EasyChair Preprint

№ 14969

Consistency-Preserving Text-Based One-Shot Video Tuning with Segmentation Mask Guidance

Yue Zhuo, Chunzhi Gu and Shigeru Kuriyama

EasyChair preprints are intended for rapid dissemination of research results and are integrated with the rest of EasyChair.

September 21, 2024

Consistency-Preserving Text-Based One-Shot Video Tuning with Segmentation Mask Guidance

1st Yue Zhuo
Department of Computer Science and Engineering
Toyohashi University of Technology
Toyohashi, Japan
zhuo.yue.qf@tut.jp

2nd Chunzhi Gu
Department of Computer Science and Engineering
Toyohashi University of Technology
Toyohashi, Japan
gu@cs.tut.ac.jp

3rd Shigeru Kuriyama
Department of Computer Science and Engineering
Toyohashi University of Technology
Toyohashi, Japan
sk@tut.jp

***Abstract*—Recent text-to-video (T2V) techniques have achieved remarkable success using the text-to-image (T2I) diffusion-based generation diagram. These methods have also been extended to tune video content using the existing T2I model in a one-shot manner. However, these models still struggle with temporal consistency preservation and tend to cause severe jitters, especially for moving objects. To address this issue, in this study, we propose to incorporate segmentation guidance into the diffusion pipeline to promote temporal stability. In particular, we first extract the positions of user-specified objects in each frame using an object segmentation model and generate a sequence of mask images. Then, we utilize the features of the mask image sequence as the query for the cross-attention mechanism in the diffusion model, while the content features of the original video serve as the key and value to generate the edited image sequence. As such, the object position information in the mask guidance can effectively guide the video generation process to reduce jitter. Experiments demonstrate that our method contributes to improved video quality compared to prior video tuning methods in terms of temporal smoothness.**

***Index Terms*—text-to-video generation, one-shot video tuning, segmentation guidance**

I. Introduction

In recent years, diffusion model-based image generation techniques have made significant progress and have gained wide attention. The ability of diffusion models [1] to generate high-resolution, high-quality, and highly diverse realistic images can be primarily attributed to the iterative denoising process, which gradually transforms random noise into realistic images. Such a generative paradigm has been widely developed to address various tasks, including image editing, image inpainting, or image translation, using representative frameworks such as Stable Diffusion [2], DALL-E-2 [3] utilize CLIP [4] for creating images from textual descriptions.

Besides the task of image generation, other efforts have been made to extend diffusion-based text-to-image (T2I) models to tackle the task of text-to-video (T2V) generation. Unlike static image generation, T2V requires generating coherent and smooth image sequences along the temporal dimension. Among these methods, Tune-A-Video (TAV) [5] is a pioneering work for T2V. It leverages pre-trained T2I models and adopts a one-shot fine-tuning approach to circumvent the burden of training with large-scale video datasets. Specifically, TAV extends the original spatial self-attention [6] mechanism to spatio-temporal self-attention to ensure content consistency in generating video sequences. During inference, it employs an implicit denoising inversion process to provide structural guidance to maintain the overall temporal coherence. Despite the promising results achieved by TAV in the T2V, it still suffers from some issues. For example, when occlusions exist in the scene where the foreground and background of the input video cannot be clearly separated and the resulting video generated by TAV can appear jittery, which crucially damages the quality required in real-world applications.

To overcome such limitations of existing T2V methods, this study proposes a novel text-based video tuning method that focuses on promoting visual smoothness. Our core idea is to explicitly guide the diffusion-based T2V model with the exact location of the object in the moving scene. To this end, we first extract the positions of user-specified objects in each frame of the input video with an off-the-shelf object segmentation approach and then generate the corresponding mask image sequence. Next, based on the content features of the original video, which serve as the key and value, we regard such mask sequence as the query in the cross-attention mechanism to eventually generate the target video sequence. By explicitly incorporating the temporal positional information of the object, our method can effectively mitigate jitter in the tuned video, thus improving the overall visual quality regarding smoothness and the preservation of original contents. Moreover, benefiting from the rich information within the mask sequence, besides the selected objects in the front scene, our method jointly contributes to smoother and stabler background generation. Our method can be utilized in a plug-and-play manner for existing video tuning models. Extensive experimental results and ablative evaluations on various types

of public video content demonstrate that our proposed mask-guided diffusion paradigm can improve the performance of existing T2V models.

II. Related Work

Diffusion-based T2V Generation. The field of T2I generation [7]–[9] has seen remarkable progress with the advancement of diffusion models. Furthermore, recent efforts are also made to T2I models to T2V generation [10]–[12] by expanding the spatial diffusion mechanism into the spatial-temporal domain. The Video Diffusion Models (VDM) by Ho et al. [13] introduced a space-time factorized U-Net [14] architecture, leveraging joint training on both image and video data. Based on this model, they also used in [15] the cascaded diffusion models with v-prediction parameterization to produce high-quality videos. Other approaches focused on transferring the progress made in T2I generation to T2V tasks, such as Make-A-Video [16] and MagicVideo [17]. Despite the impressive results, these methods rely heavily on training with large-scale video datasets, which inevitably incurs prohibitive computational expense.

Video Editing. Another direction of diffusion-based approaches [18]–[20] focus on editing the given input video, especially using a slightly modified text prompt compared to the original one. Bar et al. [18] introduced a texture-based video editing using text prompts which enables augmenting the scene with artistic visual effects, yet it can sometimes fail to reflect the intended edits due to its reliance on Layered Neural Atlases [21]. Molad et al. [19] merged the hierarchical feature representation from low- to high-resolution to boost source video fidelity. Gen-1 [20] showed awareness of structure and content during video editing. Recently, a one-shot tuning T2V method [5] has been proposed to perform high-quality editing by leveraging spatial-temporal cross attention to learn the temporal dependency for natural video generation. Although it largely saves computational costs, it can easily induce heavy jittering for the fast-moving scene or occluded objects. Our method builds upon T2V by developing a novel consistency-preserving module, which greatly mitigates the non-smooth inter-frame transition and thus contributes to high visual quality.

Guidance-injected Video Generation. Providing additional guidance as prior knowledge in video generation is also an effective way to improve quality and allow for control over the generation [22]–[26]. The guidance signal can be flexibly designed considering the inherent target of different tasks. Xu et al. [22] proposed to combine human pose and appearance guidance to jointly maintain character identity for high-quality character animation. Hu et al. [24] introduced an end-to-end framework to transform static human images into videos with arbitrary action and viewpoint guidance. The method by Chen et al. [25] enables diverse forms of guidance regarding text, image, and motion for controlled human motion synthesis. Our method falls into this category by introducing mask guidance to improve video quality. Our method also entails a cross-attention module to effectively utilize such guidance and focuses centrally on the promotion of video smoothness.

III. Method

Let $\mathcal{V} = \{I_i | i = 1, \cdots, N\}$ be an input video explained by a text prompt P, which is composed by N frames. Our goal is to generate an edited video $\mathcal{V}^{\mathcal{T}} = \{I_i^{\mathcal{T}} | i = 1, \cdots, N\}$ using a new text prompt $P^{\mathcal{T}}$. Here, $P^{\mathcal{T}}$ is an edited text description of P. Basically, $P^{\mathcal{T}}$ has a similar context to P but with additional or modified detail expressions. Although TAV [5] can directly handle this task, it tends to induce jittery results. To address this issue, we propose to incorporate mask guidance to promote smoothness. In the following, we first explain diffusion-based video tuning, and then detail our method that explicitly guides the generation process with a mask sequence.

A. Preliminaries

Denoising Diffusion Probabilistic Models (DDPMs). Recent image/video generation has been centrally resolved using DDPMs by considering the remarkable capacity. Specifically, DDPMs are generative models based on a gradual denoising process, which learns to iteratively map a Gaussian noise to clean data. During training, it starts from a forward process to convert the real data to a pure Gaussian noise over T timesteps in a Markovian fashion. The transition probability at each step is defined as

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I), \tag{1}$$

where $\beta_t \in (0, 1)$ is a predefined variance schedule. As such, the conditional probability at an arbitrary timestep can be expressed in a closed form $q(x_t|x_0)$:

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)I), \tag{2}$$

where $\bar{\alpha}_t$ is another hyperparameter given by: $\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$. Note that the diffusion-based formulation regards the sample at the T-th forward timestep as a pure Gaussian noise.

Then, beginning with a standard Gaussian distribution $p(x_T) = \mathcal{N}(x_T; 0, I)$, the model learns to generate the original data by performing iterative denoising process (i.e., reverse process):

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)), t = T, \cdots, 1. \tag{3}$$

Here, θ represents the learnable model parameters. The training objective is to minimize the KL divergence between the real data distribution and the generated distribution. In practice, this is equivalent to training a series of denoising mapping $\epsilon_\theta(x_t, t)$ to predict the added noise:

$$\mathcal{L} = \mathbb{E}_{t,x_0,\epsilon}\left[\left\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t)\right\|^2\right], \tag{4}$$

where ϵ denotes standard Gaussian noise. Eq. 4 drives the DDPMs to learn complex data distributions, which allows for the generation of high-quality samples during inference.

Tune-A-Video (TAV). TAV is a recent advancement in text-to-video generation that leverages pre-trained T2I diffusion models for one-shot tuning. TAV requires only a single video

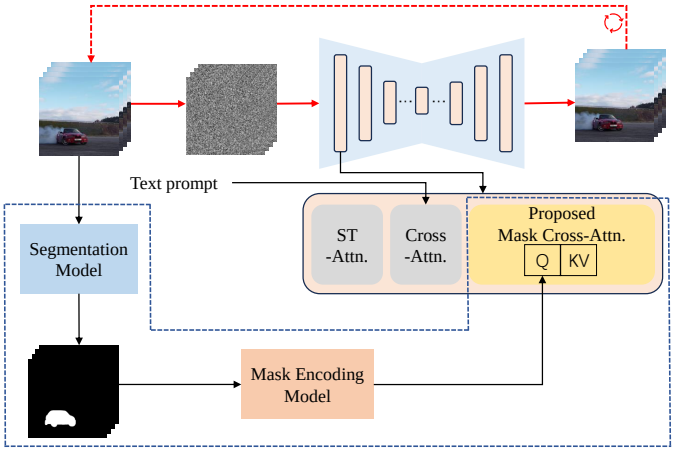

Fig. 1. **Model Overview**. Our proposed smoothness-promoting module for video tuning is illustrated in the dotted area.

to avoid the need for large-scale training video datasets. In particular, TAV tunes the T2I model parameters so that it can learn to produce temporally coherent image sequences. Given the tuned model parameterized by $\theta^{\mathcal{T}}$, it can readily generate the tuned video sample $\mathcal{V}^{\mathcal{T}}$ with modified text $P^{\mathcal{T}}$ following the reverse process.

While TAV has shown promising results in text-guided video editing, we notice that it struggles to maintain temporal consistency, especially for moving objects when the foreground and background are not clearly distinguishable or occlusions occur. We next explain our method, which further improves TAV for high-quality video results.

B. Proposed Method

To encourage smooth and natural edit results, we propose explicitly guiding diffusion-based video generation with the location of the moving object. To this end, we prepare a mask sequence that identifies the object of interest and incorporates the mask feature into the T2V framework, as shown in the dotted area in Fig. 1.

Mask Extraction. Our method starts with foreground extraction to obtain the mask sequence. Given the input video, we let users specify the target object and use the state-of-the-art interactive object segmentation model, called Segment Anything Model (SAM) [27], to perform frame-wise segmentation. This yields a mask sequence $\mathcal{M} = \{M_i | i = 1, ..., N\}$ composed by N binary images, which can be given by

$$M_i = \text{SAM}(I_i, U_i), \quad i = 1, ..., N, \tag{5}$$

where SAM represents the pre-trained SAM model, and $U_i = (x_i, y_i)$ denotes the user-specified coordinates for the i-th frame. In particular, U_i is expected to be directly clicked within the target masking object to encourage accurate segmentation. The mask sequence $\mathcal{M}$ is then encoded into feature space with a mask encoder E^M: $F^M = \text{E}^M(\mathcal{M})$ to obtain the embedding F^M. We next need to discuss how to leverage the mask feature F^M to guide video tuning.

Mask-guided Cross-Attention. In the T2V implementation, the video feature F^V is learned via spatial-temporal attention and prompt-feature attention. Building on top of the structure, we design an additional Mask-guided Cross-Attention (MCA) module to further merge the F^V and F^M. Specifically, we regard the F^M as query (Q) and the F^V as key (K) and value (V) for the MCA learning:

$$F^{\text{MG}} = \text{MCA}(Q = F^M W^Q, K = F^V W^K, V = F^V W^V), \tag{6}$$

which produces the merged feature F_{MG}. W^* refers to the corresponding learnable matrix. The MCA simply follows the dot-product-based attention calculation

$$\text{MCA}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V, \tag{7}$$

where d_k is a scaling factor to balance the influence of high feature dimensionality.

In summary, by integrating the mask information via our MCA module during training, the diffusion-based generation formulation can be guided to pay more attention to the regions specified by the masks that provide additional spatial guidance. This approach enables a more natural movement of the selected objects or areas, addressing the potential instability issues in the original TAV method. As can be seen in the dotted area in Fig. 1, our smoothness-promoting learning can be introduced in a plug-and-play manner to be effectively adapted to the existing diffusion-based generation module.

IV. Experiments

A. Implementation

To evaluate the effectiveness of our proposed mask-guided approach, we conduct a series of experiments to compare our method against the baseline TAV model. Our implementation builds upon the TAV architecture, utilizing a pre-trained Stable Diffusion [2] as the generation backbone. Following TAV, for all experiments, we process the input videos by uniformly sampling 24 frames from each video and resizing all frames to 512×512 pixels. The mask encoder simply adopts the ResNet architecture [28]. The extracted mask sequence is processed similarly for size alignment. Our training configuration included 500 epochs with a learning rate of 3×10^{-5} and a batch size of 1. We eventually collect 10 video clips from free online content and benchmark dataset [29]. To showcase the strength of our method completely, the videos are selected to contain rapidly moving objects or occlusions.

In the following sections, we present extensive experimental results focusing on various aspects of video editing quality and consistency, including overall visual quality, object stability, spatial accuracy, and temporal coherence.

B. Qualitative results

We first present qualitative results against TAV in Figs. 2 $\sim$ 5. It can be confirmed that in all cases, the edited results are consistent with the mask guidance regarding the object location. In particular, we can see in Fig. 2(h) that although the

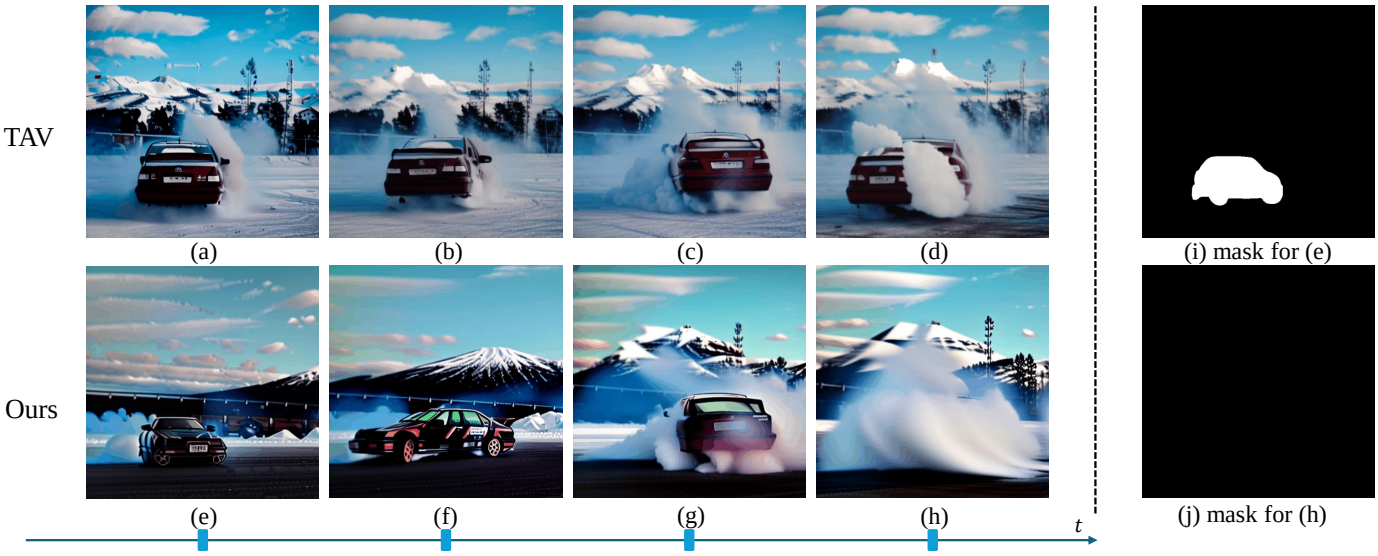

Fig. 2. **Tuned result with the text prompt being *"a car is drifting on the snow"*.**

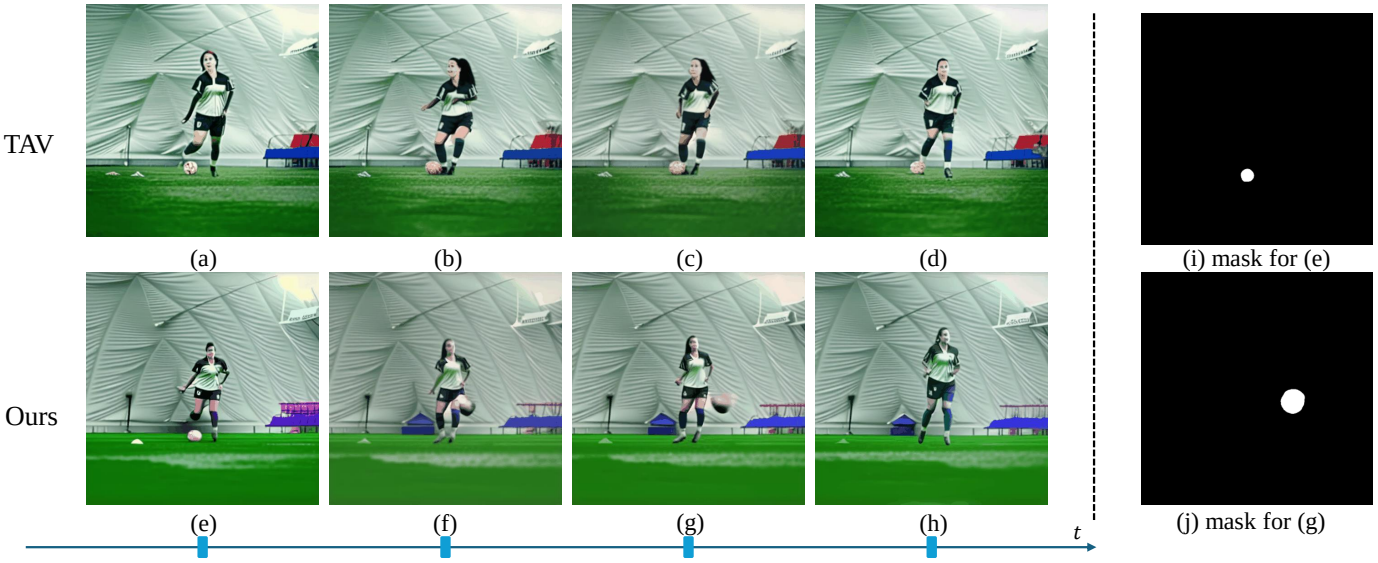

Fig. 3. **Tuned result with the text prompt being *"a woman is playing football"*.**

mask does not contain car (j), the result still seems realistic, in which the snow completely covers the entire car. By contrast, the results by TAV appear highly unnatural in (d) because the car seems to be stretched and partitioned by the snow. Another typical example is shown in Fig. 4, in which the result by TAV Fig. 4(b) is missing the basketball and a part of the astronaut's right leg. Benefiting from the mask guidance in (i), on the contrary, Our method produces stable visual effects even for such moving objects with frequent interaction with humans. *Please refer to the supplementary videos for a clear visual inspection*.

Discussion. While our method produces visually satisfactory results, it can sometimes sacrifice prompt faithfulness to gain high smoothness. An example would be Fig. 5. Although our method produces a stabler trajectory for the ball, the keyword "forest" is less reflective in the result compared to the results by TAV. We assume this trade-off between object focus and background (i.e., prompt) fidelity demonstrates one of the characteristics of our mask-guided cross-attention mechanism, which directs the model's attention to specific regions of interest.

C. Effect of Mask Setting

Since our mask guidance incorporation depends on the selected target object, we here investigate the influence of mask selection on the target object's generation quality. The results are shown in Fig. 6. In particular, we prepare three types of mask selection: Ball (B), Human (H), and Ball & Human (B+H). We can see that the mask selection plays an important role in the generation. For the B case (Fig.

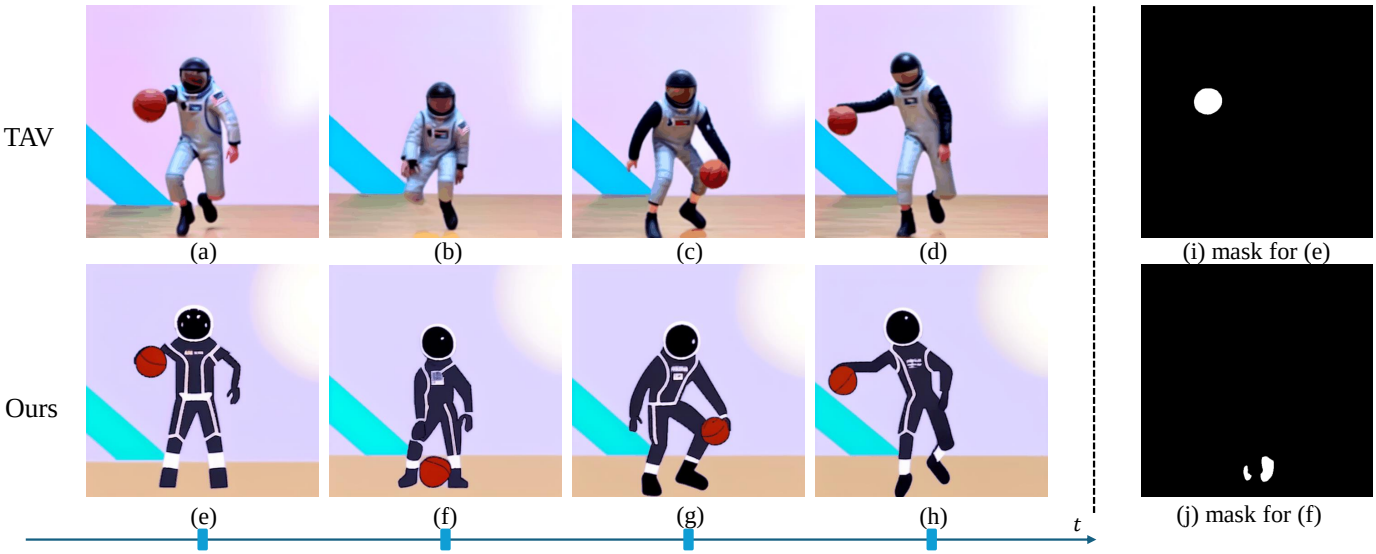

Fig. 4. **Tuned result with the text prompt being *"an astronaut is dribbling basketball, cartoon style"*.**

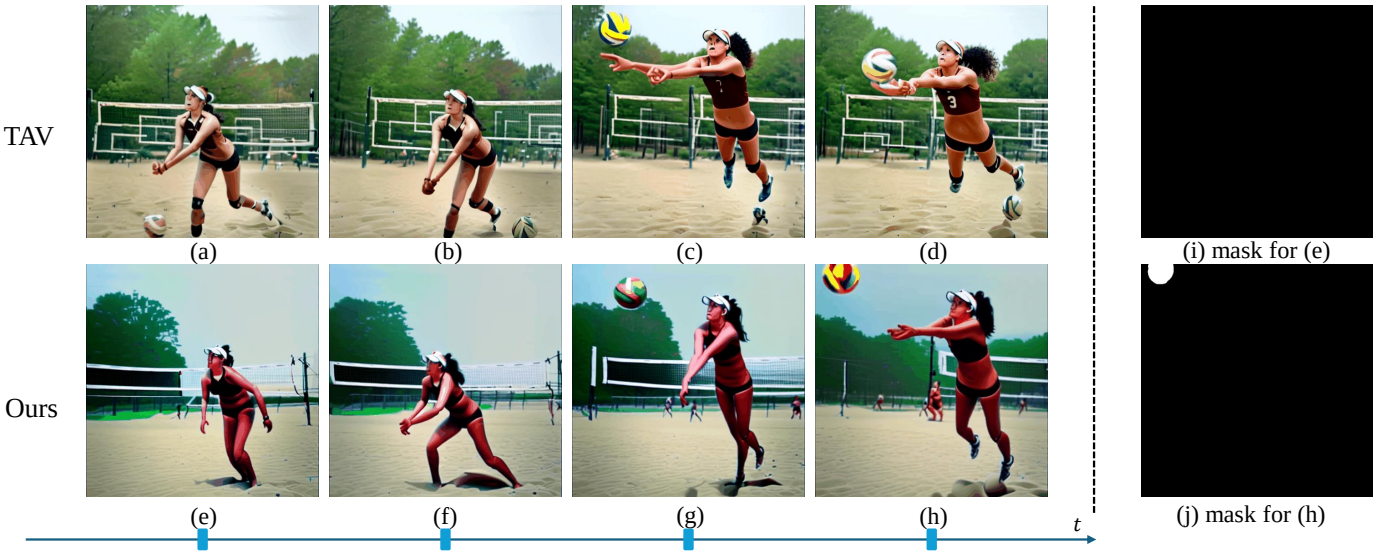

Fig. 5. **Tuned result with the text prompt being *"a woman is playing volleyball in the forest"*.**

6(a)), the model focuses on the ball movement and causes the human arm part to be less realistic, while for the H case (Fig. 6(b)), it seems that it appears another ball in the ground. As for B+H, the model performs the best among these three cases, but it can slightly degrade the background quality. We can thus confirm that a different selection for the mask target provides different regional guidance for the model in improving smoothness, and there exists a trade-off in fore- and background quality. Anyhow, the smoothness for all three cases improves significantly compared to TAV [5].

V. Conclusion

In this paper, we have introduced a novel mask-guided cross-attention mechanism (MCA) for text-guided video editing. Our approach addresses the challenges of temporal consistency and object stability in video editing tasks by incorporating user-specified mask information into the editing process. By leveraging the segmentation model for accurate mask extraction and integrating such positional information through our proposed MCA, the diffusion-based generation process can be explicitly guided to respect the mask guidance to yield smoothness. Experimental results demonstrate that our method outperforms existing video-tuning methods in terms of visual naturalness and temporal consistency.

While our approach is generally effective, it still sometimes sacrifices text fidelity for smoothness gain. We would like to devise more powerful attention-merging schemes to hopefully circumvent this issue in the future.

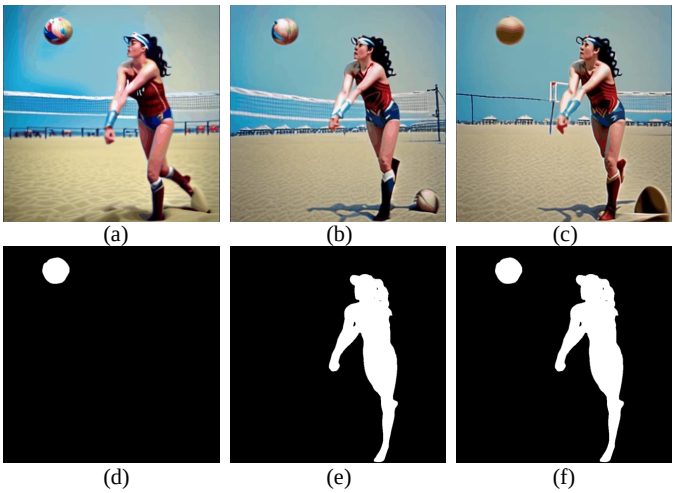

Fig. 6. **Effect of mask object selection.** (a,d), (b,e), and (c,f) show the animated frame and the corresponding mask image, respectively.

ACKNOWLEDGMENT

This work was supported in part by JSPS KAKENHI Grant Numbers JP24K15247.

REFERENCES

[1] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[2] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[3] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, vol. 1, no. 2, p. 3, 2022.

[4] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[5] J. Z. Wu, Y. Ge, X. Wang, S. W. Lei, Y. Gu, Y. Shi, W. Hsu, Y. Shan, X. Qie, and M. Z. Shou, "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.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[7] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," *Advances in neural information processing systems*, vol. 35, pp. 36 479–36 494, 2022.

[8] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen, "Glide: Towards photorealistic image generation and editing with text-guided diffusion models," *arXiv preprint arXiv:2112.10741*, 2021.

[9] S. Gu, D. Chen, J. Bao, F. Wen, B. Zhang, D. Chen, L. Yuan, and B. Guo, "Vector quantized diffusion model for text-to-image synthesis," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 696–10 706.

[10] M. Ding, W. Zheng, W. Hong, and J. Tang, "Cogview2: Faster and better text-to-image generation via hierarchical transformers," *Advances in Neural Information Processing Systems*, vol. 35, pp. 16 890–16 902, 2022.

[11] W. Hong, M. Ding, W. Zheng, X. Liu, and J. Tang, "Cogvideo: Large-scale pretraining for text-to-video generation via transformers," *arXiv preprint arXiv:2205.15868*, 2022.

[12] L. Khachatryan, A. Movsisyan, V. Tadevosyan, R. Henschel, Z. Wang, S. Navasardyan, and H. Shi, "Text2video-zero: Text-to-image diffusion models are zero-shot video generators," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 15 954–15 964.

[13] J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet, "Video diffusion models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 8633–8646, 2022.

[14] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer, 2015, pp. 234–241.

[15] J. Ho, W. Chan, C. Saharia, J. Whang, R. Gao, A. Gritsenko, D. P. Kingma, B. Poole, M. Norouzi, D. J. Fleet *et al.*, "Imagen video: High definition video generation with diffusion models," *arXiv preprint arXiv:2210.02303*, 2022.

[16] U. Singer, A. Polyak, T. Hayes, X. Yin, J. An, S. Zhang, Q. Hu, H. Yang, O. Ashual, O. Gafni *et al.*, "Make-a-video: Text-to-video generation without text-video data," *arXiv preprint arXiv:2209.14792*, 2022.

[17] D. Zhou, W. Wang, H. Yan, W. Lv, Y. Zhu, and J. Feng, "Magicvideo: Efficient video generation with latent diffusion models," *arXiv preprint arXiv:2211.11018*, 2022.

[18] O. Bar-Tal, D. Ofri-Amar, R. Fridman, Y. Kasten, and T. Dekel, "Text2live: Text-driven layered image and video editing," in *European conference on computer vision*. Springer, 2022, pp. 707–723.

[19] E. Molad, E. Horwitz, D. Valevski, A. R. Acha, Y. Matias, Y. Pritch, Y. Leviathan, and Y. Hoshen, "Dreamix: Video diffusion models are general video editors," *arXiv preprint arXiv:2302.01329*, 2023.

[20] P. Esser, J. Chiu, P. Atighehchian, J. Granskog, and A. Germanidis, "Structure and content-guided video synthesis with diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 7346–7356.

[21] Y. Kasten, D. Ofri, O. Wang, and T. Dekel, "Layered neural atlases for consistent video editing," *ACM Transactions on Graphics (TOG)*, vol. 40, no. 6, pp. 1–12, 2021.

[22] Z. Xu, J. Zhang, J. H. Liew, H. Yan, J.-W. Liu, C. Zhang, J. Feng, and M. Z. Shou, "Magicanimate: Temporally consistent human image animation using diffusion model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 1481–1490.

[23] C. Gu, C. Zhang, and S. Kuriyama, "Orientation-aware leg movement learning for action-driven human motion prediction," *Pattern Recognition*, vol. 150, p. 110317, 2024.

[24] L. Hu, "Animate anyone: Consistent and controllable image-to-video synthesis for character animation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 8153–8163.

[25] W. Chen, Y. Ji, J. Wu, H. Wu, P. Xie, J. Li, X. Xia, X. Xiao, and L. Lin, "Control-a-video: Controllable text-to-video generation with diffusion models," *arXiv preprint arXiv:2305.13840*, 2023.

[26] C. Gu, J. Yu, and C. Zhang, "Learning disentangled representations for controllable human motion prediction," *Pattern Recognition*, vol. 146, p. 109998, 2024.

[27] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4015–4026.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[29] J. Pont-Tuset, F. Perazzi, S. Caelles, P. Arbeláez, A. Sorkine-Hornung, and L. Van Gool, "The 2017 davis challenge on video object segmentation," *arXiv preprint arXiv:1704.00675*, 2017.