

Blended Latent Diffusion under Attention Control for Real-World Video Editing

Deyin Liu

School of Artificial Intelligence, Anhui University

Hefei, China

Department of Computer Science, Swansea University

Swansea, United Kingdom

iedyzzu@outlook.com

Lin Yuanbo Wu

Department of Computer Science

Swansea University

Swansea, United Kingdom

l.y.wu@swansea.ac.uk

Xianghua Xie*

Department of Computer Science

Swansea University

Swansea, United Kingdom

x.xie@swansea.ac.uk

*Corresponding author

Abstract—Due to lack of fully publicly available text-to-video models, current video editing methods tend to build on pre-trained text-to-image generation models, however, they still face grand challenges in dealing with the local editing of video with temporal information. First, although existing methods attempt to focus on local area editing by a pre-defined mask, the preservation of the outside-area background is non-ideal due to the spatially entire generation of each frame. In addition, specially providing a mask by user is an additional costly undertaking, so an autonomous masking strategy integrated into the editing process is desirable. Last but not least, image-level pretrained model hasn't learned temporal information across frames of a video which is vital for expressing the motion and dynamics. In this paper, we propose to adapt a image-level blended latent diffusion model to perform local video editing tasks. Specifically, we leverage DDIM inversion to acquire the latents as background latents instead of the randomly noised ones to better preserve the background information of the input video. We further introduce an autonomous mask manufacture mechanism derived from cross-attention maps in diffusion steps. Finally, we enhance the temporal consistency across video frames by transforming the self-attention blocks of U-Net into temporal-spatial blocks. Through extensive experiments, our proposed approach demonstrates effectiveness in different real-world video editing tasks.

Index Terms—Local Video Editing, Blended Latent Diffusion, DDIM inversion.

I. INTRODUCTION

Diffusion-based generation and editing models represent a cutting-edge research area within visual content editing. Current approaches in diffusion-based editing primarily leverage large-scale pretrained text-to-image generative models, such as Stable Diffusion [1], Imagen [2], DALL-E 2 [3]. While current methods excel in global image manipulation, there is a noticeable gap in addressing local editing needs. Local editing, which involves modifying specific regions or attributes within an image while preserving the rest, is essential for numerous practical applications, such as attribute editing and specified object manipulation.

To enable local editing, several methods are developed: DALL-E 2 [3], GLIDE [4], Blended Diffusion [5], Blended

Latent Diffusion [6], FateZero [7], and Video-P2P [8]. Among these methods, Blended Diffusion [5] stands out as a fully publicly available solution. Building upon this foundation, Blended Latent Diffusion seamlessly integrates it into the latent space of the Latent Diffusion Model [1]. Subsequent advancements, including FateZero [7] and Video-P2P [8], further extend and refine the capabilities of these techniques. The basic idea of Blended Latent Diffusion [6] is to spatially blend the foreground latent (i.e., each of the noisy latents progressively generated in the latent denoising steps conditioned directly on the guiding text prompt) with the background latent (i.e., corresponding noised version of the original latent of the input image), by using a user-provided mask to yield the latent for the next diffusion step. However, it is problematic to employ Blended Latent Diffusion [6] for local video editing due to the following reasons: **1)** For the background latent, Blended Latent Diffusion [6] leverages the noised version of the original latent at the same noise level as the foreground latent. However, the added noise is stochastic, leading to inaccurate blended latent and thus affecting the local editing outcome. **2)** The blend diffusion [5], [6] requires the user to provide a mask to specify the area to edit. This requires a user interaction or an automatic detection/segmentation method, which is of additional workload, not desirable in practice. **3)** When one extends the original Blend Latent Diffusion method [6] for video editing, it is imperative to keep the temporal consistency of the video frames, which is not learned by the pretrained text-to-image diffusion model.

In this paper we propose to manipulate local video editing based on Blended Latent Diffusion. In specific, we first improve the preservation for the background (outside-mask area) by choosing the deterministic DDIM [9] inverted latents of each frame which can be used for reconstructing the input. Secondly, to avoid the need for user to provide masks, we consider an autonomous mask manufacture mechanism, which leverages the cross-attention maps [10] from the U-Net [11] that provides the semantic layout of the image. Such cross-attention maps can be used to produce a mask by thresholding to locate the area corresponds to the edited words. Finally, we transform the self-attention blocks in U-Net [11] into

temporal-spatial attention blocks to capture the inter-frame temporal information and enhance the temporal consistency of video appearance.

Our proposed video editing method termed Blend Latent Diffusion under Attention Control consists of the above three modules: DDIM inversion based latents for background, cross-attention control for automatic masking and temporal-spatial attention for temporal consistency enhancement. We conduct extensive experiments to verify the superior performance of our method by comparing with the state-of-the-art video editing methods and ablation studies.

II. RELATED WORK

A. Text-to-Image Synthesis

Text-to-image synthesis has garnered increasing interest in recent years, which generates images that match the given textual description in terms of both semantic consistency and image realism [12]. Seminal works based on RNNs [13] and GANs [14]–[16], were later superseded by transformer-based approaches [17]. For example, DALL-E [18] proposed a two-stage approach where in they first trained a discrete VAE [19] to learn a rich semantic context, and then they trained a transformer model to autoregressively model the joint distribution over the text and image tokens.

Diffusion models were also used for various global image-editing applications. ILVR [20] demonstrates how to condition a DDPM model [21] on an input image for image translation tasks. Palette [22] trains a designated diffusion model to perform four image-to-image translation tasks, namely colorization, inpainting, uncropping, and JPEG restoration. SDEdit [23] demonstrates stroke painting to image, image compositing, and stroke-based editing. RePaint [24] uses a diffusion model for free-form inpainting of images.

B. Text Driven Video Editing

It becomes more challenging to edit the object shape in real-world videos. Current methods exhibit artifacts even with the optimization on generative priors [25]. The stronger prior of the diffusion-based model also draws the attention from researchers, e.g., gen1 [9] trains a conditional model for depth and text-guided video generation, which can edit the appearance of the generated images on the fly. Dreamix [36] finetunes a stronger diffusion-based video model [18] for editing with stronger generative priors. Nonetheless, both of these methods need private and powerful video diffusion models for editing.

III. THE METHOD

A. Preliminary

Latent Diffusion Models [5] (e.g., Stable Diffusion) extend the diffusion models [21] into the latent space of an auto-encoder. Firstly, an encoder E compresses an image x to a lower dimensional latent $z = E(x)$, which can be reconstructed back to image $D(z) \approx x$ via a decoder D . Thereafter, a U-Net [11] ϵ_θ consisting of cross-attention and self-attention blocks

is trained to predict the artificial noise using the following objective:

$$\min_{\theta} E_{z_0, \epsilon \sim N(0, I), t \sim \text{Uniform}(1, T)} \|\epsilon - \epsilon_\theta(z_t, t, p)\|_2^2, \quad (1)$$

where p is the embedding of the conditional text prompt and z_t is a noisy sample of z_0 at timestep t .

Denoising Diffusion Implicit Models (DDIM) [9] is a deterministic sampling technique, which is employed during inference to convert a random noise z_T to a clean latent z_0 in a sequence of timestep $t : T \rightarrow 1$:

$$z_{t-1} = \sqrt{\alpha_{t-1}} \frac{z_t - \sqrt{1 - \alpha_t} \epsilon_\theta}{\sqrt{\alpha_t}} + \sqrt{1 - \alpha_{t-1}} \epsilon_\theta, \quad (2)$$

where α_t is a parameter for noise scheduling and each ϵ_θ stands for the noise $\epsilon_\theta(z_t, t, p)$ predicted at timestep t .

Blended Latent Diffusion [6] follows the idea of Blended Diffusion [5] and repeatedly blends two parts, i.e., foreground and background, into the latent space as the diffusion progresses. The foreground (fg) refers to the part that one wishes to modify, which is to be restricted by a given mask m , while the background (bg) refers to the rest. Specifically, in the latent space, due to the convolutional nature of the autoencoder, the width and the height are smaller than those of the input image (by a scalability factor of 8). The input mask m is therefore downsampled to such spatial dimensions to obtain the latent space binary mask m_{latent} , which will be used to perform the blending. Then, the denoising diffusion process is manipulated in the following way: at each step, a latent denoising step is first performed, conditioned directly on the embedding of the guiding text prompt p_{edit} , to obtain a less noisy foreground latent denoted as z_{fg} . Meanwhile, the original latent z_0 is noised and injected into the current noise level in a step (via $\sqrt{\alpha_t} z_0 + \sqrt{1 - \alpha_t} \epsilon$) to obtain a noisy background latent z_{bg} . The two latents z_{bg} and z_{fg} are then blended using the resized mask to yield the latent for the next latent diffusion step via:

$$z_{t-1} = z_{fg} \odot m_{latent} + z_{bg} \odot (1 - m_{latent}), \quad (3)$$

where \odot is element-wise multiplication. At each denoising step the latent is modified corresponding to the edit prompt, whilst the subsequent blending enforces the part outside m_{latent} to remain the same. Though the resulting blended latent is not guaranteed to be coherent between foreground and background, the next latent denoising step can address it. Once the latent diffusion process terminates, the final latent is decoded to the output image using the decoder $D(z)$.

B. Blended Latent Diffusion under Attention Control

DDIM Inversion for Background Latents: As Blended Latent Diffusion [6] suggested, the background latent z_{bg} is chosen as the noised version of the initial latent of the image at the same noise level as the foreground latent, although the foreground and the background latents can be viewed as riding on the same manifold, the added noise for the background latent is random, not deterministic, thus the original intention of preserving the outside-mask area in the image is not well achieved. In contrast, DDIM inverted latents, which can be

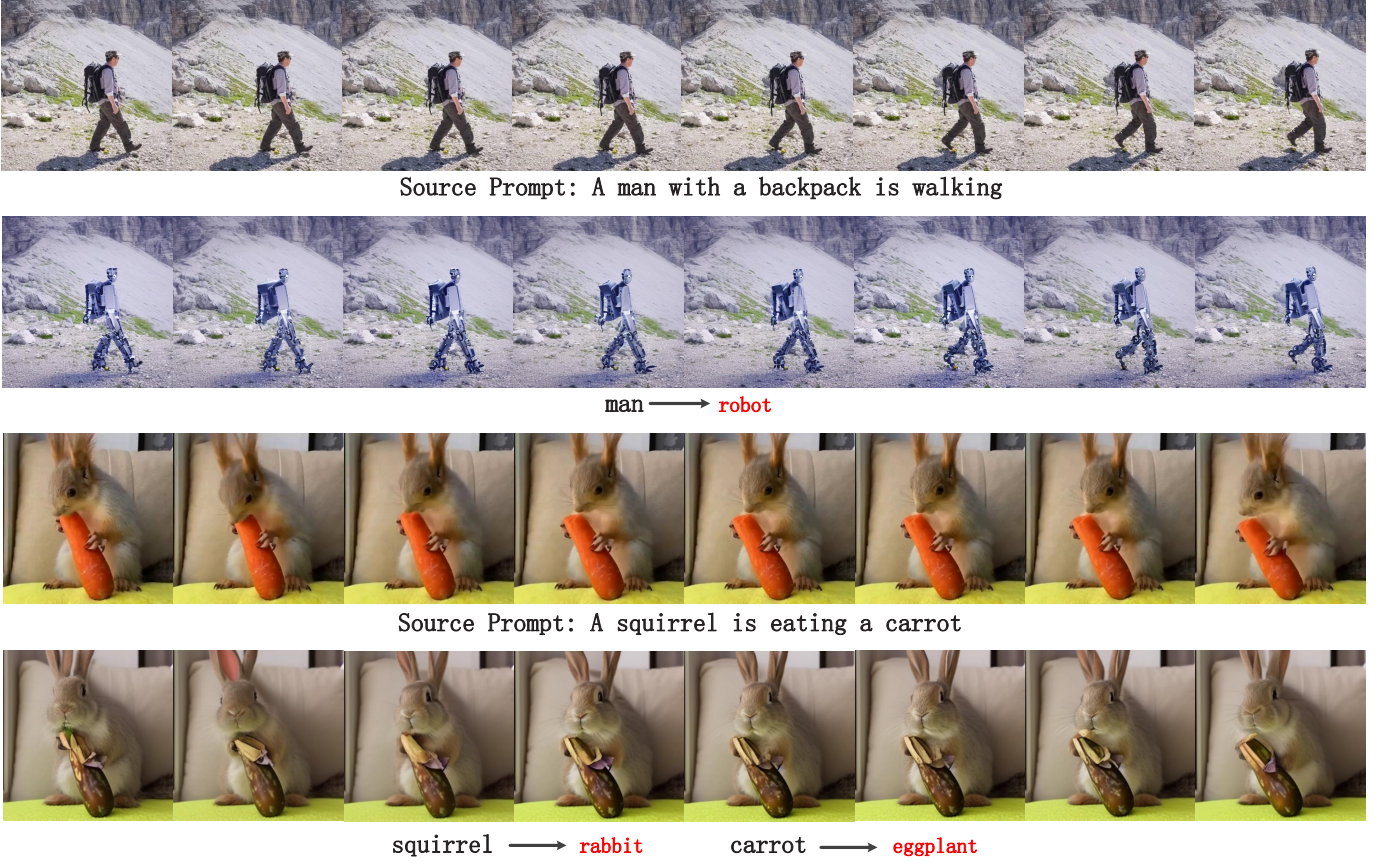


Fig. 1. Examples of local video editing achieved by our proposed method

used to progressively reconstruct the original latent, will be a better alternative for the background. According to the ODE limit analysis of the diffusion process, DDIM inversion [9], [26] is able to map the initial latent z_0 to a sequence of noised latents z_t in the steps $t : 1 \rightarrow T$, which is reverse to DDIM sampling in Equation (2):

$$\hat{z}_t = \sqrt{\alpha_t} \frac{\hat{z}_{t-1} - \sqrt{1 - \alpha_{t-1}} \varepsilon_\theta}{\sqrt{\alpha_{t-1}}} + \sqrt{1 - \alpha_t} \varepsilon_\theta. \quad (4)$$

In such a way, the acquired latents z_t , $t : 1 \rightarrow T$, can be used to accurately reconstruct the initial latent. Thus, we can choose them as the background z_{bg} to replace the randomly noised latents to preserve the background area.

Cross Attention Thresholding Mask: Unlike Blended Latent Diffusion [6] that needs a user-provided mask to localize the edit to the specified area, we take advantage of the cross-attention map which provides the semantic layout of the image closely related to the prompt text [10]. Hence, we can obtain a binary mask m_t by thresholding the cross-attention maps of the edited words during the diffusion process by a constant τ . Specifically, we first compute the average attention map $\bar{M}_{t,w}$ (averaged over steps $T, T-1, \dots, t$) of the source word w during the DDIM inversion conditioned on the embedding of the source text prompt p_{src} , and then we calculate the average attention map \bar{M}_{t,w^*}^* of the new word w^* during the diffusion

process conditioned on the embedding of the target text prompt p_{edit} . A threshold is set to produce the binary maps, where $B(x) := x > \tau$ and $\tau = 0.3$ throughout all the experiments. To achieve seamless local editing, the edited region should include the silhouettes of both the source and the newly edited area. To this end, the final mask m_{latent} is a union of the two binary maps:

$$m_{latent} = B(\bar{M}_{t,w}) \cup B(\bar{M}_{t,w^*}^*). \quad (5)$$

Using such an autonomous mask manufacture mechanism can avoid the cumbersome provision of a mask from a pre-defined method such as image segmentation.

Temporal-Spatial Attention: The previous two designs construct a strong local editing method that can preserve the structure of source image with the background almost retained. When applied to video editing, however, denoising each frame individually readily produces temporally inconsistent video since the pretrained text-to-image model used cannot learn the temporal information regarding inter-frame consistency. Inspired by the recent one-shot video generation method [27], we reshape the original self-attention to temporal-spatial attention with the pretrained weights unchanged. Specifically, for a video consisting of n frames, we implement the ATTENTION(Q, K, V) [28] for the original self attention

input feature z_i at frame index $i \in [1, n]$ as:

$$Q = W^Q z^i, K = W^K [z^{i-1}; z^i], V = W^V [z^{i-1}; z^i], \quad (6)$$

where $[\cdot]$ denotes the concatenation operation. W^Q , W^K , W^V are the projection matrices from the pretrained model. Then, the temporal-spatial attention map is represented as $s_t \in R^{hw \times fwh}$, where $f=2$ is the number of frames used as key and value. It captures both the structure of a single frame and the temporal correspondence with the nearest neighbor frame features. In such a training-free way, we build up the relationships of each frame with its nearest neighbor, maintaining the temporal continuity and inter-frame consistency.

IV. EXPERIMENTS

A. Implementation Details

We choose the pretrained CompVis stable diffusion v1.4 [1] as the base model, and the DDIM [9] sampling and inversion are generally with total timestep $T = 50$. The threshold for the binary mask is set to $\tau = 0.3$. For evaluation, we use the videos from DAVIS [29], Edinburgh office monitoring video dataset [30] and other in-the-wild videos, and generally we sample 8 frames from each video for use. The source prompts of the videos are generated via the image caption model [31] while the target prompt for each video is designed mainly by replacing several words.

B. Applications

Based on the pretrained text-to-image diffusion model [1], our proposed method can be used for local attribute editing, local object property editing, or local object category replacement in a video as shown in Fig.1, Fig.2 and Fig.3. In the first and second rows of Fig.1, the proposed method changes the whole appearance of an original object, the hiker, to a robotic one with the same dynamics of walking, by just replacing the word "man" in the source prompt into "robot" in the target prompt. Similarly, in Fig.2, the attribute of the floor of the playground is changed from "cement floor" into "ice surface" corresponding to the change in the prompts. In Fig.3, the "swan" is changed into "duck" with local modifications especially the shape of the beak. And in the last two rows of Fig.1, the video is edited from "A squirrel is eating a carrot" to "A rabbit is eating an eggplant", completely replacing two local objects' categories into another two. When editing the categories of objects, it is challenging because of the thorough change of shapes and appearances, with the action, pose and position similar to the input video, keeping the motion or temporal dynamics unchanged.

C. Comparisons with State-of-the-Arts

We compare our method with the following state-of-the-art methods: (1) Frame-wise Null-text optimization [32] combined with prompt-to-prompt [10]; (2) Frame-wise editing method SDEdit [23]; (3) Tune-A-Video [27]; (4) The Neural Layered Atlas (NLA) [33] based method combined with key frame

TABLE I
QUANTITATIVE EVALUATION FOR DIFFERENT METHODS

Inversion and Editing Methods	CLIP Metrics	
	Frame-wise Accuracy	Temporal Consistency
Frame-wise Null-text inversion and Prompt-to-prompt [10], [32]	0.96	0.85
Frame-wise SDEdit [23]	0.82	0.91
NLA, Null-text inversion and Prompt-to-prompt [10], [32], [33]	0.60	0.95
DDIM inversion and Tune-A-Video [9], [27]	0.75	0.96
FateZero [7]	0.90	0.97
The proposed	0.91	0.95

editing [10], [32]; (5) Fusing Attentions for Zero-shot Text-based Video Editing (FateZero) [7]. For attention-based editing methods, the timestep parameters are set to identical to us.

In our experiments, quantitative evaluation are also conducted utilizing the trained CLIP [34] model, following existing methods [27], [35], [36]. The results are shown in Table 1. The "Frame-wise Accuracy" [37] is the frame-wise editing accuracy and it represents the percentage of frames for which the edited image has a higher CLIP score to the target prompt than to the source prompt. The "Temporal Consistency" [35] measures the temporal consistency among frames, which is an average of cosine similarities of all pairs of consecutive frames. Table 1 shows that the proposed method achieves comparable frame-wise editing accuracy and temporal consistency as the state-of-the-art methods in local video editing tasks. Although Frame-wise null inversion [32] achieves the best Frame-wise Accuracy, it costs 500 iterations of optimization for each frame, with low temporal consistency. It's known that NLA-based method [33] needs to take hours to optimize the neural atlas for each input video. Tune-A-Video [27] combined with DDIM inversion and FateZero [7] show impressive Temporal Consistency, which benefits from the finetune for spatial-temporal self attention and the temporal attention modules, especially when applied in shape-aware editing. In contrast, our method just reshapes the self attention into temporal-spatial attention with all the weights unchanged, achieving a comparably superior temporal consistency in a training-free way.

D. Ablation Studies

In this section, we ablate our method at different components in the video editing. Firstly, we study the effect of DDIM inversion for background latents. In the blended latent diffusion, for the choice of the background latents, we first use the previous setting, the randomly noised version of the original latent at the same noise level of the each foreground latent, and then the performance of one edited video example is shown in the third row of Fig.2. The target is to edit a local attribute of the original video, changing the cement floor to an ice surface. Although the target is achieved basically in the

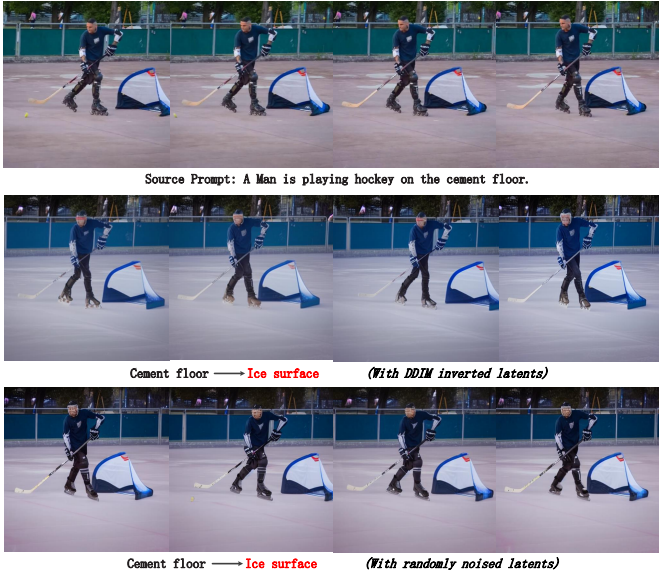


Fig. 2. Performance comparisons between using the DDIM inverted latents and previous randomly noised ones

TABLE II
QUANTITATIVE COMPARISONS REGARDING THE EFFECT OF MASK

Mask Schemes	CLIP Metrics	
	Frame-wise Accuracy	Temporal Consistency
Without Mask	0.55	0.81
User-provided Mask (with the dataset)	0.92	0.95
Cross Attention Thresholding Mask	0.91	0.95

third row, as you can see by zooming in, some details in the background area (here it refers to other part other than the floor in the video) are changed too, for example, the face of the player has been "worn" a face shield, and his shoes are changed from inline skates to blade ice skates. In contrast, as shown in the second row, we use DDIM inverted latents instead of the randomly noised to overcome these problems, keeping the background area unchanged even in such details. This indicates that the latents resulted from DDIM inversion are better choice to preserve the background information.

Secondly, we evaluate the effect of the mask in the local video editing. The quantitative comparisons among different cases are shown in Table 2. Obviously, the one without any mask has the worst local editing performance since there is no any means to limit the target edit into a local area. We can also find that our proposed autonomous mask manufacture scheme, Cross Attention Thresholding Mask presents almost the same local editing performance as the User-provided Mask, while the autonomous scheme will save a lot of extra workload in providing such masks.

Thirdly, we analyze the role of the Temporal-spatial attention module. As shown in Fig.3, the performance of the second row is with Temporal-spatial attention, while the third row

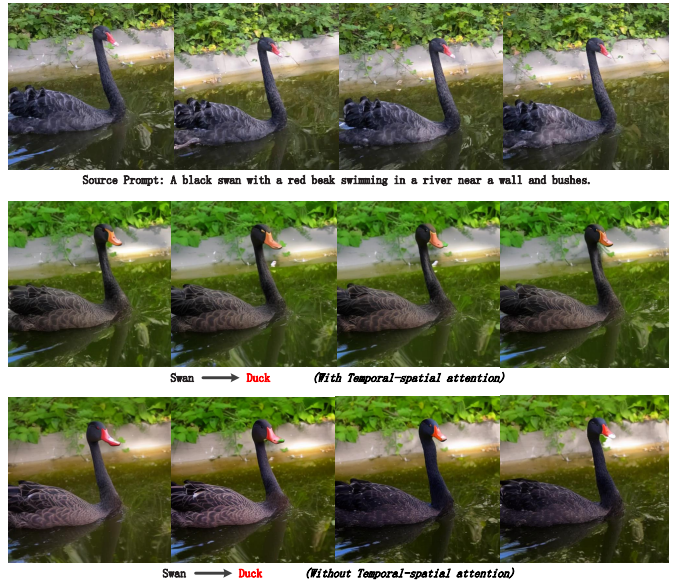


Fig. 3. Performance comparisons with/without the proposed temporal-spatial attention

is without it. Through careful observation, it is not hard to find that, in the third row, the shape and color layout (red part and white point) of the edited duck's beak are very inconsistent among the four frames, and the color tones of the duck are also different (the former two frames are light black, while the last two are heavy black). In contrast, the second row shows all consistent in these details, illustrating the significance of the Temporal-spatial attention in keeping the inter-frame consistency.

V. CONCLUSION

In this paper, we propose a novel text-driven video editing framework that performs local video object editing subject to user-prompts. We utilise the DDIM inverted latents to serve as background and blends with the foreground latents during denoising steps. To accurately localise the local area, we develop an automatic masking mechanism leveraging the cross-attention maps corresponding to the edited words. We further transform the self-attention block in U-Net architecture into temporal-spatial attention, which enhances the temporal consistency in the video. We also demonstrate the proposed method's impressive effectiveness in various local video editing tasks on attribute change, object replacement and so on.

ACKNOWLEDGMENT

This project was funded by Airbus Endeavr Wales, the Foreign Expert Program of the Ministry of Science and Technology (G2023019011L), and the National Natural Science Foundation of China (62372150).

REFERENCES

- [1] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022*,

- New Orleans, LA, USA, June 18-24, 2022. IEEE, 2022, pp. 10 674–10 685.
- [2] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, S. K. S. Ghasemipour, R. G. Lopes, B. K. Ayan, T. Salimans, J. Ho, D. J. Fleet, and M. Norouzi, “Photorealistic text-to-image diffusion models with deep language understanding,” in *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., 2022.
 - [3] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, “Hierarchical text-conditional image generation with CLIP latents,” *CoRR*, vol. abs/2204.06125, 2022.
 - [4] A. Q. 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,” in *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, ser. Proceedings of Machine Learning Research, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvári, G. Niu, and S. Sabato, Eds., vol. 162. PMLR, 2022, pp. 16 784–16 804.
 - [5] O. Avrahami, D. Lischinski, and O. Fried, “Blended diffusion for text-driven editing of natural images,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 2022, pp. 18 187–18 197.
 - [6] O. Avrahami, O. Fried, and D. Lischinski, “Blended latent diffusion,” *ACM Trans. Graph.*, vol. 42, no. 4, pp. 149:1–149:11, 2023.
 - [7] C. Qi, X. Cun, Y. Zhang, C. Lei, X. Wang, Y. Shan, and Q. Chen, “Fatezero: Fusing attentions for zero-shot text-based video editing,” in *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*. IEEE, 2023, pp. 15 886–15 896.
 - [8] S. Liu, Y. Zhang, W. Li, Z. Lin, and J. Jia, “Video-p2p: Video editing with cross-attention control,” *CoRR*, vol. abs/2303.04761, 2023.
 - [9] J. Song, C. Meng, and S. Ermon, “Denoising diffusion implicit models,” in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.
 - [10] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or, “Prompt-to-prompt image editing with cross-attention control,” in *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.
 - [11] 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*, ser. Lecture Notes in Computer Science, N. Navab, J. Hornegger, W. M. W. III, and A. F. Frangi, Eds., vol. 9351. Springer, 2015, pp. 234–241.
 - [12] D. Liu, L. Wu, F. Zheng, L. Liu, and M. Wang, “Verbal-person nets: Pose-guided multi-granularity language-to-person generation,” *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–13, 2022.
 - [13] E. Mansimov, J. B. Emilio Parisotto, and R. Salakhutdinov, “Generating images from captions with attention,” in *CoRR abs/1511.02793*, 2016.
 - [14] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He, “AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks,” in *CVPR*, 2018, pp. 1316–1324.
 - [15] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. Metaxas, “Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks,” in *ICCV*, 2017, pp. 5907–5915.
 - [16] —, “Stackgan++: Realistic image synthesis with stacked generative adversarial networks,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 8, pp. 1947–1962, 2019.
 - [17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in *NIPS*, 2017, pp. 5998–6008.
 - [18] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, “Zero-shot text-to-image generation,” in *ICML*, 2021, pp. 139:8821–8831.
 - [19] A. Razavi, A. V. den Oord, and O. Vinyals, “Generating diverse high-fidelity images with vq-vae-2,” in *Advances in neural information processing systems*, 2019.
 - [20] J. Choi, S. Kim, Y. Jeong, Y. Gwon, and S. Yoon, “Ilvr: Conditioning method for denoising diffusion probabilistic models,” in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, p. 14347–14356.
 - [21] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., 2020.
 - [22] C. Saharia, W. Chan, H. Chang, C. Lee, J. Ho, T. Salimans, D. Fleet, and M. Norouzi, “Palette: Image-to-image diffusion,” in *ACM SIGGRAPH*, 2022, pp. 2909–2918.
 - [23] C. Meng, Y. He, Y. Song, J. Song, J. Wu, J.-Y. Zhu, and S. Ermon, “Sdedit: Guided image synthesis and editing with stochastic differential equations,” in *International Conference on Learning Representations*, 2021.
 - [24] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, and L. V. Gool, “Repaint: Inpainting using denoising diffusion probabilistic models,” in *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 461–11 471.
 - [25] Y.-C. Lee, J.-Z. G. J. Jang, Y.-T. Chen, E. Qiu, and J.-B. Huang, “Shape-aware text-driven layered video editing,” in *CVPR*, 2023, pp. 14 317–14 326.
 - [26] P. Dhariwal and A. Q. Nichol, “Diffusion models beat gans on image synthesis,” in *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, and J. W. Vaughan, Eds., 2021, pp. 8780–8794.
 - [27] 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 *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*. IEEE, 2023, pp. 7589–7599.
 - [28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, Eds., 2017, pp. 5998–6008.
 - [29] J. Pont-Tuset, F. Perazzi, S. Caelles, P. Arbelaez, A. Sorkine-Hornung, and L. V. Gool, “The 2017 davis challenge on video object segmentation,” *CoRR*, vol. abs/1704.00675, 2017.
 - [30] [Online]. Available: <https://homepages.inf.ed.ac.uk/rbf/OFFICEDATA/>
 - [31] J. Li, D. Li, C. Xiong, and S. C. H. Hoi, “BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation,” in *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, ser. Proceedings of Machine Learning Research, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvári, G. Niu, and S. Sabato, Eds., vol. 162. PMLR, 2022, pp. 12 888–12 900.
 - [32] R. Mokady, A. Hertz, K. Aberman, Y. Pritch, and D. Cohen-Or, “Null-text inversion for editing real images using guided diffusion models,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. IEEE, 2023, pp. 6038–6047.
 - [33] Y. Kasten, D. Ofri, O. Wang, and T. Dekel, “Layered neural atlases for consistent video editing,” *ACM Trans. Graph.*, vol. 40, no. 6, pp. 210:1–210:12, 2021.
 - [34] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, “Learning transferable visual models from natural language supervision,” in *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 2021, pp. 8748–8763.
 - [35] P. Esser, J. Chiu, P. Atighehchian, J. Granskog, and A. Germanidis, “Structure and content-guided video synthesis with diffusion models,” in *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*. IEEE, 2023, pp. 7312–7322.
 - [36] G. Parmar, K. K. Singh, R. Zhang, Y. Li, J. Lu, and J. Zhu, “Zero-shot image-to-image translation,” in *ACM SIGGRAPH 2023 Conference Proceedings, SIGGRAPH 2023, Los Angeles, CA, USA, August 6-10, 2023*, E. Brunvand, A. Sheffer, and M. Wimmer, Eds. ACM, 2023, pp. 11:1–11:11.
 - [37] J. Hessel, A. Holtzman, M. Forbes, R. L. Bras, and Y. Choi, “Clipscore: A reference-free evaluation metric for image captioning,” in *Proceedings*

of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021, M. Moens, X. Huang, L. Specia, and S. W. Yih, Eds. Association for Computational Linguistics, 2021, pp. 7514–7528.