Patch'n Play: Zero-Shot Video Editing by Fusing Local and Global Patches

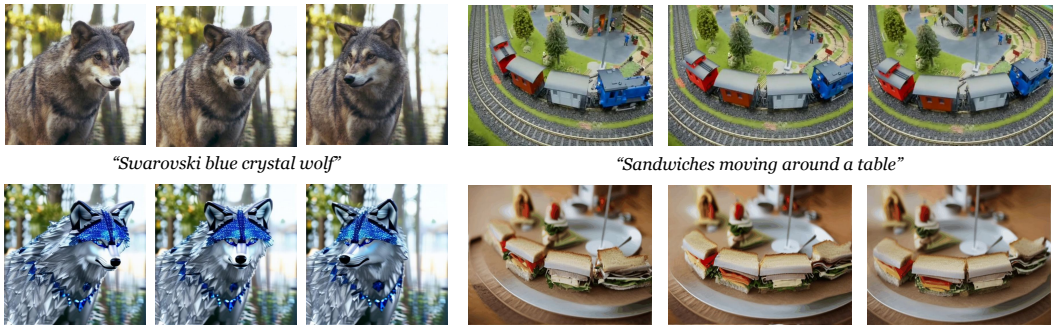

Figure 1: **Video Editing Samples Generated by Patch'n Play.** We present Patch'n Play, a novel video editing method that leverages local diffusion features, aggregating information based on inter-frame correspondences by fusing diffusion paths. Patch'n Play enhances both spatial and temporal consistencies in video edits in a zero-shot manner using pre-trained text-to-image models like Stable Diffusion [21]. It can apply style edits, like making a *wolf* look like a *Swarovski blue crystal*, and complex edits, like turning a *train* into a *sandwich*.

Abstract

Recent progress in diffusion-based models has shown remarkable achievements in generating images from text prompts. Despite these advancements, video editing methods have lagged in achieving comparable visual quality and editing capabilities. This paper introduces `Patch'n Play`, a novel zero-shot video editing method that leverages both local and global latent features to enhance temporal consistency. Unlike previous approaches that prioritize global consistency at the expense of local consistency, our method aggregates and fuses local features from each frame along with global information shared across multiple frames. Compatible with pre-trained text-to-image diffusion models, our approach does not require prompt-specific training or user-generated masks. Our qualitative and quantitative analysis underscores `Patch'n Play`'s superior performance across a wide array of video editing contexts against existing methods.

1 Introduction

Diffusion-based generative models [25, 11, 21] excel in high-quality image generation and editing through text prompts [22, 10, 1]. These successes have led to applications such as object editing, personalized content creation, and inpainting. Extending such advances to video remains challenging. Text-to-video methods [24, 12] require costly large-scale training, while atlas-based approaches [14, 3] rely on labor-intensive processes. Recent efforts adapt pre-trained text-to-image (T2I) diffusion models [13, 20, 28, 4], but often struggle with temporal consistency [16] or require additional training [28, 17].

We propose a zero-shot video editing method that fuses local and global diffusion features across frames to achieve temporal consistency without training or user masks. Unlike prior works that

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: The First Workshop on Generative and Protective AI for Content Creation.

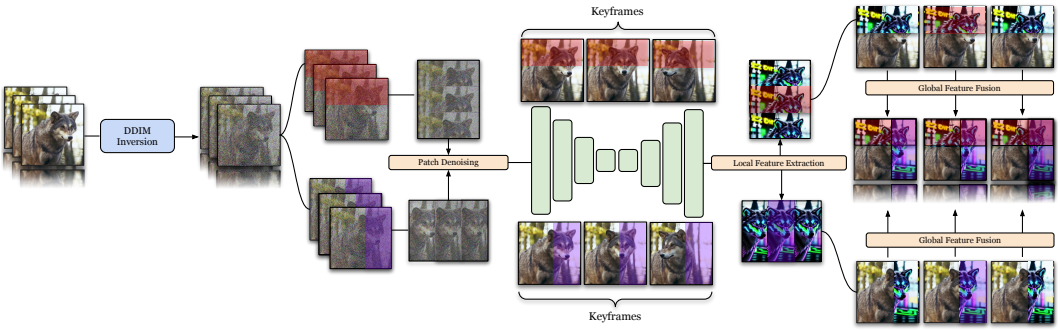

Figure 2: **Patch'n Play Framework.** From left to right, we first invert input frames into noisy latents via DDIM inversion. Keyframes (top row) remain unedited to anchor temporal consistency. Next, local feature extraction partitions each frame's latent into overlapping sub-frames, with red regions representing horizontal scans and purple regions representing vertical scans. These sub-frames are collectively denoised (center), leveraging U-Net and ControlNet for structural guidance. Finally, in global feature fusion (right), the overlapping areas of the vertical and horizontal sub-frames (intersection) are aggregated to form a unified latent, which is decoded to produce the edited video. This pipeline preserves both local details (via sub-frame denoising) and global coherence (via overlap fusion and keyframe anchoring).

emphasize global coherence, our approach preserves both fine-grained details and overall consistency (Fig. 3). Compatible with any pre-trained T2I model such as Stable Diffusion [21], our framework also leverages spatial guidance via ControlNet [31] and pre-trained style models [5]. Experiments show effectiveness across diverse edits, from style transfer (e.g., *wolf* $\rightarrow$ *crystal*) to complex transformations (e.g., *train* $\rightarrow$ *sandwich*).

2 Methodology

Our framework, **Patch'n Play**, performs zero-shot video editing through a five-stage pipeline (Fig. 2): inverting input frames, sampling keyframes, extracting local features, denoising sub-frames, and fusing global features. Each stage addresses a specific challenge of video editing, balancing efficiency, temporal consistency, and spatial coherence.

Inverting Input Frames. Given an input video $\mathcal{V} = \{I^{(1)}, \ldots, I^{(N)}\}$, each frame is first encoded into a latent space with a pretrained VAE:

$$x_0^{(m)} = \mathcal{E}(I^{(m)}), \tag{1}$$

where $x_0^{(m)} \in \mathbb{R}^{C \times H/8 \times W/8}$. This step compresses pixel space into a compact representation while preserving semantic details. To enable diffusion-based editing, the latents are inverted into noisy states via DDIM inversion [26]:

$$x_T^{(m)} = \sqrt{\alpha_T} x_0^{(m)} + \sqrt{1 - \alpha_T}\, \epsilon, \quad \epsilon \sim \mathcal{N}(0, I). \tag{2}$$

The inversion provides a deterministic mapping from clean to noisy latents, ensuring edits remain reproducible across runs. No editing is applied at this stage; ControlNet is introduced later during denoising for structural guidance.

Keyframe Sampling. Editing all frames independently often leads to flickering or motion discontinuities. To counter this, we select M_k uniformly spaced keyframes that remain unedited and serve as temporal anchors. During denoising, their latents are concatenated with editable frames and passed through U-Net self-attention:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V. \tag{3}$$

By participating in attention, the keyframes transmit unaltered motion and structure to their neighbors. This ensures the edited frames evolve consistently with the original dynamics while preserving the underlying scene geometry.

Local Feature Extraction. While keyframes enforce global alignment, preserving *local* details such as textures, edges, or small objects requires finer control. Each noisy latent is partitioned into overlapping sub-frames using dual scanning strategies:

$$z_{t,v}^{(m)} = x_t^{(m)}[:, h_s : h_e, w_s : w_e], \tag{4}$$

where v indexes a sliding window. Horizontal and vertical scans ensure coverage across both axes, while overlaps prevent seams at patch boundaries. This design allows the framework to capture fine-grained details and to remain robust to complex or diagonal motion, as objects appear across multiple overlapping sub-frames.

Patch Denoising. For each view v, sub-frames from all frames in the batch are stacked together:

$$s_{t,v} = \text{stack}(z_{t,v}^{(1)}, \ldots, z_{t,v}^{(M)}), \tag{5}$$

creating a temporally aligned bundle of local regions. These are jointly denoised with U-Net and ControlNet, conditioned on the prompt $\mathcal{P}$ and cropped control maps C_v:

$$s_{t-1,v} = \psi(s_{t,v}, t, \mathcal{P}, C_v). \tag{6}$$

Unlike frame-wise denoising, this stacked formulation allows attention layers to exchange information across time, ensuring consistency in both object appearance and motion. ControlNet provides additional structure (e.g., depth maps), ensuring that the edits remain faithful to the original spatial layout.

Global Feature Fusion. After denoising, the sub-frames must be recombined into full latents. Overlapping contributions are averaged:

$$x_{t-1}^{(m)} = \frac{\sum_v z_{t-1,v}^{(m)}}{\sum_v \mathbf{1}_v}, \tag{7}$$

where $\mathbf{1}_v$ counts sub-frames covering each position. This averaging eliminates seams, harmonizes textures, and achieves frame-wide coherence. Finally, the refined latents are decoded back to pixels:

$$\hat{I}^{(m)} = \mathcal{D}(x_0^{(m)}). \tag{8}$$

This stage ensures that local details (from sub-frame denoising) and global structures (from overlap fusion and keyframe anchoring) come together in a coherent final output. The iterative combination of temporal and spatial consistency mechanisms produces smooth, flicker-free, and visually unified edited videos.

3 Experiments

Baselines. We compare Patch'n Play with recent video editing methods: Pix2Video [4], RAVE [13], TokenFlow [9], Rerender [29], Text2Video-Zero [16], FLATTEN [6], FRAG [30], and Video-P2P [18]. Approaches vary: flow-guided attention (FLATTEN), feature smoothing (TokenFlow), noise shuffling (RAVE), random warping (Text2Video-Zero), optical flow (Rerender), dynamic receptive fields (FRAG), and unconditional embeddings (Video-P2P). Pix2Video and Video-P2P are restricted to 8 frames due to GPU limits. All baselines are run with official code and default settings.

Implementation Details. Patch'n Play uses batch size $M = 8$, $T = 50$ diffusion steps, and guidance scale 7.5. Latents are extracted with overlapping vertical ($k_h = H/16, s_h = H/128$) and horizontal ($k_w = W/8, s_w = W/256$) scans to ensure smooth fusion. ControlNet with depth maps provides structural guidance. For fairness, all methods (including ours) use Stable Diffusion v1.5 with DDIM inversion and the same guidance scale, without negative prompts. Experiments run on a single NVIDIA L40 GPU.

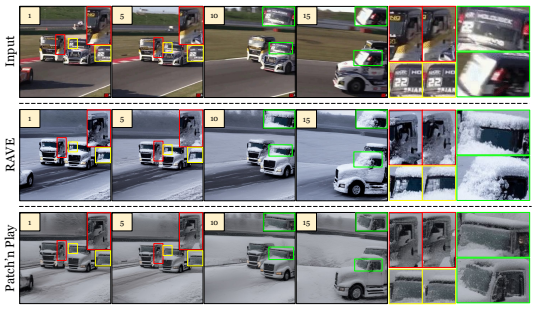

Figure 3: **Comparison between RAVE and Patch'n Play**

Table 1: **Quantitative Comparisons.** Higher is better. "-" indicates a method cannot process more than 8 frames. Best results in **bold**.

Method	CLIP-F ↑			WarpSSIM ↑			CLIP-T ↑			Q_{edit} ↑		
	8f	36f	90f	8f	36f	90f	8f	36f	90f	8f	36f	90f
Text2Video-Zero [16]	93.77	92.15	93.58	65.41	40.58	68.36	26.60	27.11	27.23	17.39	11.00	18.61
Rerender [29]	91.54	86.42	89.65	70.13	46.32	72.48	23.54	24.21	25.63	16.50	11.21	18.57
TokenFlow [9]	94.73	92.86	93.83	76.19	**53.46**	81.15	29.38	30.02	31.07	**22.38**	16.04	25.21
Pix2Video [4]	89.31	-	-	58.17	-	-	28.93	-	-	16.82	-	-
RAVE [13]	94.83	94.65	95.32	72.87	51.19	81.06	29.26	29.89	**31.13**	21.32	15.30	25.23
FLATTEN [6]	94.57	91.36	93.44	75.23	51.03	80.37	28.11	29.16	30.39	21.14	14.88	24.42
FRAG [30]	95.19	93.42	93.86	**76.21**	53.38	80.31	29.63	30.45	30.95	22.58	16.31	24.85
Video-P2P [18]	92.13	-	-	72.43	-	-	24.14	-	-	19.17	-	-
Patch'n Play	**95.34**	**95.73**	**95.91**	74.55	52.85	**82.37**	**29.89**	**30.61**	31.08	22.28	**16.17**	**25.60**

Qualitative Results. Examples are shown in Figs. 1 and 6, covering style edits (e.g., Picasso), structural edits (e.g., man→skeleton), and combined edits. In comparisons (Fig. 5), Patch'n Play achieves stronger global consistency and text alignment. Text2Video-Zero fails under background motion, Rerender suffers from style drift, TokenFlow blurs details due to smoothing, and Pix2Video produces flicker. Patch'n Play avoids these issues and can also enhance existing methods: Fig. 8(e) shows our framework improving temporal stability when applied to RAVE.

Quantitative Evaluation. Following prior work [9, 4, 29, 13], we evaluate using CLIP-F (frame similarity), WarpSSIM (flow-based SSIM), CLIP-T (text alignment), and Q_{edit} (WarpSSIM $\times$ CLIP-T). We test on 128 video–prompt pairs from previous benchmarks [9, 13, 3] and DAVIS [19], with 8-, 36-, and 90-frame sequences. Results (Table 1) show Patch'n Play consistently achieves the best or near-best scores, especially on long sequences where temporal consistency is hardest.

User Study. To complement metrics, we conducted a user study following [13]. Participants compared outputs on three criteria: **General Editing (GE)**, **Temporal Consistency (TC)**, and **Textual Alignment (TA)**. Results (Table 4) show Patch'n Play is most frequently preferred across all categories.

Ablation Studies. Fig. 8 summarizes ablations. (a) Increasing keyframes improves temporal consistency. (b) Larger overlapping windows reduce seams. (c) Our method remains robust under ego/exo-motion, partial framing, and occlusions. (d) Scanning both horizontally and vertically yields best results. (e) ControlNet ablations (lineart, softedge, depth) show stable performance. (f) Applied to RAVE, our approach improves both spatial and temporal consistency.

Method	GE	TC	TA
RAVE	70.1%	55.2%	26.6%
TokenFlow	39.4%	52.6%	24.1%
Rerender	14.9%	24.8%	13.0%
FLATTEN	33.3%	42.3%	33.3%
FRAG	66.7%	61.9%	71.4%
Video-P2P	14.3%	9.5%	4.8%
Patch'n Play	**85.7%**	**76.1%**	**90.4%**

Figure 4: **User Study.** Percentage of times each method was ranked among top two for GE/TC, and top for TA.

4 Discussion

Our method obtains the highest CLIP-F scores across all frame lengths, indicating strong temporal consistency. While TokenFlow achieves the highest WarpSSIM score for 36-frame sequences, our method surpasses all competitors for 90-frame sequences, confirming that our approach ensures better structural consistency for longer videos. For CLIP-T, which measures alignment with the textual prompt, our method outperforms all baselines in 8-frame and 36-frame sequences. For the 90-frame sequences Patch'n Play obtains the second best score with 0.05 difference, indicating improved semantic coherence across frames. Additionally, our Q_{edit} scores surpass all baselines across 36-frame and 90-frame sequences, highlighting that our method strikes a superior balance between structural integrity (WarpSSIM) and textual alignment (CLIP-T). Overall, while some methods, such as TokenFlow, show strong performance in short sequences, our method excels in both short and long sequences, achieving the best trade-off between temporal consistency, structural coherence, and textual alignment.

References

[1] Avrahami, O., Fried, O., Lischinski, D.: Blended latent diffusion. arXiv preprint arXiv:2206.02779 (2022)

[2] Avrahami, O., Fried, O., Lischinski, D.: Blended latent diffusion. ACM Transactions on Graphics (TOG) **42**(4), 1–11 (2023)

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

[4] Ceylan, D., Huang, C.H.P., Mitra, N.J.: Pix2video: Video editing using image diffusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 23206–23217 (2023)

[5] CivitAI: Civitai. `https://civitai.com/` (2024), accessed: 2023-11-16

[6] Cong, Y., Xu, M., Simon, C., Chen, S., Ren, J., Xie, Y., Perez-Rua, J.M., Rosenhahn, B., Xiang, T., He, S.: Flatten: optical flow-guided attention for consistent text-to-video editing. arXiv preprint arXiv:2310.05922 (2023)

[7] Couairon, G., Verbeek, J., Schwenk, H., Cord, M.: Diffedit: Diffusion-based semantic image editing with mask guidance. In: The Eleventh International Conference on Learning Representations (2023), `https://openreview.net/forum?id=3lge0p5o-M-`

[8] Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. In: The Eleventh International Conference on Learning Representations (2023), `https://openreview.net/forum?id=NAQvF08TcyG`

[9] Geyer, M., Bar-Tal, O., Bagon, S., Dekel, T.: Tokenflow: Consistent diffusion features for consistent video editing. arXiv preprint arXiv:2307.10373 (2023)

[10] Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-or, D.: Prompt-to-prompt image editing with cross-attention control. In: The Eleventh International Conference on Learning Representations (2023), `https://openreview.net/forum?id=_CDixzkzeyb`

[11] Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)

[12] Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. arXiv:2204.03458 (2022)

[13] Kara, O., Kurtkaya, B., Yesiltepe, H., Rehg, J.M., Yanardag, P.: Rave: Randomized noise shuffling for fast and consistent video editing with diffusion models. arXiv preprint arXiv:2312.04524 (2023)

[14] Kasten, Y., Ofri, D., Wang, O., Dekel, T.: Layered neural atlases for consistent video editing. ACM Transactions on Graphics (TOG) **40**(6), 1–12 (2021)

[15] Kawar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., Mosseri, I., Irani, M.: Imagic: Text-based real image editing with diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6007–6017 (2023)

[16] Khachatryan, L., Movsisyan, A., Tadevosyan, V., Henschel, R., Wang, Z., Navasardyan, S., Shi, H.: Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 15954–15964 (October 2023)

[17] Liew, J.H., Yan, H., Zhang, J., Xu, Z., Feng, J.: Magicedit: High-fidelity and temporally coherent video editing. arXiv preprint arXiv:2308.14749 (2023)

[18] Liu, S., Zhang, Y., Li, W., Lin, Z., Jia, J.: Video-p2p: Video editing with cross-attention control. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8599–8608 (2024)

[19] Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., Sorkine-Hornung, A.: A benchmark dataset and evaluation methodology for video object segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 724–732 (2016)

[20] Qi, C., Cun, X., Zhang, Y., Lei, C., Wang, X., Shan, Y., Chen, Q.: Fatezero: Fusing attentions for zero-shot text-based video editing. arXiv preprint arXiv:2303.09535 (2023)

[21] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)

[22] Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22500–22510 (2023)

[23] Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems **35**, 36479–36494 (2022)

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

[25] Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International Conference on Machine Learning. pp. 2256–2265. PMLR (2015)

[26] Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)

[27] Valevski, D., Kalman, M., Molad, E., Segalis, E., Matias, Y., Leviathan, Y.: Unitune: Text-driven image editing by fine tuning a diffusion model on a single image. ACM Transactions on Graphics (TOG) **42**(4), 1–10 (2023)

[28] Wu, J.Z., Ge, Y., Wang, X., Lei, S.W., Gu, Y., Shi, Y., Hsu, W., Shan, Y., Qie, X., Shou, M.Z.: Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7623–7633 (2023)

[29] Yang, S., Zhou, Y., Liu, Z., Loy, C.C.: Rerender a video: Zero-shot text-guided video-to-video translation. arXiv preprint arXiv:2306.07954 (2023)

[30] Yoon, S., Koo, G., Kim, G., Yoo, C.D.: Frag: Frequency adapting group for diffusion video editing. arXiv preprint arXiv:2406.06044 (2024)

[31] Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847 (2023)

A Related Work

Text-Driven Image Editing. Methods such as Dream-Booth [22] and Textual Inversion [8] demonstrate diverse image generation through fine-tuning in a few-shot manner. UniTune [27] and Imagic [15], both based on the Imagen [23] model, exhibit strong editing performances. Recent training-free methods like Prompt-to-Prompt [10], DiffEdit [7], Blended Diffusion [1], and Blended Latent Diffusion [2] achieve local and detailed editing by leveraging attention properties.

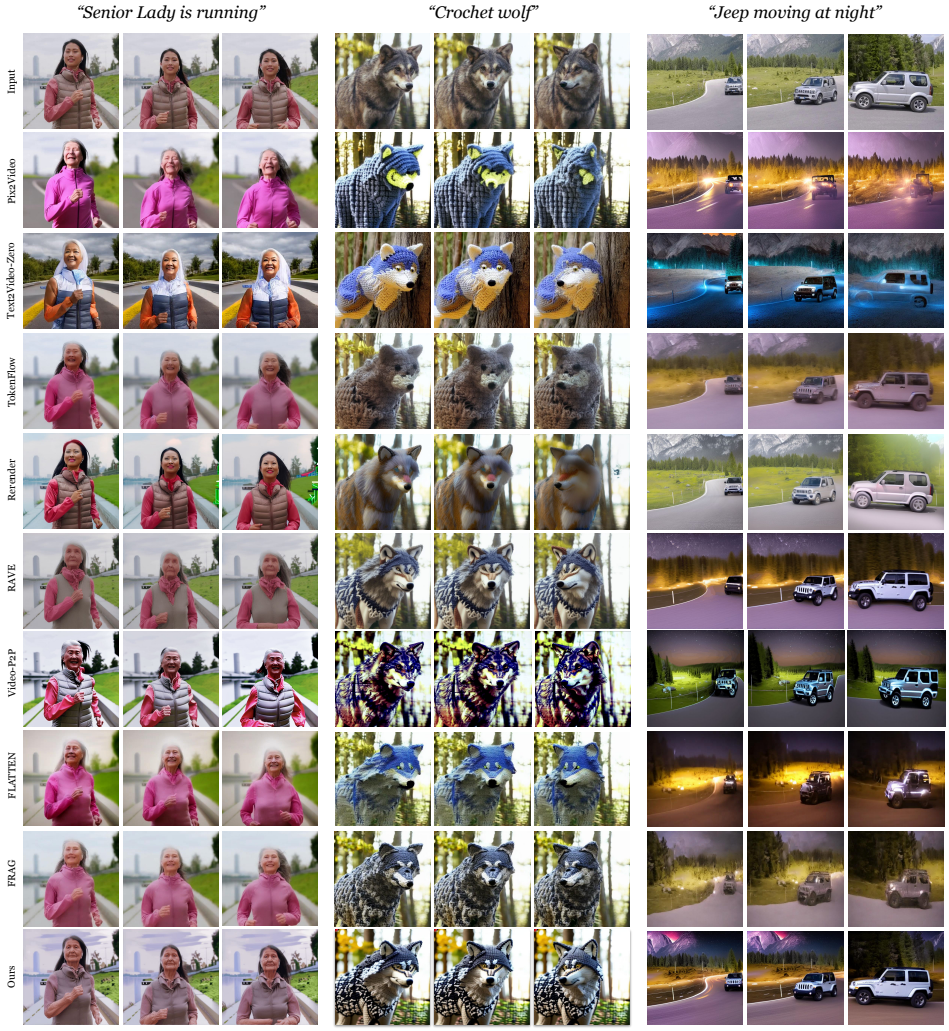

Figure 5: **Qualitative Comparison.** We compare our approach with other methods on short and long videos featuring diverse motions and objects. We demonstrate style editing (e.g., "Jeep moving at night") as well as complex edits (e.g., transforming a wolf into a crochet style). We use videos featuring backgrounds in motion ("a woman running") and different objects engaged in various activities. For a fair comparison, all methods used SD 1.5 as the base method. Please zoom in for clarity, and see the supplementary material for the full videos.

Text-Driven Video Editing. Recent studies emphasize zero-shot, training-free approaches for practical applicability. Pix2Video [4] employs sparse-causal attention for temporal consistency along with latent guidance, using the predictions of the original images as a proxy at each denoising step. FateZero [20] uses attention features during inversion for spatio-temporal preservation and blending, claiming that those are better in preserving motion and structure compared to that of during sampling. FateZero [20], similar to Pix2Video [4], requires source prompts as it is built on the Prompt-to-Prompt [10] editing technique. This method necessitates specific types of prompts on the source prompt, thereby limiting editing diversity. Additionally, both are constrained to shorter

clips due to memory limitations. Text2Video-Zero [16] synthesizes and edits videos with cross-frame attention, initial frame integration, and background smoothing. Text2Video-Zero [16] and Rerender [29] heavily rely on off-the-shelf methods and optical flow, limiting consistency over longer videos. Rerender-A-Video [29] employs hierarchical cross-frame constraints for temporal consistency, while TokenFlow [9] focuses on feature-level smoothing to reduce the effects of flickering. RAVE [13] uses a grid-based strategy to perform noise-shuffling to achieve temporal consistency. FLATTEN [6] utilizes a flow-guided attention mechanism, while FRAG [30] uses a dynamic receptive field during the diffusion process to improve the quality of edited frames. Furthermore, Video-P2P [18] makes the inversion process more efficient by using an optimized, shared unconditional embedding technique.

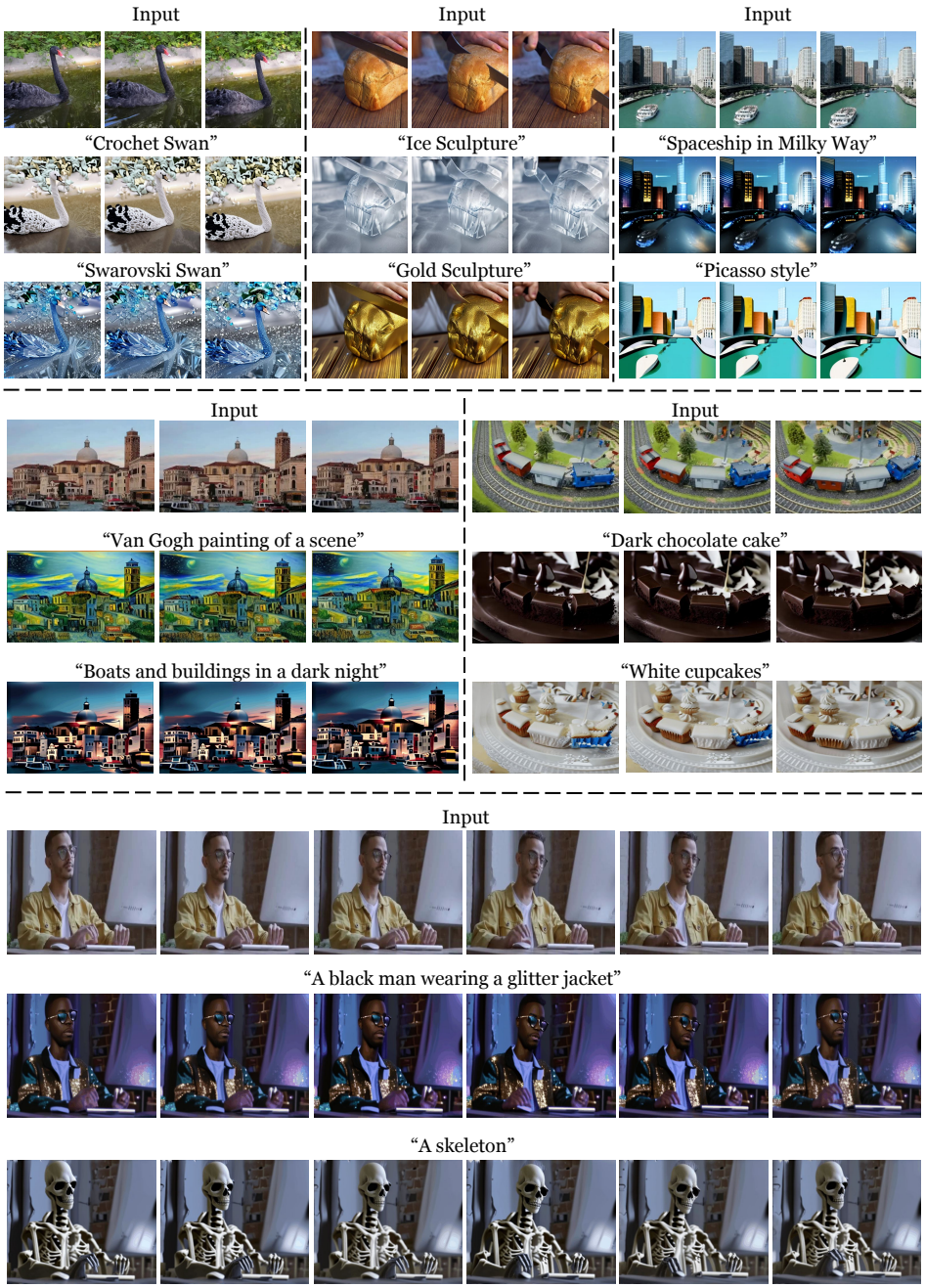

Figure 6: **Qualitative Results.** Our method works on a variety of editing scenarios. Please see the appendix for the full videos.

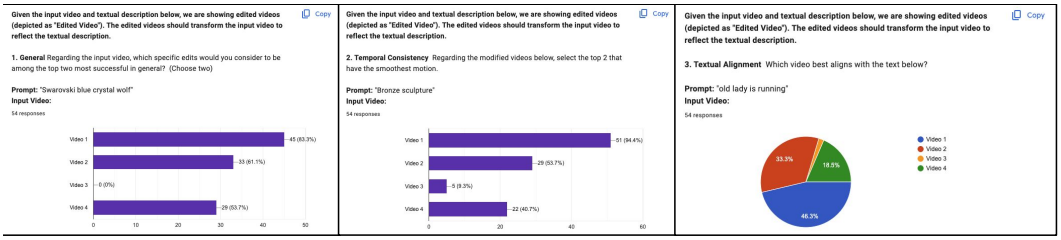

Figure 7: **Screenshots from the User Study.** To better assess how well our method compares to competing approaches, we conducted a user study involving 50 participants on the Prolific.com platform. We categorized evaluation questions into 3: (Left) General Editing, (Middle) Temporal Consistency, (Right) Textual Alignment. Our figure shows a single question and user answer for each type of evaluation question. In these survey questions, Video 1, Video 2, Video 3, and Video 4 correspond to Patch'n Play, RAVE [13], Rerender [29], and Tokenflow [9] respectively. In all questions, the options were randomly shuffled.

B User Study

Fig. 7 shows a screenshot of the survey form, featuring a single question alongside the input video and edited videos. Note that the order of the videos in each question is randomly shuffled to provide a fair comparison. The order of the questions are also randomly shuffled for every participant. A total of 21 video-prompt pairs were shown to the users. Our results in Tab. 4 show that Patch'n Play was consistently among the top methods selected for successful editing (GE), temporal consistency (TC), and textual alignment (TA).

C More Qualitative Results

Figures 1 and 6 showcase examples of video editing using Patch'n Play. We demonstrate style editing (e.g., 'Picasso style'), complex shape editing (e.g., transforming a man into a skeleton), and both applied simultaneously (e.g., 'Spaceship in Milky Way'). We use videos featuring backgrounds in motion and different objects engaged in various activities. Furthermore, we present a qualitative comparison in Fig. 5 with baseline methods. Text2Video-Zero appears effective with videos featuring constant backgrounds, despite a significant color change in the woman's jacket in Fig. 5; however, it encounters challenges when there is motion in the background, such as the trees in the same figure. Rerender heavily utilizes optical flow along with keyframe propagation; however, these are not enough to perform editing while keeping style over time (observe the color change of the car in Fig. 5). Although TokenFlow can maintain structural consistency over time, it experiences overall blurring in the entire image due to 'feature-level smoothing' being applied. While Pix2Video's edits seem to apply the style in a relevant manner, their videos have significant flickering issues (please refer to the supplementary material to view the videos for all methods). In contrast, our method preserves global consistency while aligning with the text prompt. Moreover, our method can be seamlessly integrated into existing video-editing frameworks to enhance temporal consistency, showcasing its versatility and adaptability.

D Ablations

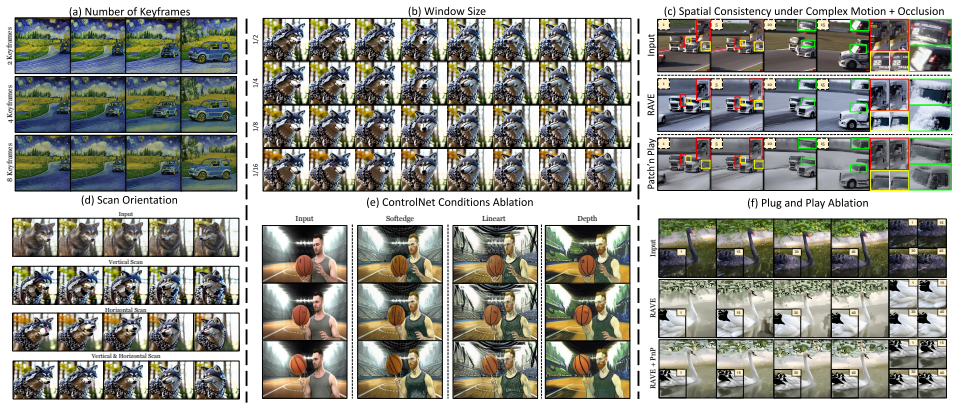

Figure 8: **Ablation Studies** Please zoom in for more clarity. (a) We show how keyframe count affects temporal consistency. (b) Window size ablation is illustrated by showing [window size / spatial size] over the y-axis. (c) We demonstrate the robustness of our method under complex motions: Ego-motion, Exo-motion, Partial Framing and Occlusion of the main objects. (d) We ablate scan orientation in horizontal, vertical and horizontal & vertical directions, where best results are obtained by scanning in both directions. (e) We demonstrate the robustness of our method in a range of ControlNet conditions. (f) We demonstrate the plug-and-play nature of our method when taking RAVE as the backbone model.