maximum scores with the MAX(k, i)-th pixel in z^i. Here, c, h, and w indicate the channel size and spatial dimensions of z^i respectively. Then, we calculate a distance between MAX(k, i) and k. When two locations are close to each other, the object in $z^{1,k}$ can be regarded as mostly stationary until the i-th frame. On the other hand, if MAX(k, i) is far from k, then the object is highly likely to be largely moving in the video. We go through the same process for each pixel in the i-th frame and produce a map with the estimated distance which amounts to the moving distance of the pixel from the first frame. Eventually, we extract masks of motion-related area for each frame $i \in \{2, \cdots, N\}$ by retaining only pixels with a large moving distance. We denote the masks as $M = \{M^2, \cdots, M^N\}$. Utilizing these motion masks, we add the following regularization term to encourage cross-attention map of S_{mot} to follow M:

$$\mathcal{L}_{\text{attn}} = \frac{1}{N-1} \sum_{i=2}^{N} \|CA(z^i, v^i_{mot}) - M^i\|_2^2 \tag{4}$$

Adopting Eq. 1 and Eq. 4, the overall optimization objective is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{ldm}} + \lambda_{\text{attn}} \mathcal{L}_{\text{attn}} \tag{5}$$

where λ_{attn} is a hyperparameter balancing between a reconstructive ability and a motion focusing. We illustrate the overall regulating process in Fig. 4. Some flow estimation models [46] share a similar concept to viewing optical flow estimation as a feature-matching problem. Meanwhile, the novelty of our pseudo optical flow lies in its ability to estimate flows without incurring extra costs by adapting pre-computed self-attention maps to flow estimation in video diffusion models.

4.4 Appearance Pre-registration Strategy

To resolve the problem that the motion and the protagonist get easily entangled, we propose a two-stage training strategy to separate the two properties. We newly define a pseudo-word S_{pro} that represents the appearance and texture features of the protagonist. At the first stage, we find a text embedding of S_{pro}, namely v_{pro}, in the textual embedding space before inflating T2I models. v_{pro} is optimized with the LDM loss as in Eq. 1 considering video frames as batch of images. At the second stage, we inflate the T2I model to a T2V model and optimize W_{mot} and v_b in Eq. 3. As the protagonist and its appearances are already registered in the text encoder, v_{mot} can be effectively learned using disentangled motion information for the video.

5 Experiments

Dataset. We evaluate our method and baselines on videos collected from DAVIS dataset [28] and YouTube [22] following previous works [3, 22, 42]. Each video consists of 8–32 frames at the resolution of 512 × 512. As one of the important

Fig. 5: Protagonist editing results comparing ours with baselines. Our method successfully reproduces the motion in a source video identifying S_{mot} as both the head and the mouth movements (left column) while editing a protagonist faithfully with a natural video of a raccoon *doing* S_{mot} (right column). Meanwhile, other baselines either generate an inaccurate motion *e.g.* a dog opening its mouth across all frames, or fail to combine a new protagonist with the motion resulting in incomplete editing.

aspects to evaluate is the editing ability of a protagonist with large structural changes, we additionally provide object-changed prompts for hard cases (*e.g.* changing a cat in a source video to Pikachu). Ultimately, we composed 48 pairs of videos and text prompts to evaluate. We also conduct additional experiments on the open-sourced benchmark released by the LOVEU-TGVE competition at CVPR 2023 [43]. More details and qualitative results are in the Supplementary.

Baselines. We compare our method with the state-of-the-art video editing and generation approaches. *(1) Tune-A-Video (TAV)* [42] is the conventional video editing method that fine-tunes the inflated T2I model on a given source video. *(2) Video-P2P* [22] improves upon TAV applying Prompt-to-Prompt [10] and Null-text Inversion [25]. *(3) Fate-Zero* [29] proposes to blend the attention maps stored during inversion. Following the one-shot editing version, we adopt the

Table 1: Quantitative comparison on videos collected from DAVIS dataset [3, 28, 42] and YouTube [22].

Method	Automated Metrics		
	CLIP-Txt ↑	CLIP-Img ↑	Flow sim. ↑
Tune-A-Video [42]	25.84	93.37	55.61
Video-P2P [22]	25.27	93.89	64.59
Fate-Zero [29]	24.14	94.05	**80.92**
Ours	**25.99**	**94.21**	79.10

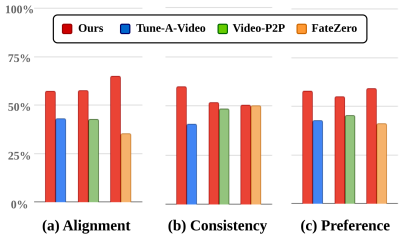

Fig. 6: User study results. Our method is preferred over other baseline methods across all evaluation criteria.

TAV weights pretrained on the source video when evaluating Fate-Zero. We exclude zero-shot methods [6, 8, 47] from our baselines since they exhibit weakness in video editing with a structure change as a trade-off for the efficiency. During inference, we also adopt P2P technique from Video-P2P to maintain the background of the source video as the motion word learns to focus on the motion.

Metrics. In line with previous works [6, 39, 42], we evaluate the baselines using the pretrained CLIP [11] model as follows: *(1) CLIP-Text similarity* is the average CLIP score between frames of generated video and the corresponding editing prompt, representing a textual alignment of the outputs. We evaluate the model's ability to edit protagonists using this metric. *(2) CLIP-Image similarity* computes cosine similarity between the CLIP image embeddings of pairs of video frames, representing frame consistency. To score faithfulness to the motion of the source video, we measure *(3) Flow similarity* that computes cosine similarity between optical flows of the source and the edited video using the estimation model [37]. This metric reflects how well motion is preserved from the source video. We further evaluate the methods through five human raters for each example conducted with Amazon Mechanical Turk. The following three questions were asked. *(1) Textual alignment*: "Which video better matches the text?" *(2) Consistency*: "Which video has higher consistency?" *(3) Preference*: "From the perspective of video editing, which video do you prefer?"

5.1 Qualitative results

Fig. 5 illustrates qualitative comparisons between our method and baselines. As our method (bottom row) effectively learns the motion of the original protagonist, it generates a new protagonist that reproduces the motion in the source video seamlessly despite having a significantly different structure from that of the original one. For example, as shown in Fig. 5-left, our method ables to grasp the accurate motion in the source video and successfully associates those movements to S_{mot}. Meanwhile, other baselines commonly miss the mouth movements. TAV [42] also produces an inaccurate head motion in the third frame. Our method effectively reflects the editing prompts compared to other baselines

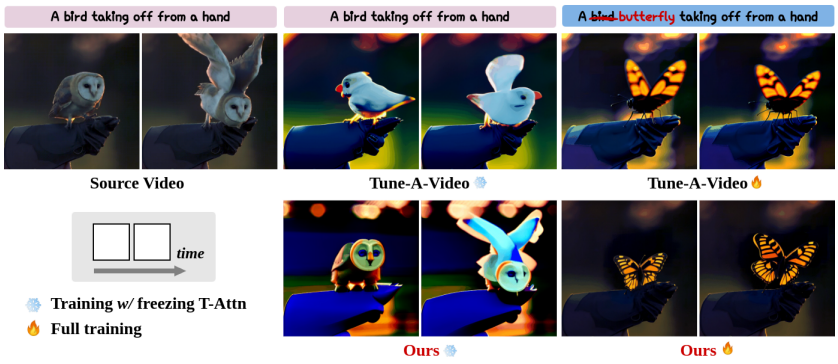

Fig. 7: Analysis on motion learning and additional editing results. As S_{mot} effectively learns the motion in a source video, our method still reproduces a proper motion with frozen T-Attn layers, resulting the more flexible editing.

as shown in the right column of Fig. 5. Baseline methods are unable to overcome the discrepancy between 'man' and 'raccoon' in the structure and generate a raccoon video where certain segments maintain the man's appearance. On the other hand, our method disentangles the appearance and the motion with separate S_{pro} and S_{mot} and renders a new raccoon doing S_{mot} from the text encoder from the start, successfully applying the motion features to a new protagonist.

5.2 Quantitative results

In Tab. 1 and Fig. 6, our method quantitatively exhibits the highest ability in textual alignment, temporal consistency, and user preferences. Our method closely associates the learned motion with a new protagonist, successfully generating a natural video that is faithfully aligned with an editing prompt. Meanwhile, Fate-Zero [29] shows the highest flow similarity and frame consistency on par with our method. As shown in Fig. 5 and Supplementary, when an editing prompt requires a large structural change for a new protagonist, Fate-Zero shows a tendency to adhere closely to the source video. This leads to low CLIP-Text scores in Tab. 1 and less voted Alignment in Fig. 6, while the edited video still achieves a highly similar optical flow to that of the source video and a high frame consistency. On the other hand, our method attains high scores in *both* text alignment and flow similarity & consistency demonstrating a general editing ability.

5.3 Analysis

To demonstrate that incorporating S_{mot} actually alleviates the burden on the T-Attn layers in learning the motion, we conduct the following experiment: we freeze the T-Attn layers in TAV and our method respectively when training the networks. After training, we reconstruct the source video using the motion-related words ('taking off' and S_{mot}). As shown in the left two columns in Fig. 7,

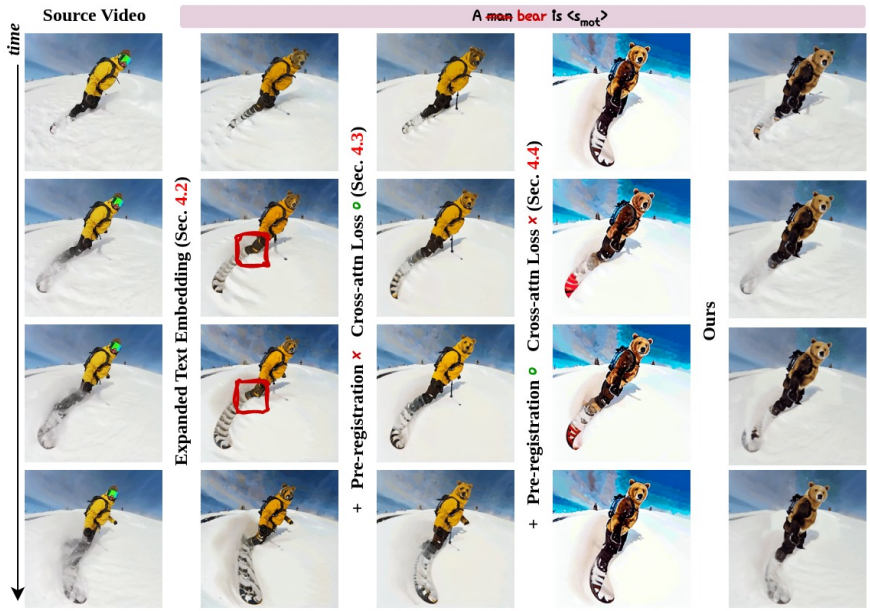

Fig. 8: The impact of each component. By the temporally expanded text embedding, S_{mot}, our method is able to learn the motion across frames. Cross-attention regularization and pre-registration of the protagonist word further enhance an ability of S_{mot} to understand the motion.

Table 2: Quantitative ablation on each component.

	CLIP-Txt	CLIP-Img	Flow sim.		CLIP-Txt	CLIP-Img	Flow sim.
Expanded Emb.	24.72	90.16	77.53	*w/* Pre-reg.	25.38	89.93	77.89
w/ Cross-attn $\mathcal{L}$	24.86	91.20	**79.83**	Ours	**25.99**	**94.21**	79.10

TAV cannot properly reproduce a motion in the source video, heavily depending on the T-Attn layers in regard of learning the motion. On the other hand, our method is able to generate the accurate motion in the source video by using S_{mot}. The right two columns in Fig. 7 indicate editing results from each fully-trained method. Our method successfully renders a new protagonist undergoing structural modifications (*e.g.* editing a bird to a butterfly) as S_{mot} actively exploits the spatial information of the motion.

5.4 Ablation Studies

We isolate each component in our method and verify the effect respectively. As shown in the second column in Fig. 8, with temporally expanded text embedding v_{mot}, a new protagonist in the edited video well follows the overall pose of the original protagonist in the source video. However, the new protagonist exhibits

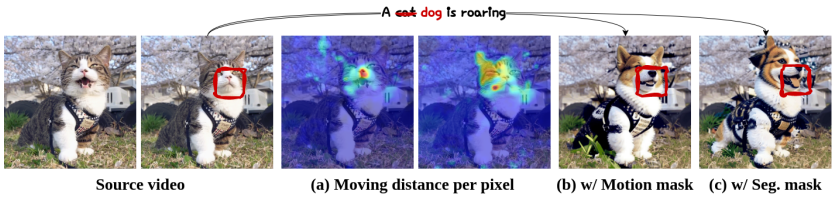

Fig. 9: Ablation on motion masks M. Encouraging S_{mot} to focus on motion masks guides S_{mot} to learn the motion more effectively than using segmentation masks alone.

an awkward leg appearance while the movement is also slightly different. By adopting cross-attention regularization, as shown in the third column, a generated bear retains more accurate movements in the source video. Meanwhile, pre-registration of S_{pro} effectively decouples the motion from the appearance and generates a new protagonist faithfully as shown in the fourth column. The last column indicates the results of our method which disentangles the motion from the appearance in a source video and effectively digests the information on the motion. We report quantitative results in Tab. 2. A Slight decrease in Flow similarity in our method compared to applying only cross-attention loss would have resulted from improved text alignment, *i.e.*, editing the protagonist successfully and changing the corresponding flows. We also refer to Supplementary for ablations on other examples.

We also investigate the efficiency of our motion masks M introduced in Sec. 4.3. Fig. 9 illustrates an estimated moving distance of each pixel and our results using the motion masks M in (a) and (b) respectively while comparing the results using the object segmentation masks [18] instead in (c). Narrowing down the moving area with these motion masks effectively guides S_{mot} to focus on the motion itself.

6 Conclusion

In this paper, we propose a new method to diversify a protagonist that reproduces the motion in a source video. We first reveal a location bias in the existing methods that hinders flexible editing. To resolve this problem, we introduce a motion word that encompasses temporal relationships among frames. We also adopt a couple of approaches to effectively train the motion word focusing on a target motion. Our method paves the way for broader editing, enriching the video editing task.

Limitation & Future work. We found that our method struggles to learn the motion of multiple protagonists as can be found in the failure cases in the Supplementary. Also, in specific cases, video P2P technique [22] during inference results in some artifacts around the protagonist if some incorrect attention maps occur. Future works to expand into broader movements and alleviate the artifact will be an interesting topic.

Acknowledgments.

This work was supported by NRF (2021R1A2C3006659), KOCCA (RS-2024-00398320) and IITP (RS-2021-II211343), all funded by the Korean Government.

References

1. Avrahami, O., Aberman, K., Fried, O., Cohen-Or, D., Lischinski, D.: Break-a-scene: Extracting multiple concepts from a single image. arXiv preprint arXiv:2305.16311 (2023)
2. Bai, J., He, T., Wang, Y., Guo, J., Hu, H., Liu, Z., Bian, J.: Uniedit: A unified tuning-free framework for video motion and appearance editing. arXiv preprint arXiv:2402.13185 (2024)
3. Bar-Tal, O., Ofri-Amar, D., Fridman, R., Kasten, Y., Dekel, T.: Text2live: Text-driven layered image and video editing. In: European conference on computer vision. pp. 707–723. Springer (2022)
4. Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S.W., Fidler, S., Kreis, K.: Align your latents: High-resolution video synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22563–22575 (2023)
5. Chen, H., Zhang, Y., Cun, X., Xia, M., Wang, X., Weng, C., Shan, Y.: Videocrafter2: Overcoming data limitations for high-quality video diffusion models (2024)
6. Esser, P., Chiu, J., Atighehchian, P., Granskog, J., Germanidis, A.: Structure and content-guided video synthesis with diffusion models. arXiv preprint arXiv:2302.03011 (2023)
7. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion (2023)
8. Geyer, M., Bar-Tal, O., Bagon, S., Dekel, T.: Tokenflow: Consistent diffusion features for consistent video editing. arXiv preprint arXiv:2307.10373 (2023)
9. Guo, Y., Yang, C., Rao, A., Wang, Y., Qiao, Y., Lin, D., Dai, B.: Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. arXiv preprint arXiv:2307.04725 (2023)
10. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control (2023)
11. Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y.: Clipscore: A reference-free evaluation metric for image captioning. arXiv preprint arXiv:2104.08718 (2021)
12. Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D.P., Poole, B., Norouzi, M., Fleet, D.J., et al.: Imagen video: High definition video generation with diffusion models. arXiv preprint arXiv:2210.02303 (2022)
13. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
14. Hong, W., Ding, M., Zheng, W., Liu, X., Tang, J.: Cogvideo: Large-scale pretraining for text-to-video generation via transformers. arXiv preprint arXiv:2205.15868 (2022)
15. Huang, Z., Wu, T., Jiang, Y., Chan, K.C., Liu, Z.: Reversion: Diffusion-based relation inversion from images. arXiv preprint arXiv:2303.13495 (2023)

16. Jin, Y., Sun, Z., Xu, K., Xu, K., Chen, L., Jiang, H., Huang, Q., Song, C., Liu, Y., Zhang, D., Song, Y., Gai, K., Mu, Y.: Video-lavit: Unified video-language pre-training with decoupled visual-motional tokenization (2024)
17. Kara, O., Kurtkaya, B., Yesiltepe, H., Rehg, J.M., Yanardag, P.: Rave: Randomized noise shuffling for fast and consistent video editing with diffusion models (2023)
18. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023)
19. Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.Y.: Multi-concept customization of text-to-image diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1931–1941 (2023)
20. Li, X., Ma, C., Yang, X., Yang, M.H.: Vidtome: Video token merging for zero-shot video editing (2023)
21. Liang, F., Wu, B., Wang, J., Yu, L., Li, K., Zhao, Y., Misra, I., Huang, J.B., Zhang, P., Vajda, P., Marculescu, D.: Flowvid: Taming imperfect optical flows for consistent video-to-video synthesis (2023)
22. Liu, S., Zhang, Y., Li, W., Lin, Z., Jia, J.: Video-p2p: Video editing with cross-attention control. arXiv preprint arXiv:2303.04761 (2023)
23. Ma, X., Wang, Y., Jia, G., Chen, X., Liu, Z., Li, Y.F., Chen, C., Qiao, Y.: Latte: Latent diffusion transformer for video generation. arXiv preprint arXiv:2401.03048 (2024)
24. Ma, Z., Zhou, D., Yeh, C.H., Wang, X.S., Li, X., Yang, H., Dong, Z., Keutzer, K., Feng, J.: Magic-me: Identity-specific video customized diffusion (2024)
25. Mokady, R., Hertz, A., Aberman, K., Pritch, Y., Cohen-Or, D.: Null-text inversion for editing real images using guided diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6038–6047 (2023)
26. Molad, E., Horwitz, E., Valevski, D., Acha, A.R., Matias, Y., Pritch, Y., Leviathan, Y., Hoshen, Y.: Dreamix: Video diffusion models are general video editors. arXiv preprint arXiv:2302.01329 (2023)
27. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741 (2021)
28. Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., Sorkine-Hornung, A.: A benchmark dataset and evaluation methodology for video object segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 724–732 (2016)
29. Qi, C., Cun, X., Zhang, Y., Lei, C., Wang, X., Shan, Y., Chen, Q.: Fatezero: Fusing attentions for zero-shot text-based video editing (2023)
30. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 **1**(2), 3 (2022)
31. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
32. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22500–22510 (2023)

33. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems **35**, 36479–36494 (2022)
34. Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., et al.: Make-a-video: Text-to-video generation without text-video data. arXiv preprint arXiv:2209.14792 (2022)
35. Sohn, K., Ruiz, N., Lee, K., Chin, D.C., Blok, I., Chang, H., Barber, J., Jiang, L., Entis, G., Li, Y., et al.: Styledrop: Text-to-image generation in any style. arXiv preprint arXiv:2306.00983 (2023)
36. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations (2021)
37. Teed, Z., Deng, J.: 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. pp. 402–419. Springer (2020)
38. Tewel, Y., Gal, R., Chechik, G., Atzmon, Y.: Key-locked rank one editing for text-to-image personalization. In: ACM SIGGRAPH 2023 Conference Proceedings. pp. 1–11 (2023)
39. Wang, W., Xie, K., Liu, Z., Chen, H., Cao, Y., Wang, X., Shen, C.: Zero-shot video editing using off-the-shelf image diffusion models. arXiv preprint arXiv:2303.17599 (2023)
40. Wei, Y., Zhang, Y., Ji, Z., Bai, J., Zhang, L., Zuo, W.: Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation (2023)
41. Wu, C., Liang, J., Ji, L., Yang, F., Fang, Y., Jiang, D., Duan, N.: Nüwa: Visual synthesis pre-training for neural visual world creation. In: European conference on computer vision. pp. 720–736. Springer (2022)
42. Wu, J.Z., Ge, Y., Wang, X., Lei, W., Gu, Y., Hsu, W., Shan, Y., Qie, X., Shou, M.Z.: Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation (2023)
43. Wu, J.Z., Li, X., Gao, D., Dong, Z., Bai, J., Singh, A., Xiang, X., Li, Y., Huang, Z., Sun, Y., He, R., Hu, F., Hu, J., Huang, H., Zhu, H., Cheng, X., Tang, J., Shou, M.Z., Keutzer, K., Iandola, F.: Cvpr 2023 text guided video editing competition (2023)
44. Wu, R., Chen, L., Yang, T., Guo, C., Li, C., Zhang, X.: Lamp: Learn a motion pattern for few-shot-based video generation. arXiv preprint arXiv:2310.10769 (2023)
45. Xiao, G., Yin, T., Freeman, W.T., Durand, F., Han, S.: Fastcomposer: Tuning-free multi-subject image generation with localized attention. arXiv preprint arXiv:2305.10431 (2023)
46. Xu, H., Zhang, J., Cai, J., Rezatofighi, H., Tao, D.: Gmflow: Learning optical flow via global matching. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8121–8130 (2022)
47. Yang, S., Zhou, Y., Liu, Z., Loy, C.C.: Rerender a video: Zero-shot text-guided video-to-video translation. arXiv preprint arXiv:2306.07954 (2023)
48. Yuan, H., Zhang, S., Wang, X., Wei, Y., Feng, T., Pan, Y., Zhang, Y., Liu, Z., Albanie, S., Ni, D.: Instructvideo: Instructing video diffusion models with human feedback (2023)
49. Zhang, D.J., Wu, J.Z., Liu, J.W., Zhao, R., Ran, L., Gu, Y., Gao, D., Shou, M.Z.: Show-1: Marrying pixel and latent diffusion models for text-to-video generation. arXiv preprint arXiv:2309.15818 (2023)

50. Zhang, Y., Wei, Y., Jiang, D., Zhang, X., Zuo, W., Tian, Q.: Controlvideo: Training-free controllable text-to-video generation. arXiv preprint arXiv:2305.13077 (2023)
51. Zhang, Y., Xing, Z., Zeng, Y., Fang, Y., Chen, K.: Pia: Your personalized image animator via plug-and-play modules in text-to-image models (2023)
52. Zhang, Z., Li, B., Nie, X., Han, C., Guo, T., Liu, L.: Towards consistent video editing with text-to-image diffusion models. Advances in Neural Information Processing Systems (2024)
53. Zhao, R., Gu, Y., Wu, J.Z., Zhang, D.J., Liu, J., Wu, W., Keppo, J., Shou, M.Z.: Motiondirector: Motion customization of text-to-video diffusion models. arXiv preprint arXiv:2310.08465 (2023)