DRIVE: Diffusion Refinement and Instance-level Video Editing

Pin-Han Huang
b10902065@csie.ntu.edu.tw

Yi-Ru Wu
b10902135@csie.ntu.edu.tw

Te-Wei Chen
b10902138@csie.ntu.edu.tw

Ren-Wei Liang
b10902050@csie.ntu.edu.tw

Cheng-Yu Lin
b10902024@csie.ntu.edu.tw

***Abstract*—Recent advancements in video editing have greatly enhanced quality by incorporating pre-trained diffusion models into pipelines that maintain temporal consistency across frames. In this work, we build upon previous research to develop a training-free video editing pipeline featuring instance-level control by leveraging pre-processed object masks. Our pipeline, DRIVE (Diffusion Refinement and Instance-level Video Editing), advances the current state-of-the-art training-free video editing framework, RAVE [15]. We introduce several key modifications to the original framework, allowing a precise instance-level video editing framework that remains training-free. A qualitative demo video showcasing our pipeline is available at: https://www.youtube.com/watch?v=-LebLeMJJ-0.**

I. Introduction

Diffusion-based generative models [12], [26] have achieved great success in generating and editing high-quality images guided by text prompts. These methods enable tasks such as object editing [10], image inpainting [2], personalized generation [23], and image-to-image translation [20]. DDIM inversion [27] is employed to perform precise editing by image-to-noise inversion and regenerating edited images based on target text prompts. Recent research has extended these methods to video editing, though significant challenges remained. Some research has focused on text-to-video (T2V) generation [11], [13], [25]using large datasets, but these methods are often costly and unsuitable for general use. Other methods rely on conventional video editing techniques, such as keyframe selection [14] or atlas editing [3], [16], which are time-consuming. More recent works have explored using pre-trained text-to-image (T2I) diffusion models [3], [4], [21], [29], [31], but applying these methods to video-editing is challenging since videos require consistent edits across frames to maintain temporal coherence.

Motivation RAVE [15] is a zero-shot video editing framework that leverages pre-trained text-to-image (T2I) diffusion models, such as Stable Diffusion [22]. It enables editing of style, attributes, and shapes in videos while maintaining temporal consistency through a novel noise shuffling strategy. However, we find that RAVE struggles with instance-level editing and preserving details unrelated to the edit. In this project, we build upon RAVE and propose a new framework for instance-level video editing. Our method combines Mask-Guided Compositional Denoising, Foreground-Background Composition, Harmonization Post-Processing [17], and EasyInv [33] for DDIM Inversion. Our main contributions are as follows:

- **Instance-Level Video Editing:** We introduce Mask-Guided Compositional Denoising to apply edits to specific objects.
- **Foreground-Background Alignment:** We use Harmonization post-processing [17] to align the edited target with the background, ensuring smooth and natural integration.
- **Preservation of Details:** We adopt EasyInv [33] for DDIM inversion to preserve distinctive details that are unrelated to the edit prompt.

II. Related Work

Training-free Image Editing Several training-free methods have been proposed for text-driven image editing, leveraging attention mechanisms for precise and localized modifications. Prompt-to-Prompt [10] enables edits by adjusting cross-attention layers based on textual changes, while DiffEdit [6] uses automatically generated masks to guide semantic edits. Blended Diffusion [2] integrates local text-driven edits with spatial blending for seamless results, and Blended Latent Diffusion [1] further accelerates the process by operating in a lower-dimensional latent space. These methods achieve efficient and controllable editing without requiring extensive retraining.

Training-free Video Editing with T2I Models Several training-free methods have been proposed for text-driven video editing, leveraging T2I diffusion models for efficient and consistent results. Pix2Video [4] and TokenFlow [9] focus on temporal consistency, with Pix2Video [4] using sparse-causal attention and latent guidance to ensure coherence, while TokenFlow [9] employs feature-level smoothing to reduce flickering by propagating diffusion features based on inter-frame correspondences. Recently, RAVE [15] introduces randomized noise shuffling to achieve faster and more consistent edits, enabling practical applications without additional training. VidToMe [19] enhances temporal consistency through video token merging, improving the

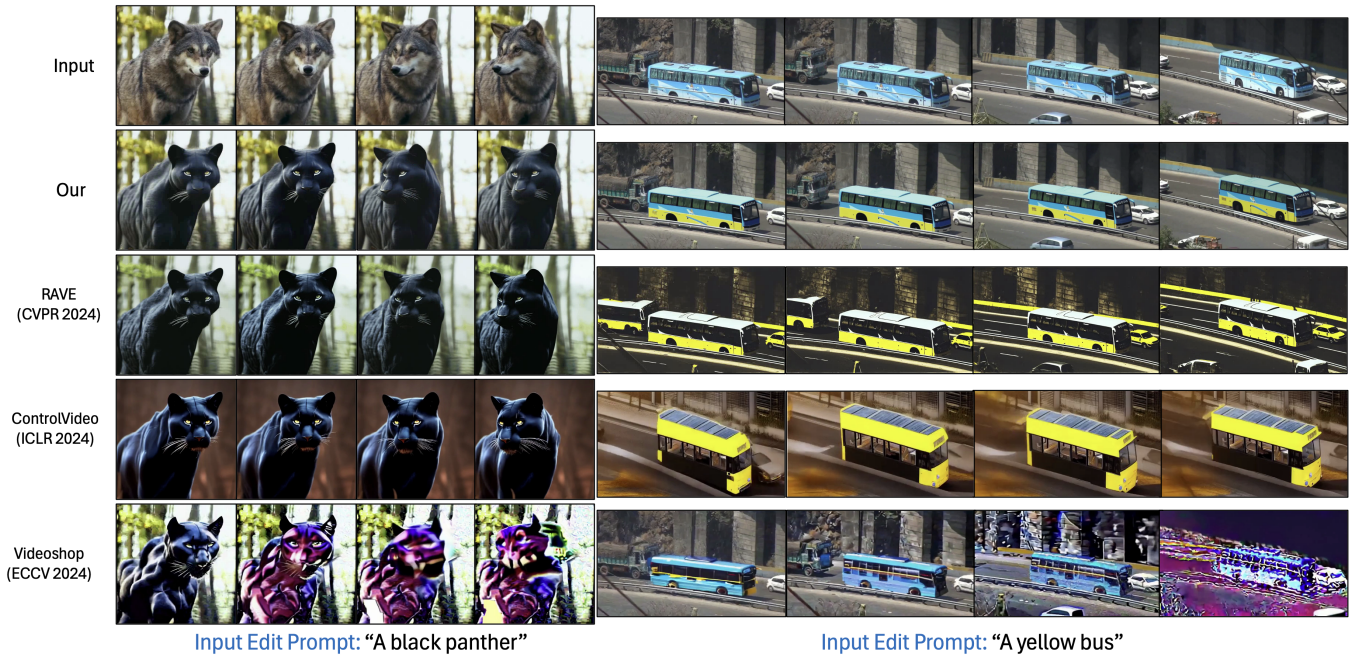

Fig. 1. ***DRIVE:*** A training-free framework for precise instance-level video editing.

performance of zero-shot video editing. Other approaches, such as FLATTEN [5] and FRESCO [32], emphasize maintaining spatial-temporal consistency across frames. FLATTEN [5] uses optical flow-guided attention within the diffusion model's U-Net, ensuring patches along the same motion path attend to one another, while FRESCO [32] improves correspondence to achieve seamless video translations. For more precise and localized modifications, ControlVideo [34] introduces conditional controls for user-guided text-driven edits, and UniVST [28] provides a unified framework for localized video style transfer, enabling efficient style modifications without retraining.

Training-free Video Editing with Video-based Models With the advancements in video-based models, more works have chosen to leverage their power for video editing. FateZero [21] utilizes attention features during inversion to preserve motion and structure, enabling zero-shot editing without requiring per-prompt training. Text2Video-Zero [18] adapts text-to-image diffusion models for video generation, enriching latent codes with motion dynamics and reprogramming frame-level self-attention to maintain temporal consistency. Similarly, Rerender-A-Video [31] employs hierarchical cross-frame constraints to ensure coherence in video-to-video translation, while BIVDiff [24] bridges image and video diffusion models to provide a training-free framework for general-purpose video synthesis. Videoshop [8], on the other hand, enables localized semantic edits through noise-extrapolated diffusion inversion, balancing precision and consistency.

Although video-based models have shown increasing capabilities in text-driven video editing, they often face significant computational costs and stability issues. Their high resource consumption, particularly memory usage, makes it challenging to edit longer videos, while occasional instability can result in inconsistent edits across frames. Consequently, our work focuses on efficient image-based approaches to address these challenges and ensure broader applicability.

III. RAVE

In this section, we introduce the RAVE [15] framework. Fig. 2 provides an overview. RAVE is a zero-shot video editing framework that builds on pre-trained text-to-image diffusion models to edit video styles, attributes, and shapes. RAVE integrates DDIM inversion to accelerate diffusion model sampling by using a non-Markov process. This allows us to map video frames to the diffusion latent space for efficient editing. Additionally, ControlNet [34] is employed to guide the editing process with spatial and temporal conditions. The main difference of RAVE compared to prior works is its use of grid trick and noise-shuffling technique. The grid trick arranges video frames into a grid and processes them as a single image, significantly improving temporal and style consistency across frames. During denoising, the noise-shuffling technique extends the grid trick, allowing it to handle longer videos with minimal memory requirements. This approach maintains smooth transitions between frames without needing extra memory and achieves 25% faster processing than existing methods while handling 90 frames at 512×512 resolution in just 5 minutes.

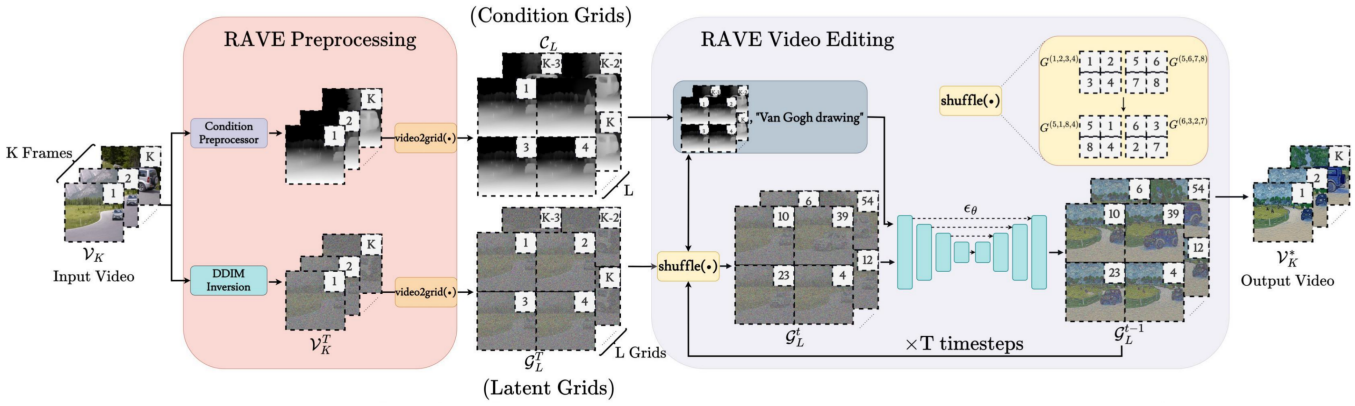

Fig. 2. ***An illustration of RAVE.*** RAVE starts by performing a DDIM inversion with the pre-trained T2I model and condition extraction with a conditional preprocessor to the input video. RAVE performs diffusion denoising for T timesteps using conditional grids, latent grids and the target prompt as inputs for ControlNet. Random shuffling is applied to the latent grids and condition grids at each denoising step. After completing T timesteps, the latent grids are rearranged to generate the final output video. Source: [15]

IV. Methods

In this paper, we propose a series of methods to address the challenges of realistic video editing and instance-aware content generation. Our approach builds upon the RAVE architecture, leveraging its grid-based processing capabilities to maintain temporal consistency while enabling precise and seamless manipulation of target objects. The key components of our methodology are: Mask-Guided Compositional Denoising, Foreground-Background Composition, Harmonization Post-Processing [17], and EasyInv [33] to replace vanilla DDIM Inversion, each addressing specific aspects of the editing pipeline. Together, these components enable high-quality video generation with instance-level control, ensuring compositional coherence and preserving the original background when user prompts focus on foreground elements. The overview of the pipeline is illustrated in Fig. 3. In the preprocessing stage, we leverage a zero-shot visual tracking model SAMURAI [30] to generate precise binary masks for tracking the target object across frames. These masks are later utilized in both the denoising stage and the foreground-background composition. Details of our pipeline are presented below.

A. Mask-Guided Compositional Denoising

Mask-Guided Compositional Denoising utilizes semantic guidance and spatial masking to enable instance-level control in video generation. The denoising process is designed to ensure that style transformations applied to the target (foreground) objects are consistent with the overall scene, mitigating stark differences between the transformed foreground and the surrounding context.

At each denoising step, two types of noise are generated using a U-Net model: **conditional noise** $\hat{\epsilon}_{\text{cond}}$ and **unconditional noise** $\hat{\epsilon}_{\text{uncond}}$. Conditional noise focuses exclusively on the target object, while unconditional noise represents the background. These two noise maps are composited using an object mask (M) as follows:

$$\hat{\epsilon} = M\hat{\epsilon}_{\text{cond}} + (1 - M)\hat{\epsilon}_{\text{uncond}}, \tag{1}$$

where M is a binary mask indicating the region of the target object, and $\hat{\epsilon}$ is the final noise prediction used for denoising.

Conditional grids, latent grids, and text embeddings are the primary inputs to the U-Net model, which generates both types of noise in parallel. By incorporating textual information, the model ensures that the semantic context of the target object is aligned with the intended modifications. During the denoising process, the object mask guides the separation of noise application, enabling precise edits to the object while leaving the background unaffected.

The denoising process incorporates spatial masks to isolate target objects and conditionally generates noise to align the foreground's appearance with its surroundings. By doing so, it ensures that the style of the transformed foreground elements remains visually coherent and natural when later composed with the original background.

B. Foreground-Background Composition

While the generated video background is produced through unconditional denoising, it often contains artifacts introduced during the diffusion process. These artifacts can degrade the overall quality and realism of the output, making it unsuitable for seamless video editing. To address this, following the denoising stage, the refined foreground will be composited with the original background, preserved in its unaltered form.

To facilitate this composition, we utilize masks generated through a method inspired by the SAMURAI framework. By incorporating motion-aware memory selection and leveraging temporal motion cues, we generate precise, instance-level masks that maintain spatial and temporal consistency across

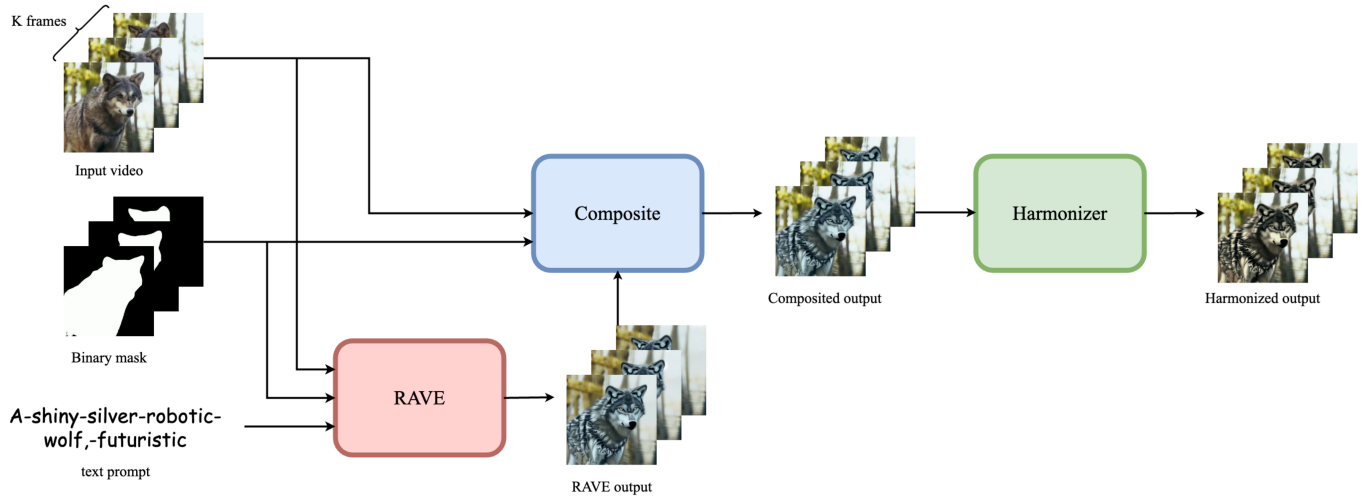

Fig. 3. ***An illustration of our pipeline.*** Our process begins by passing the input video through the RAVE model, where Mask-Guided Compositional Denoising and EasyInv techniques are applied. The transformed foreground is then composited with the original, unaltered background to preserve spatial integrity and maintain the scene's authenticity. Finally, the composited video undergoes Harmonization Post-Processing to refine color, tone, and overall visual balance, producing the final output video.

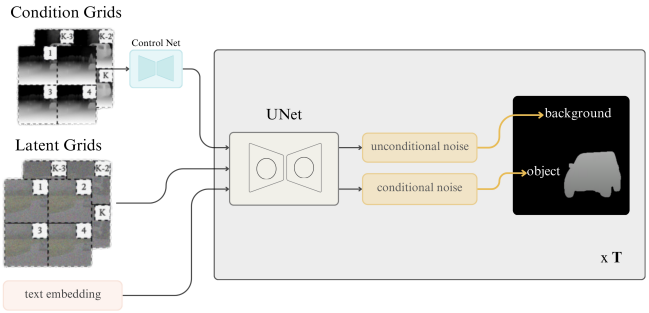

Fig. 4. ***Denoising process*** Illustration of denoising process in Mask-Guided Compositional Denoising, generating and compositing conditional and unconditional noise with an object mask to ensure coherent style transformations.

frames. These masks ensure that the refined foreground aligns accurately with the original background during the compositing process, even in challenging scenarios such as occlusion or fast motion.

By discarding the generated background and relying on the authentic background, this approach preserves the natural fidelity and structural integrity of the scene. Additionally, it ensures that the stylistic transformations applied to the foreground blend seamlessly with the preserved context, resulting in a cohesive and visually appealing final output.

C. Harmonization

While Mask-Guided Compositional Denoising enables precise instance-aware editing by applying modifications exclusively to target objects, it may result in noticeable discrepancies in lighting, texture, and color tone between the foreground and background regions of the video.

To resolve these inconsistencies, we leverage a method inspired by the Harmonizer framework, which combines neural networks with white-box image filters for efficient and reliable image-level adjustments. The process involves a cascade regressor that predicts filter arguments for critical attributes such as brightness, contrast, color temperature, and shadow refinement. These filters are then applied sequentially to align the visual properties of the target objects with the rest of the scene, ensuring a cohesive appearance.

The cascade regressor ensures that the interdependencies between filters are respected, leading to stable and accurate adjustments. Additionally, a dynamic loss strategy is employed to balance the learning process, emphasizing filters with higher complexity to enhance precision. This step not only refines the visual coherence of each frame but also maintains temporal consistency across frames, eliminating flickering artifacts.

D. EasyInv

After applying harmonization, the edited video may still suffer from a loss of distinctive characteristics and fine details. This issue arises due to noise introduced during the DDIM process, which can accumulate and degrade the quality of the final output. To address this, we employ the EasyInv technique, which strategically reduces error by reinforcing the influence of the initial latent state during the inversion process.

EasyInv enhances the vanilla DDIM Inversion by incorporating exponential moving average (EMA) at selected time steps. Specifically, it aggregates the current latent state with the previous one, prioritizing the original latent state and mitigating the impact of noise. This aggregation ensures that the generated output remains closer to the original, retaining fine details and distinctive features that are often compromised

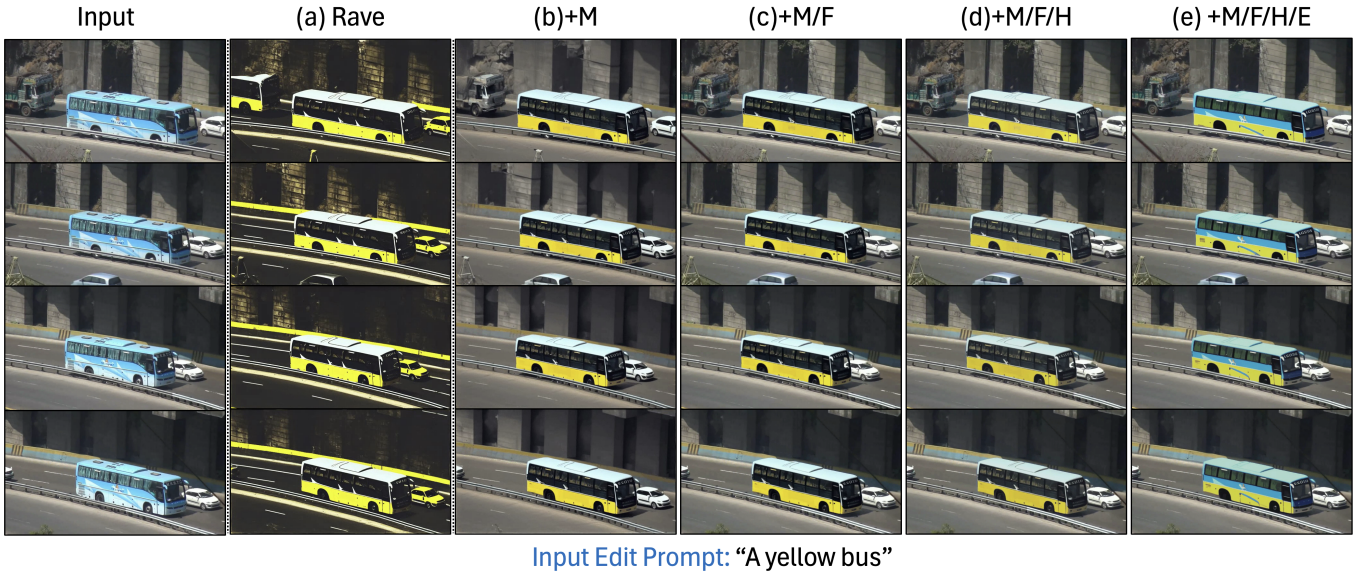

Fig. 5. ***Ablation Study.*** We analyze the contributions of components by incrementally adding: (b) Mask-guided Compositional Denoising, (c) Foreground-Background Composition, (d) Harmonization, and (e) EasyInv.

during the diffusion process. Mathematically, EasyInv modifies the latent state z_t at specific time steps $\bar{t}$ as follows:

$$z_{\bar{t}} = \eta z_{\bar{t}} + (1-\eta) z_{\bar{t}-1}, \tag{2}$$

where η is a trade-off parameter controlling the balance between the current and previous states. By applying this blending approach, EasyInv effectively reduces reconstruction errors without relying on iterative optimizations, significantly improving computational efficiency.

The integration of EasyInv into our pipeline enables the recovery of lost details and ensures that the output video maintains its distinctive characteristics, enhancing both realism and overall quality.

V. Experiments

In this section, we access the effectiveness of our pipeline through qualitative results and an ablation study on each components of our pipeline.

A. Qualitative Comparison

For qualitative comparison, we select the `A black panther` demo video from the RAVE demo website, which contains 27 frames. Additionally, we choose a challenging video from the LaSOT (Large-scale Single Object Tracking) dataset [7], featuring a fast-moving bus on a highway with 90 frames to edit.

For the baselines, we consider our base framework RAVE [15], ControlVideo [34], and Videoshop [8]. Due to the computational limitations of Videoshop when editing long videos, we divide the video into segments, with each segment conditioned on the last frame of the preceding section. The results are presented in Fig. 1.

For the black panther example, our method demonstrates precise and natural instance-level editing without altering the original background, outperforming RAVE in this regard. For Videoshop, the editing quality deteriorates across frames as errors accumulate and propagate through the segments.

In the challenging bus example, the two baselines other than RAVE struggle to produce natural editing results due to the fast-moving foreground objects. In contrast, our method achieves both natural and precise instance-level editing, effectively improve upon the base framework RAVE.

B. Ablation Study

We evaluate the effect of each component in our pipeline by incrementally adding them. The results are presented in Fig. 5. Adding Mask-guided Compositional Denoising significantly improves instance-level editing, producing precise and natural results.

As discussed in Section IV-B, the background would change and may contain artifacts if we did not composite the foreground and background. To address this, we ensure background invariance by compositing the edited foreground with the original background. Incorporating Harmonization post-processing further aligns the visual attributes of the foreground and background, resulting in a more seamless output.

Finally, replacing the standard DDIM inversion with EasyInv enables the preservation of more distinctive details from the original input. We also observe that using EasyInv slightly reduces flickering in the edited video. This improvement is particularly noticeable in the rear part of the target object in the bus instance demo video.

VI. Conclusion

In this project, we built upon RAVE to develop a training-free video editing pipeline that produces natural edited results with precise instance-level control. Our pipeline begins with the

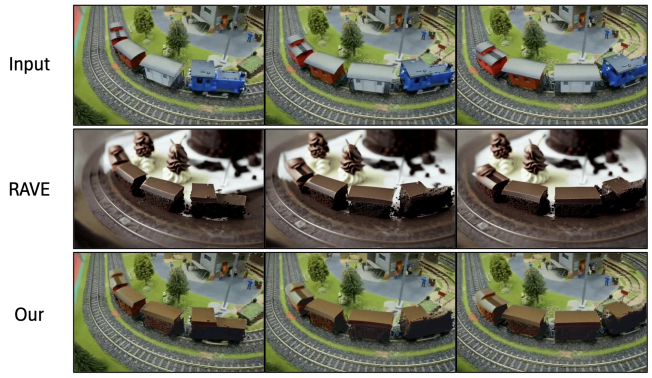

Fig. 6. ***Limitation*:** Our pipeline relies on precise object tracking to generate instance-level masks. Here, we present cases where foreground-background composition fails due to imprecise masks. A closer look at the last two carriages reveals unnatural editing results.

core concept of using masks to composite unconditional and conditional noise during the denoising steps. This approach applies edits only to the target while avoiding drastic style shifts. Additionally, we enhance the pipeline by incorporating several existing works, such as post-processing harmonization for foreground-background alignment. We also replace the vanilla DDIM inversion to better preserve details, while improving the edited video quality by mitigating flickering. Qualitative accessments showcase the effectiveness of our pipeline.

VII. Limitations

While our pipeline successfully enables a training-free approach for precise and natural instance-level video editing, several limitations remain. First, our method depends on the accuracy of object masks generated by the visual tracking model. For example, Fig. 6 illustrates a case where the object masks are not sufficiently precise, leading to suboptimal editing results. Second, similar to RAVE and other training-free frameworks, the quality of the edits largely depends on the capabilities of the pre-trained text-to-image model used for editing. Lastly, minor flickering artifacts persist, even with modifications to the vanilla DDIM inversion. This issue appears to be an inherent limitation of training-free video editing frameworks that rely on pre-trained text-to-image models.

VIII. Acknowledgment

Our project is developed based on the following open-source code bases:

- RAVE: https://github.com/RehgLab/RAVE
- Harmonizer: https://github.com/ZHKKKe/Harmonizer
- EasyInv: https://github.com/potato-kitty/EasyInv
- SAMURAI: https://github.com/yangchris11/samurai

IX. Team Contribution

Pin-Han Huang

- Presentation
- EasyInv
- Report Writing (Experiments, Limitations)

Yi-Ru Wu

- Mask-Guided Compositional Denoising
- Report Writing (Introduction, Conclusion)

Te-Wei Chen

- SAMURAI
- Demo Video

Ren-Wei Liang

- Videoshop
- Report Writing (Related Work)

Cheng-Yu Lin

- Harmonization
- Report Writing (Method)

REFERENCES

[1] Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *ACM transactions on graphics (TOG)*, 2023.

[2] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2022.

[3] Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. Text2live: Text-driven layered image and video editing. In *European Conference on Computer Vision (ECCV)*, 2022.

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

[5] Yuren Cong, Mengmeng Xu, christian simon, Shoufa Chen, Jiawei Ren, Yanping Xie, Juan-Manuel Perez-Rua, Bodo Rosenhahn, Tao Xiang, and Sen He. FLATTEN: optical FLow-guided ATTENtion for consistent text-to-video editing. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

[6] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.

[7] Heng Fan, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Hexin Bai, Yong Xu, Chunyuan Liao, and Haibin Ling. Lasot: A high-quality benchmark for large-scale single object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2019.

[8] Xiang Fan, Anand Bhattad, and Ranjay Krishna. Videoshop: Localized semantic video editing with noise-extrapolated diffusion inversion. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *European Conference on Computer Vision (ECCV)*, 2024.

[9] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

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

[11] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 6840–6851, 2020.

[13] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[14] Ondřej Jamriška, Šárka Sochorová, Ondřej Texler, Michal Lukáč, Jakub Fišer, Jingwan Lu, Eli Shechtman, and Daniel Sỳkora. Stylizing video by example. *ACM Transactions on Graphics (TOG)*, 2019.

[15] Ozgur Kara, Bariscan Kurtkaya, Hidir Yesiltepe, James M Rehg, and Pinar Yanardag. Rave: Randomized noise shuffling for fast and consistent video editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[16] Yoni Kasten, Dolev Ofri, Oliver Wang, and Tali Dekel. Layered neural atlases for consistent video editing. *ACM Transactions on Graphics (TOG)*, 2021.

[17] Zhanghan Ke, Chunyi Sun, Lei Zhu, Ke Xu, and Rynson WH Lau. Harmonizer: Learning to perform white-box image and video harmonization. In *European Conference on Computer Vision (ECCV)*, 2022.

[18] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)*, 2023.

[19] Xirui Li, Chao Ma, Xiaokang Yang, and Ming-Hsuan Yang. Vidtome: Video token merging for zero-shot video editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[20] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021.

[21] Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. Fatezero: Fusing attentions for zero-shot text-based video editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)*, 2023.

[22] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2022.

[23] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 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 (CVPR)*, 2023.

[24] Fengyuan Shi, Jiaxi Gu, Hang Xu, Songcen Xu, Wei Zhang, and Limin Wang. Bivdiff: A training-free framework for general-purpose video synthesis via bridging image and video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[25] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.

[26] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning (ICML)*, 2015.

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

[28] Quanjian Song, Mingbao Lin, Wengyi Zhan, Shuicheng Yan, Liujuan Cao, and Rongrong Ji. Univst: A unified framework for training-free localized video style transfer, 2024.

[29] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[30] Cheng-Yen Yang, Hsiang-Wei Huang, Wenhao Chai, Zhongyu Jiang, and Jenq-Neng Hwang. Samurai: Adapting segment anything model for zero-shot visual tracking with motion-aware memory. *arXiv preprint arXiv:2411.11922*, 2024.

[31] Shuai Yang, Yifan Zhou, Ziwei Liu, and Chen Change Loy. Rerender a video: Zero-shot text-guided video-to-video translation. In *SIGGRAPH Asia 2023 Conference Papers*, 2023.

[32] Shuai Yang, Yifan Zhou, Ziwei Liu, and Chen Change Loy. Fresco: Spatial-temporal correspondence for zero-shot video translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[33] Ziyue Zhang, Mingbao Lin, Shuicheng Yan, and Rongrong Ji. Easyinv: Toward fast and better ddim inversion. *arXiv preprint arXiv:2408.05159*, 2024.

[34] Min Zhao, Rongzhen Wang, Fan Bao, Chongxuan Li, and Jun Zhu. Controlvideo: Adding conditional control for one shot text-to-video editing. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.