

APPENDIX

A QUANTITATIVE RESULTS

A.1 QUANTITATIVE METRICS AND MODEL DEPENDENCIES

Table 1: **Quantitative comparison and model dependencies between STR-Match and existing methods.** For quantitative metrics, bold black and red numbers indicate the best and second-best performance for each metric, respectively. Note that FC (Frame Consistency), CS (CLIP Similarity), and VB (VE-Bench Score) are higher-is-better metrics, while BL (BG-LPIPS) and ME (Motion Error) are lower-is-better.

Method	Base model	External model	FC (\uparrow)	CS (\uparrow)	BL (\downarrow)	ME (\downarrow)
FateZero (Qi et al., 2023)	T2I (sd1.4 ²)	–	0.979	31.56	0.139	2.749
FLATTEN (Cong et al., 2024)	T2I (sd2.1 ³)	RAFT	0.980	31.43	0.277	2.748
VideoGrain (Yang et al., 2025a)	T2I (sd1.5 ⁴)	RAFT, SAM-Track, ControlNet	0.978	31.16	0.062	1.943
Ground-A-Video (Jeong & Ye, 2024)	T2I (sd1.5)	ControlNet, GLIGEN, RAFT, ZoeDepth, OWL-ViT	0.969	30.62	0.244	3.348
DMT (Yatim et al., 2024)	T2V (LaVie)	–	0.981	31.94	0.499	5.741
UniEdit (Bai et al., 2024)	T2V (LaVie)	SAM-Track	0.979	31.02	0.134	2.632
STR-Match (Ours, w/o mask)	T2V (LaVie)	–	0.981	31.61	0.216	2.402
STR-Match (Ours, w/ mask)	T2V (LaVie)	SAM-Track	0.981	31.68	0.103	1.932

Table 1 presents evaluation metrics used in the radar graph shown in Figure 6 of Section 5.3 in the main paper along with the base diffusion models and external models used by each method. While some existing algorithms rely on multiple external models such as ControlNet (Zhang et al., 2023), GLIGEN (Li et al., 2023a), and ZoeDepth (Bhat et al., 2023), STR-Match operates without the need for such models, optionally using SAM-Track only for generating binary masks. Overall, the proposed method, STR-Match with LaVie, whether applied with and without masks, achieves a balanced and superior performance across all metrics compared to other methods. Notably, while DMT generates high quality videos (evidenced by strong FC and CS), these outputs often lack fidelity to the source video (reflected in poor scores for BL and ME). Although we have provided the quantitative metrics, we encourage readers to consult the qualitative results in Figure 4 in the main paper, Appendix C, and supplementary material, as these metrics are incomplete and often fail to reflect the true quality of videos.

A.2 VE-BENCH SCORE

Table 2: **Quantitative comparison on VE-Bench score.** Bold black indicates the best VE-Bench score. STR-Match achieves the highest VE-Bench score among all training-free methods, highlighting its effectiveness across consistency, fidelity, and temporal coherence.

Method	VE-Bench
FateZero (Qi et al., 2023)	0.496
FLATTEN (Cong et al., 2024)	0.401
VideoGrain (Yang et al., 2025a)	0.503
Ground-A-Video (Jeong & Ye, 2024)	0.222
DMT (Yatim et al., 2024)	0.438
UniEdit (Bai et al., 2024)	0.583
STR-Match (Ours, w/o mask, LaVie)	0.530
STR-Match (Ours, w/ mask, LaVie)	0.595
CogVideoX-V2V (Yang et al., 2025b)	0.554
STR-Match (Ours, w/o mask, CogVideoX)	0.587

²<https://huggingface.co/CompVis/stable-diffusion-v1-4>

³<https://huggingface.co/stabilityai/stable-diffusion-2-1>

⁴<https://huggingface.co/stable-diffusion-v1-5/stable-diffusion-v1-5>

To further validate the effectiveness of STR-Match, we evaluate its performance using VE-Bench (Sun et al., 2025), a comprehensive benchmark designed to assess video editing quality from multiple perspectives. As shown in Table 2, STR-Match with LaVie achieves the highest VE-Bench score among all baselines, evaluated on a diverse set of 54 edited videos. Although UniEdit obtains a comparable VE-Bench score, our main paper demonstrates that it lacks flexibility in handling domain shifts effectively (see Figure 4 in the main paper and Figure 7, Figure 8 in the supplementary material).

Moreover, STR-Match with CogVideoX outperforms CogVideoX-V2V in terms of VE-Bench score, highlighting its strong generalization capability on recent DiT-based T2V models. This advantage is also qualitatively observable in Figure 5 of the main paper and Figure 9 in the supplementary material.

In summary, these results establish STR-Match as the new state-of-the-art in text-guided video editing, demonstrating both superior quantitative performance and greater flexibility in handling diverse editing scenarios.

B COMPUTATIONAL COMPLEXITY OF STR-MATCH ON DiT-BASED MODEL

In the DiT-based setting, STR score computation is significantly more efficient than calculating a full 3D attention map. The complexity of computing full 3D attention is $O(h \times f^2 \times n^2 \times c)$, where h is the number of heads, f the number of frames, n the number of spatial positions, and c the channel dimension. In contrast, our pseudo self-attention and pseudo temporal-attention maps have lower complexities of $O(f \times h \times n^2 \times c)$ and $O(n \times h \times f^2 \times c)$, respectively. The STR score computation itself requires only $O(f \times h \times n^2)$ operations. This design significantly reduces computational overhead while effectively preserving spatiotemporal relevance. Moreover, we further enhance efficiency by computing STR scores using only the first two attention blocks in CogVideoX-2B during optimization, striking a balance between performance and efficiency.

C QUALITATIVE RESULTS

C.1 ADDITIONAL COMPARISONS WITH OTHER METHODS

We provide video files in the supplementary material that showcase a variety of video editing examples—ranging from typical cases with minimal domain shifts to more challenging ones with significant shape transformations, as illustrated in Figures 7 and 8. For instances involving extreme shape transformations (*e.g.* ‘cat \rightarrow dragon’, ‘goldfish \rightarrow snake’), many competing methods distort the target objects to conform to the shape of the source, resulting in unnatural edited outputs. In the cases of extreme domain change (*e.g.* ‘cat \rightarrow robot dog’, ‘goldfish \rightarrow donuts’), few other methods transfer only a part of the intended concept while the majority fail to perform any meaningful editing. In contrast, STR-Match with LaVie consistently delivers successful video edits across these challenging scenarios, highlighting its flexibility and robust editing capabilities. This strong performance is also evident when STR-Match is applied to CogVideoX, outperforming CogVideoX-V2V as demonstrated in Figure 9. We strongly encourage readers to view the HTML file included in the zip archive for a more comprehensive understanding of STR-Match’s editing capabilities.

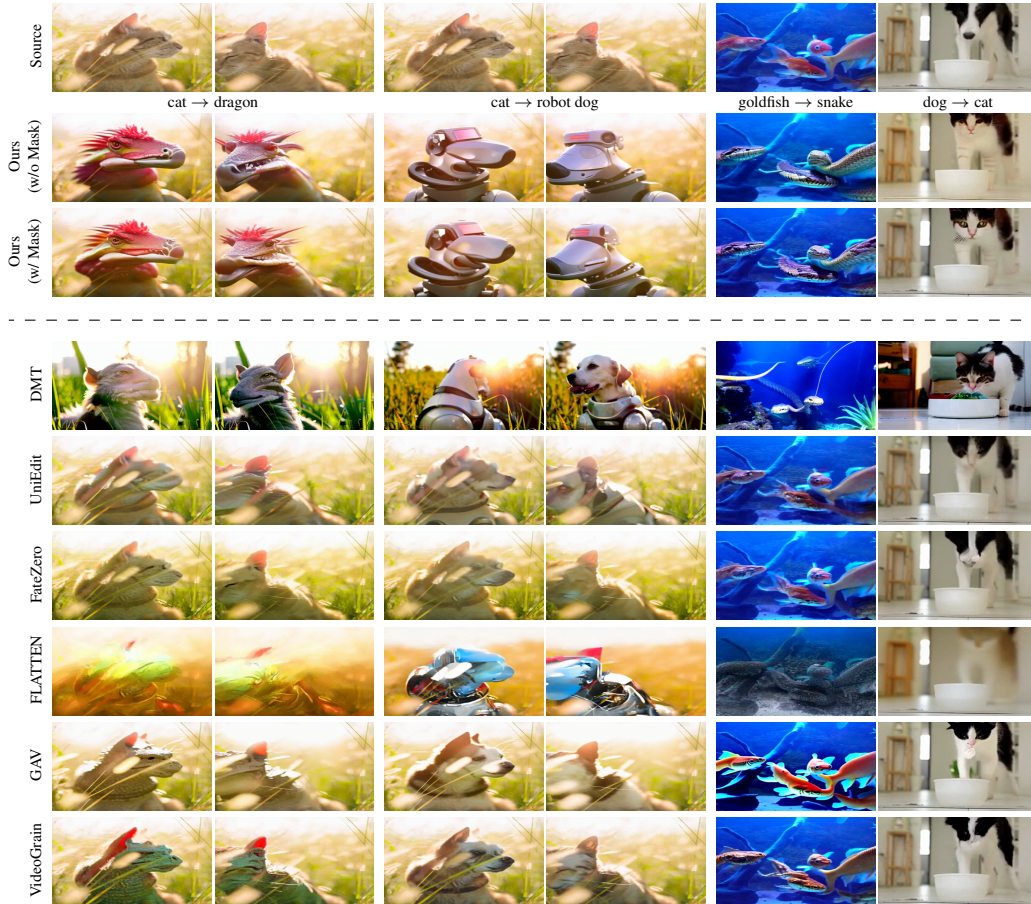


Figure 7: **Additional qualitative comparisons between STR-Match with LaVie and existing methods.** This figure illustrates the performance of STR-Match with LaVie in challenging scenarios, including cat → dragon, cat → robot dog, goldfish → snake, and dog → cat.

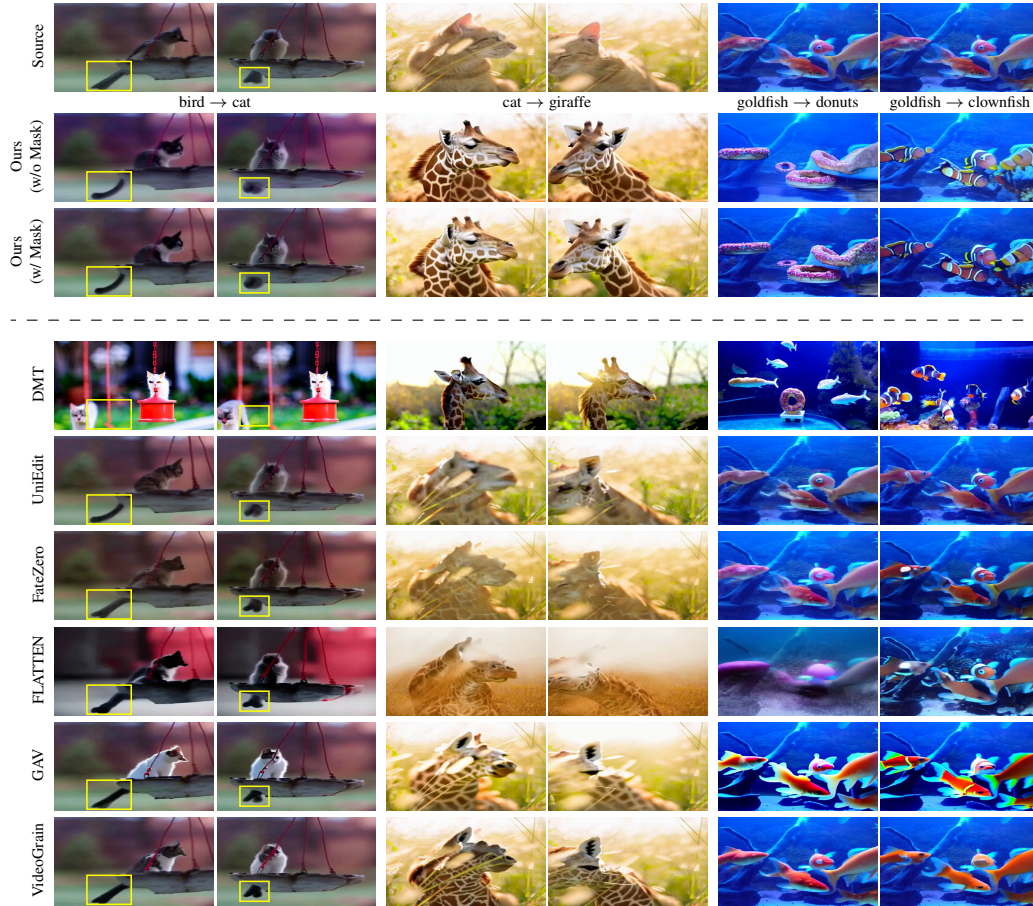


Figure 8: **Qualitative comparisons between STR-Match with LaVie and existing methods.** This figure illustrates the performance of STR-Match with LaVie in challenging scenarios, including bird \rightarrow cat, cat \rightarrow giraffe, goldfish \rightarrow donuts, and goldfish \rightarrow clownfish.



Figure 9: **Qualitative comparisons between STR-Match and CogVideoX-V2V.** This figure presents qualitative comparisons between STR-Match and CogVideoX-V2V across various editing scenarios, including balloon → cabinet, penguin → puffin, and kangaroo → wallaby.

C.2 STR-MATCH WITH ZEROSCOPE

In the main experiments based on the U-Net architecture, we adopt LaVie as the pretrained T2V model for applying STR-Match. To demonstrate the general applicability of our method, we also apply STR-Match to other T2V models, such as Zeroscope⁵. Figure 10 illustrates the results of STR-Match using Zeroscope as the base model. The results demonstrate that STR-Match can effectively edit videos with Zeroscope, achieving similar performance to LaVie.



Figure 10: **Qualitative results of STR-Match using Zeroscope.** STR-Match can be applied to Zeroscope, achieving similar performance to LaVie.

D OBJECT DELETION/ADDITION

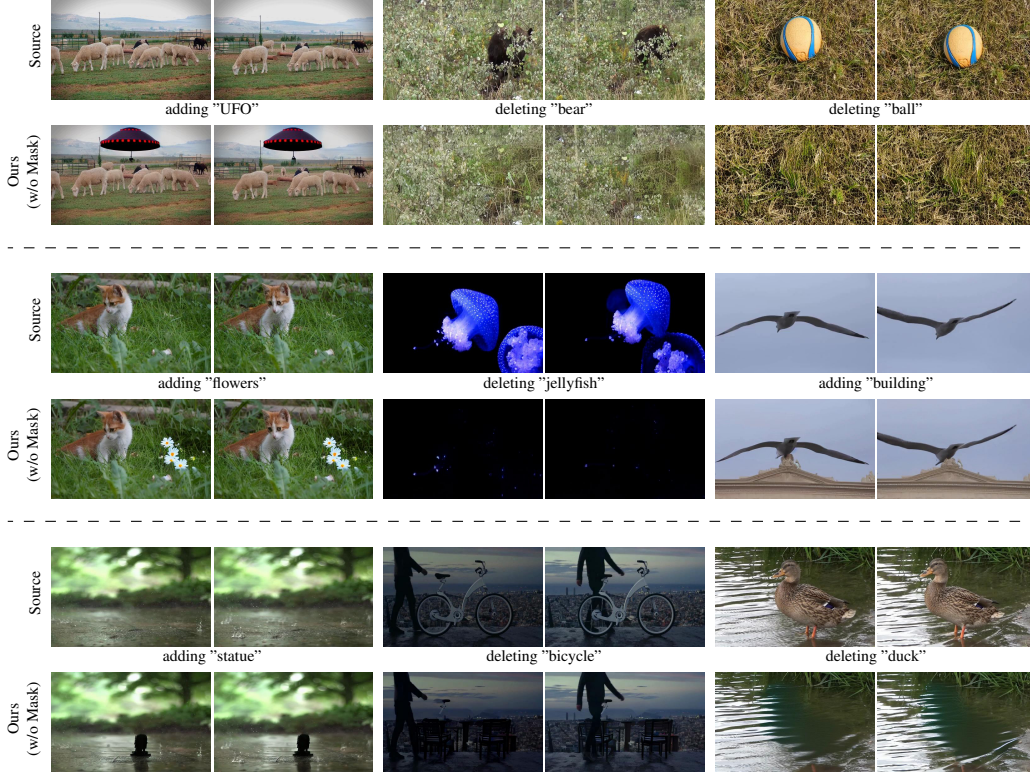


Figure 11: **Object Deletion and Addition Results of STR-Match with LaVie.** This figure presents multiple examples of object deletion and addition performed using STR-Match with LaVie. Our proposed method consistently demonstrates strong performance in both object deletion and addition tasks.

Although STR-Match was originally developed for object replacement, it generalizes naturally to other editing tasks, including object deletion and addition. For these scenarios, we adapt the latent optimization process by computing the STR score exclusively in unmasked regions, allowing the model to preserve spatiotemporal pixel relevance where needed.

⁵https://huggingface.co/cerspense/zeroscope_v2_576w

To assess generalization beyond object replacement, we conduct additional experiments on the U-Net based setting using a 30-video subset from our dataset, focusing on object deletion and addition tasks. Qualitative results (Figure 11) demonstrate that STR-Match maintains high fidelity and preserves key spatial structures from the source video, even when modifying semantic content. Quantitatively, STR-Match achieved VB scores of **0.577** for object deletion and **0.657** for object addition. These scores are comparable to those from our main object replacement experiments (0.595 with mask, 0.530 without mask), indicating consistent performance across different editing tasks.

These results confirm that STR-Match effectively extends to a broader range of video editing applications beyond object replacement, while preserving essential spatiotemporal structures.

E ABLATION STUDY

E.1 FLEXIBILITY OF STR SCORE

To evaluate the effectiveness of the proposed STR score, we compare STR-Match with a baseline that optimizes the concatenation of self- and temporal-attention maps. For the baseline, we set the guidance strength $\lambda = 0.08$ to ensure that the edited video preserves essential attributes of the source, such as motion dynamics, and perform optimization using the L2 loss. Figure 12 shows that STR-Match produces significantly higher quality videos compared to the baseline. For instance, in the ‘dog \rightarrow cat’ case, the baseline method generates oversaturated colored video and in the ‘turtle \rightarrow shark’ case, it fails to alter the sharks’ shape into that of turtles. These two examples illustrate that naïvely using self- and temporal-attention maps as guidance imposes overly strict constraints, whereas the proposed STR score effectively captures key features while providing sufficient flexibility for editing, as it optimizes values that are conceptually derived from the element-wise multiplication of self- and temporal-attention maps. Moreover, Table 3 supports this conclusion, as fidelity-related metrics (FC and CS) are higher for our method. Although the baseline better preserves background and motion, it often fails to transform objects, as demonstrated in Figure 12.

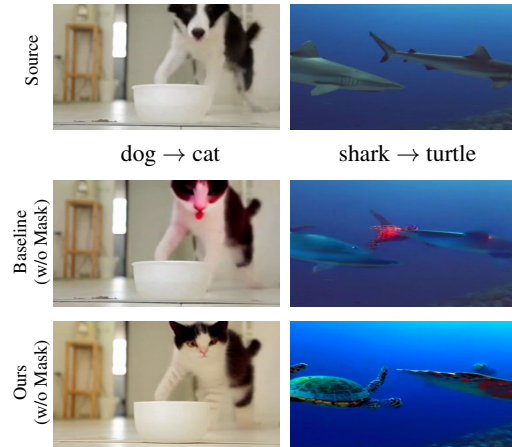


Figure 12: Quantitative comparison between STR-Match and the baseline.

Table 3: Quantitative comparison between STR-Match and the baseline without mask. Bold numbers indicate the better score for each metric.

Method	FC (\uparrow)	CS (\uparrow)	BL (\downarrow)	ME (\downarrow)
Baseline	0.979	31.24	0.117	2.293
Ours	0.981	31.61	0.216	2.402

Table 4: Ablation study on λ values. Bold black and red numbers indicate the best and second-best scores for each metric, respectively.

λ	FC (\uparrow)	CS (\uparrow)	BL (\downarrow)	ME (\downarrow)
0.005	0.982	31.60	0.271	3.120
0.01	0.981	31.61	0.216	2.402
0.015	0.979	31.33	0.196	2.225

E.2 ABLATION ON THE HYPERPARAMETER λ

λ is the only hyperparameter in STR-Match, which controls the guidance strength during optimization. To investigate its effect, we conduct an ablation study using LaVie by varying the guidance strength λ across three values: 0.005, 0.01, and 0.015. As shown in Table 4, we empirically observe that smaller values of λ yield higher fidelity scores (FC, CS) but struggle to preserve background and motion dynamics, whereas larger values promote preservation at the cost of fidelity. To balance these objectives, we adopt $\lambda = 0.01$ for all experiments.

E.3 ADDITIONAL ABLATION STUDY ON STR-MATCH: RELEVANCE TYPE AND OPTIMIZATION STEPS

Table 5: **Additional ablation study on STR-Match with LaVie.** We report VE-Bench scores across different relevance types and optimization steps. The bold black number denotes the best-performing setting. The results show that bidirectional relevance with full optimization consistently outperforms the directional and optimization step variant.

Relevance type / Optimization step	50	40	30	20	10
Bidirectional	0.506	0.389	0.480	0.361	0.118
Directional	0.492	0.345	0.384	0.392	0.315

We additionally perform detailed ablation studies to analyze the design choices of STR-Match, focusing on two key components: the relevance type used in the STR score and the number of optimization steps. We compare our default bidirectional relevance computation with a directional variant that considers only the first term in Equation (3) of the main paper. Additionally, we test a range of optimization step counts: [10, 20, 30, 40, 50], where no optimization is applied at certain time steps, and the standard DDIM reverse step is used instead. The results using LaVie are summarized in Table 5, which shows VE-Bench scores across these different settings. Our default configuration—bidirectional relevance combined with full optimization steps—consistently yields the best performance, highlighting the effectiveness and robustness of the proposed design.

F LIMITATIONS

While STR-Match produces satisfying editing results, even in the challenging scenarios like flexible shape transformations, it still has some limitations. One limitation is its inability to edit multiple objects into different targets simultaneously. Although a workaround exists—editing each object individually with its corresponding mask—this approach is highly inefficient. Additionally, while the method supports flexible shape transformations, it produces suboptimal results when the object’s size varies significantly. We plan to address remaining limitations in future work.

G SOCIETAL IMPACT

STR-Match is a training-free video editing algorithm that leverages pretrained T2V models. Since it relies heavily on these pretrained models, there is a potential risk of generating videos with unintended or inappropriate contents. However, we believe this issue can be indirectly mitigated by carefully controlling the training data used for the underlying T2V models.