

# Supplementary Material for FFP-300K: Unlocking Generalizable Video Editing via First-Frame Propagation

Xijie Huang<sup>1\*</sup>, Chengming Xu<sup>2\*</sup>, Donghao Luo<sup>2</sup>, Xiaobin Hu<sup>2</sup>, Peng Tang<sup>2</sup>, Xu Peng<sup>2</sup>, Jiangning Zhang<sup>2</sup>  
Chengjie Wang<sup>2</sup>, Yanwei Fu<sup>1,3†</sup>

<sup>1</sup>Fudan University, <sup>2</sup>Tencent YouTu Lab, <sup>3</sup>Shanghai Innovation Institute

[ffp-300k.github.io](https://github.com/ffp-300k)

## A. Additional Information about FFP-300K

The video frames shown in our figures have been packaged and uploaded. Please refer to the accompanying zip file for details.

### A.1. Dataset Construction

**Prompts used.** Our data construction pipeline follows a two-track modular pipeline and we use Qwen2.5-VL-72B-Instruct [1] to produce the prompts for both local editing and global stylization.

#### A.1.1. Local Editing

To identify the primary editable objects in each video, we use a prompt that analyzes the first frame.

##### Object Identification.

You are given a single video frame. Identify the main editable object in this frame.

Rules:

- Output THREE lowercase category word only (e.g., person, car, dog, ball, cup, bottle, phone, bag, plant, flower, sign, tableware).
- Do not describe background or actions.
- If several candidates exist, choose the smallest salient object that humans often edit/remove (e.g., ball before person in sports; cup before hands on a table).

Output strictly in JSON: {"Object": "Category"}

The original caption is constructed to preserve the scene context and serve as reference when replacing the masked object for swap tasks.

##### Original Caption.

You are given a short video. Write ONE concise caption in present tense (18–30 words) describ-

ing only the stable scene elements (location, background, persistent subject categories). Rules:

- Describe what remains visually consistent across the clip.
- Use generic categories for moving actors (e.g., “a person”, “three red cups on a white table”, “a yellow taxi by a street”).
- Avoid counts unless they are constant; avoid names, brands, emotions, camera terms.
- No negations (e.g., “no/without”).
- Keep it objective, concrete, and free of text/letters.

Output strictly in JSON: “Caption”: “The video shows ...”

The removal caption describes the scene without the target object identified earlier.

##### Removal Caption.

You are given an original video caption and a target object to remove. Rewrite the caption so it naturally describes the scene as if that object never existed.

Rules:

- Output ONE fluent sentence in present tense, 35–60 words.
- Start with: “The video shows ...”
- Do NOT mention the removed object, any pronouns referring to it, or actions tied to it (e.g., holding, touching, pouring).
- Do NOT use negations like “no/without”.
- Do NOT invent new objects or text/letters that the original background did not imply.
- Keep only concrete, persistent background elements (location, surfaces, vehicles, trees, buildings, sky, lighting, colors, furniture).

Input: “Original\_Caption”: “Original\_Caption”,

Output strictly in JSON: {"Remove\_Caption": "The video shows ... "}

### Task Discrimination.

- Video A: Source Video
- Video B: Generated after editing

- Decide:

- swap: an object in A is replaced by a different object in B. The region still contains an object, but its identity changes.
- modification: the same object remains in B, but its attributes (shape, color, texture, style, size, letters, patterns, or fine details) are edited without replacing it with a different object.

If the object's identity clearly changes, classify as "swap. If only attributes or features change while the object stays the same, classify as "modification. Output strictly in JSON: {"Task": "Swap"} or {"Task": "Modification"}

For global stylization, a cinematic caption is first constructed to summarize the scene and atmosphere of the input artistic image for source video generation.

- Provide a concise scene description (environment, atmosphere, subjects).
- Do not add new motions to the subjects; keep them static as in the image.
- Focus on cinematic camera work: wide shots, dolly-in, pans, or close-ups.
- You may suggest smooth transitions or scene framing, but no new actions.
- Limit the output to 3–4 sentences.

For stylized video generation, the following prompt produces a detailed style description based on both the reference style image and the source video.



**Style Caption.** Apply style transfer using the reference image, but keep the output cinematic and natural. Rules:

- **Style Control:** Use soft, balanced colors with reduced saturation, no overexposure.
- **Subject Preservation:** Preserve the subject's natural tones and details (do not oversaturate).
- **Lighting & Texture:** Maintain subtle textures, soft lighting, and a film-like atmosphere.
- **Constraints:** Avoid harsh highlights, neon effects, or unnatural color shifts.

**Distribution of edited objects.** We visualize the objects selected for local editing in FFP-300K in Fig. 1. The word cloud highlights substantial diversity in the edited-object space of the local-editing subset of FFP-300K: while people and hand-held items (e.g., person, microphone, guitar, phone) are prominent, there is a wide spread of categories spanning furniture and electronics (table, chair, laptop), animals and nature (horse, tree, bird), vehicles and buildings, and many everyday objects. This long-tailed, semantically rich distribution indicates the dataset supports a broad range of local-editing scenarios, from fine-grained human-centric manipulations to structurally complex scene elements. Consequently, models trained on FFP-300K are exposed to varied object types and contexts, which helps foster robustness and generalization across diverse editing tasks.

**Distribution of video content.** To demonstrate the content diversity of FFP-300K, for each source video adopted, we extract 5 frames and ask Qwen2.5-VL to classify them into 15 predefined scenes, of which the distribution is shown in Fig. 2. The scene distribution of the local-editing subset is strongly skewed toward a few dominant contexts—Indoor



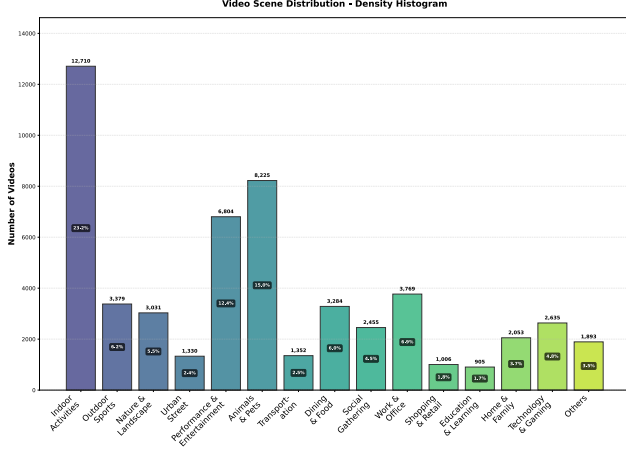


Figure 2. Scene distribution of the local editing subset of FFP-300K

Activities (12,710 videos, 29.2%), Performance & Entertainment (8,225, 19.0%) and Urban Scenes (6,904, 15.9%)—while the remaining categories (e.g., Outdoor Activities, Sports, Nature, Animals, Transport, Technology, Medical, etc.) form a long tail with individual shares typically below 8%. This composition provides dense coverage of common indoor and urban editing scenarios that are crucial for real-world applications, while still retaining broad scene diversity for generalization.

### A.3. Visualization of FFP-300K

To illustrate the visual results of FFP-300K, we provide representative examples from the two tracks of our data construction pipeline. Both tracks maintain spatial coherence and temporal consistency across all frames, enabling the model to learn strong motion priors through the first-frame propagation paradigm and supporting reliable video editing.

**Local Editing.** The local editing track constructs object-level samples using remove and swap manipulations. These samples are generated by editing specific target objects in the source video while keeping the surrounding scene unchanged, forming paired sequences that cover diverse object categories and scene contexts. As shown in Fig. 4, these examples reflect the broad coverage of fine-grained object manipulations and varied local-editing scenarios present in FFP-300K.

**Global Stylization.** The global stylization track generates full-scene style-transfer samples by applying the appearance of a reference image to the entire source video. Each source video is paired with multiple reference images, producing multiple stylized sequences that span a wide range of aesthetic styles. As illustrated in Fig. 5, these samples

expand the appearance diversity of the dataset and represent the full-scene stylization capabilities captured in FFP-300K.

## B. Additional Method Details

### B.1. Attention Head Classification Heuristic

Our proposed AST-RoPE requires pre-classification for each self-attention head. While previous methods such as SparseVidGen and Follow-your-motion utilize sample-specific classification, we find that the category of each attention head is generally sample-agnostic, which is intuitively reasonable that each head learns fixed prior knowledge. Therefore we design a simple classification strategy as follows.

**Grid-based Partitioning of the Attention Map.** For a given self-attention head and an input video with  $F$  frames, each of resolution  $H \times W$ , the total number of tokens is  $N = F' \times H' \times W'$ . The attention map is a matrix  $\mathbf{A} \in \mathbb{R}^{N \times N}$ . We conceptually partition this large matrix into a  $F' \times F'$  grid of smaller sub-matrices. Each sub-matrix  $\mathbf{A}_{ij}$  represents the attention from all tokens in the source latent frame  $i$  to all tokens in the target latent frame  $j$ .

**Quantifying Attention Density.** We measure the “activity” within each grid by calculating its attention density. The attention density  $\rho_{ij}$  for a grid  $\mathbf{A}_{ij}$  is defined as the proportion of its elements that are non-zero. In practice, due to the softmax function, all attention scores are positive. We therefore define density as the proportion of attention scores exceeding a small threshold  $\epsilon$  (e.g.,  $\epsilon = 10^{-6}$ ) to filter out negligible floating-point values.

$$\rho_{ij} = \frac{1}{H \times W \times H \times W} \sum_{u=1}^{HW} \sum_{v=1}^{HW} \mathbb{I}(\mathbf{A}_{ij}[u, v] > \epsilon) \quad (1)$$

where  $\mathbb{I}(\cdot)$  is the indicator function.

**The Classification Rule.** Our heuristic compares the strongest temporal signal against the weakest spatial signal. Let  $\mathcal{D}_{\text{diag}} = \{\rho_{ii} \mid i \in [1, F']\}$  be the set of densities for all diagonal (spatial) grids, and  $\mathcal{D}_{\text{non-diag}} = \{\rho_{ij} \mid i, j \in [1, F'], i \neq j\}$  be the set for all non-diagonal (temporal) grids.

An attention head is classified as **temporal** if its maximum non-diagonal attention density is greater than its minimum diagonal attention density. Otherwise, it is classified as **spatial**.

$$\text{Head Type} = \begin{cases} \text{Temporal} & \text{if } \max(\mathcal{D}_{\text{non-diag}}) > \min(\mathcal{D}_{\text{diag}}) \\ \text{Spatial} & \text{otherwise} \end{cases} \quad (2)$$

The intuition is that for a head to be genuinely temporal, its cross-frame attention must be meaningful and stronger than its most diffuse, weakest intra-frame attention. A head that only pays weak, noisy attention across frames but strong attention within frames will be correctly classified as spatial.

**Final Classification via Majority Voting.** The behavior of an attention head can be content-dependent. To obtain a stable and generalizable classification, we do not rely on a single video sample. Instead, we apply the classification process described above to a set of **10 diverse video samples** randomly drawn from our validation set. The final, definitive classification for each attention head is determined by a **majority vote** on the outcomes from these 10 samples. This aggregation ensures that the assigned role reflects the head’s typical behavior rather than an artifact of a specific input.

## C. Additional Experiment Results

### C.1. Additional Analysis

To help better understand the efficacy of our proposed dataset, we further conduct several experiments as follows:

**Symmetric training strategy.** While one may question whether using generated data for training can lead to high-quality results, the local editing subset of FFP-300K can actually be flipped and randomly select either the real or generated video as the target, which is adopted by us in the main experiments. We compare such a strategy with the model trained with the single-side variant which only adopts generated video as target. As shown in Tab. 1 and Fig. 3(3), symmetric training (**Ours-33f**) outperforms the single-side variant (**Ours-33f-SingleSide**), confirming that mixing real targets improves robustness.

**More Robust evaluation metrics.** We add **Flow-LPIPS** and **fg/bg VLM scores** for more robust evaluation. Specifically, Flow-LPIPS leverages the optical flow of source video to warp the edited first frame, and calculate the LPIPS between the warped videos and the generated results. Fg/bg VLM scores use SAM3 masks, decomposing original VLM prompts into editing accuracy/quality for Fg. and preservation for Bg. Conclusions hold that ours remains best.

**Application of FFP-300K to other models.** We fine-tune two FFP methods (Señorita, VACE) on FFP-300K. Tab. 1 and Fig. 3(1) show consistent gains over the originals, validating the dataset. Our model is still best, supporting the necessity of our structure design.

### C.2. Experiments on UNICBench

As a supplement to the experiment in the main paper, we further conduct experiments on UNICBench [4], which is filtered by us with the same principle as for EditVerseBench to delete samples that are not suitable for FFP. The whole test set contains 128 videos, covering tasks of add, delete, change and stylization. We adopt UNIC [4], AnyV2V [2], LucyEdit [3] and Señorita [5] as baseline methods, among which the results of UNIC and AnyV2V are provided by UNIC, and results of the other two methods are produced by us. We adopt the same metrics as EditVerseBench, which are presented in Tab. 2. Our method receives the best performance in terms of all metrics. The qualitative comparison is shown in Fig. 9, which further demonstrates that our method is not only more accurate for editing but also visually better.

### C.3. More Results on EditVerseBench

As a complement to the visual examples in the main paper, we provide additional visualization results on EditVerseBench to offer a broader view of the editing results produced by our method. Among these results is a full-task visualization that shows all four main editing tasks—add, remove, change, and stylization—together with the corresponding source video, as shown in Fig. 6. In addition, we include two orientation-specific visualizations: one for landscape orientation, as presented in Fig. 7, and one for portrait orientation, as illustrated in Fig. 8. Each visualization compares the edited videos with its corresponding source video and serves as a supplementary demonstration of our method’s editing results under different video orientations.

### C.4. More Results on UNICBench

We provide additional visualization results on UNICBench to present the editing results of our method together with the source video and UNIC under the FFP-based video editing paradigm, as shown in Fig. 10. This example offers a direct visual comparison of the editing results produced by our method and UNIC. We also include a mixed visualization that incorporates cases for which UNIC does not provide FFP-based video editing outputs, as presented in Fig. 11. In these cases, we present the instruction-based outputs from UNIC alongside our FFP-Based results to provide a broader visual reference across the different video editing types.

Method	Temporal Consistency			Text Alignment		Video Quality Pick Score $\uparrow$	VLM Evaluation		
	CLIP $\uparrow$	DINO $\uparrow$	Flow-LPIPS $\downarrow$	Frame $\uparrow$	Video $\uparrow$		Fg. $\uparrow$	Bg. $\uparrow$	Overall $\uparrow$
InsV2V	0.972	0.969	0.661	25.923	23.092	19.611	2.023	1.086	5.252
LucyEdit	0.985	0.984	0.711	26.398	23.491	19.611	2.309	1.529	5.678
EditVerse	0.986	0.986	0.662	27.776	25.293	20.132	3.148	1.513	7.104
Aleph	0.989	0.984	0.674	28.087	24.837	20.291	3.098	1.035	7.154
VACE	0.990	0.989	0.679	27.169	24.188	20.095	2.565	0.616	6.072
VACE-FT	0.989	0.989	0.671	27.341	25.323	20.227	2.841	1.255	6.660
Señorita	0.989	0.987	0.662	27.754	24.657	19.913	2.767	1.300	7.341
Señorita-FT	0.989	0.987	<b>0.653</b>	27.874	25.453	20.384	3.000	1.390	7.433
Ours-33f-SingleSide	0.990	0.990	0.667	28.166	25.508	20.481	3.330	1.411	7.308
Ours-33f	<b>0.991</b>	0.990	0.657	28.293	25.398	<b>20.419</b>	3.433	<b>1.592</b>	<b>7.631</b>
Ours-81f	<b>0.991</b>	<b>0.991</b>	0.656	<b>28.316</b>	<b>25.925</b>	20.405	<b>3.442</b>	1.573	7.600

Table 1. Results with finetuned baselines and new metrics.

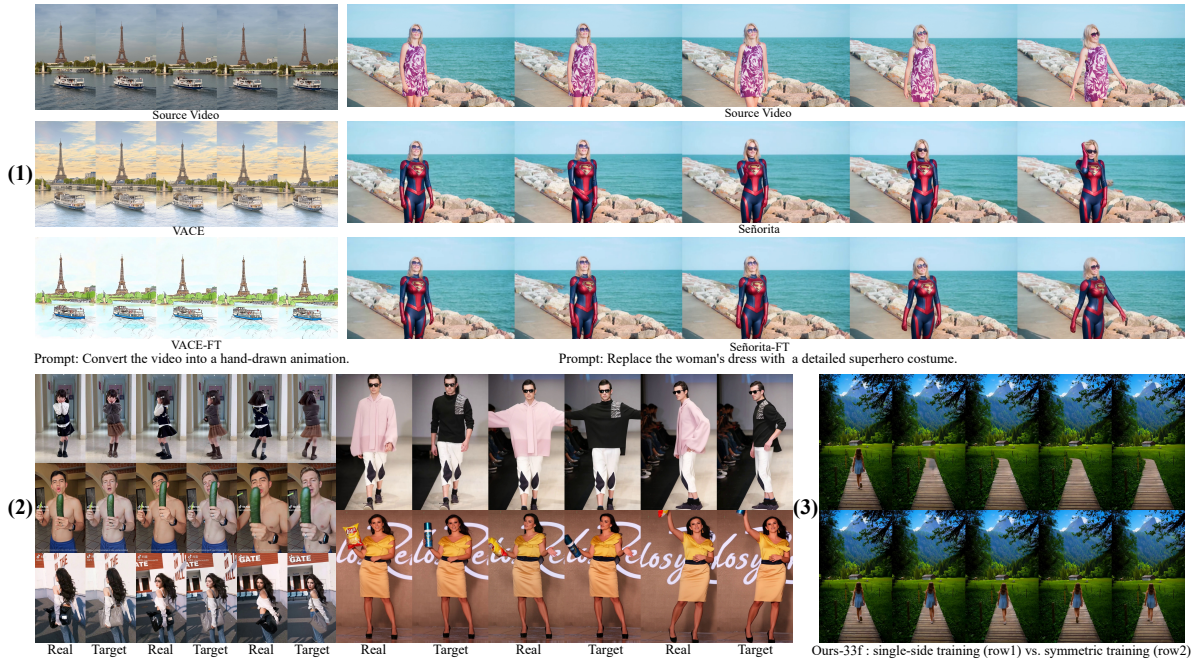


Figure 3. (1) Results of Senorita/VACE and their finetuned version on FFP-300K. (2) Real-world application of our model. (3) Single-side training v.s. symmetric training.

Type	Temporal Consistency		Text Alignment		Video Quality Pick Score $\uparrow$	VLM Evaluation VLM Score $\uparrow$
	CLIP $\uparrow$	DINO $\uparrow$	Frame $\uparrow$	Video $\uparrow$		
AnyV2V	0.941	0.92	23.597	20.138	19.864	4.132
LucyEdit	0.978	0.977	22.171	18.036	19.612	5.065
Senorita	0.985	0.981	24.197	20.273	19.950	6.648
UNIC	0.980	0.973	24.267	20.116	19.182	5.203
<b>Ours</b>	<b>0.986</b>	<b>0.982</b>	<b>24.879</b>	<b>20.733</b>	<b>19.951</b>	<b>6.672</b>

Table 2. **Quantitative comparison.** We compared three types of video editing methods on UNICBench. The best results are highlighted in **bold**.



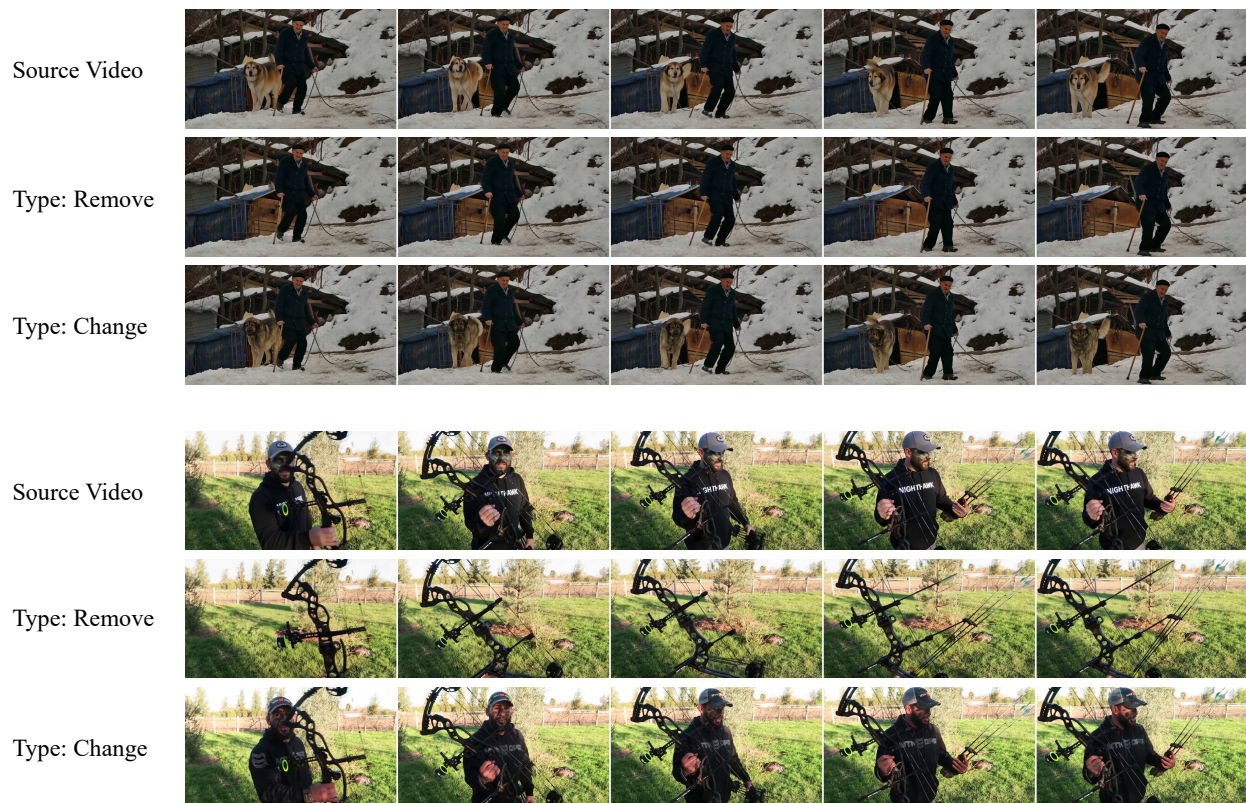


Figure 4. The visualization of local editing track in FFP-300K.



Figure 5. The visualization of global stylization track in FFP-300K.



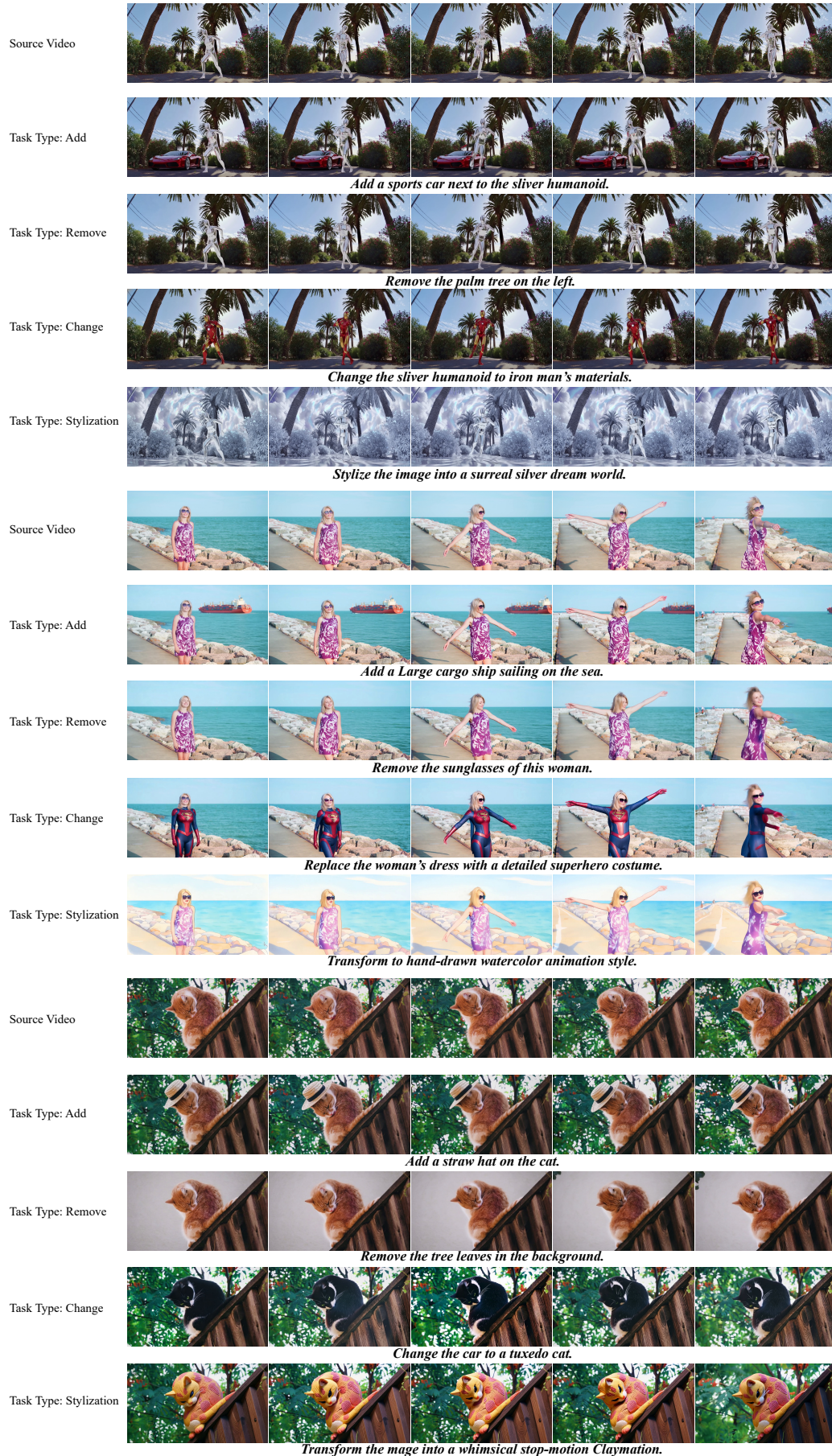


Figure 6. More results of local editing and global stylization tasks on EditVerseBench.



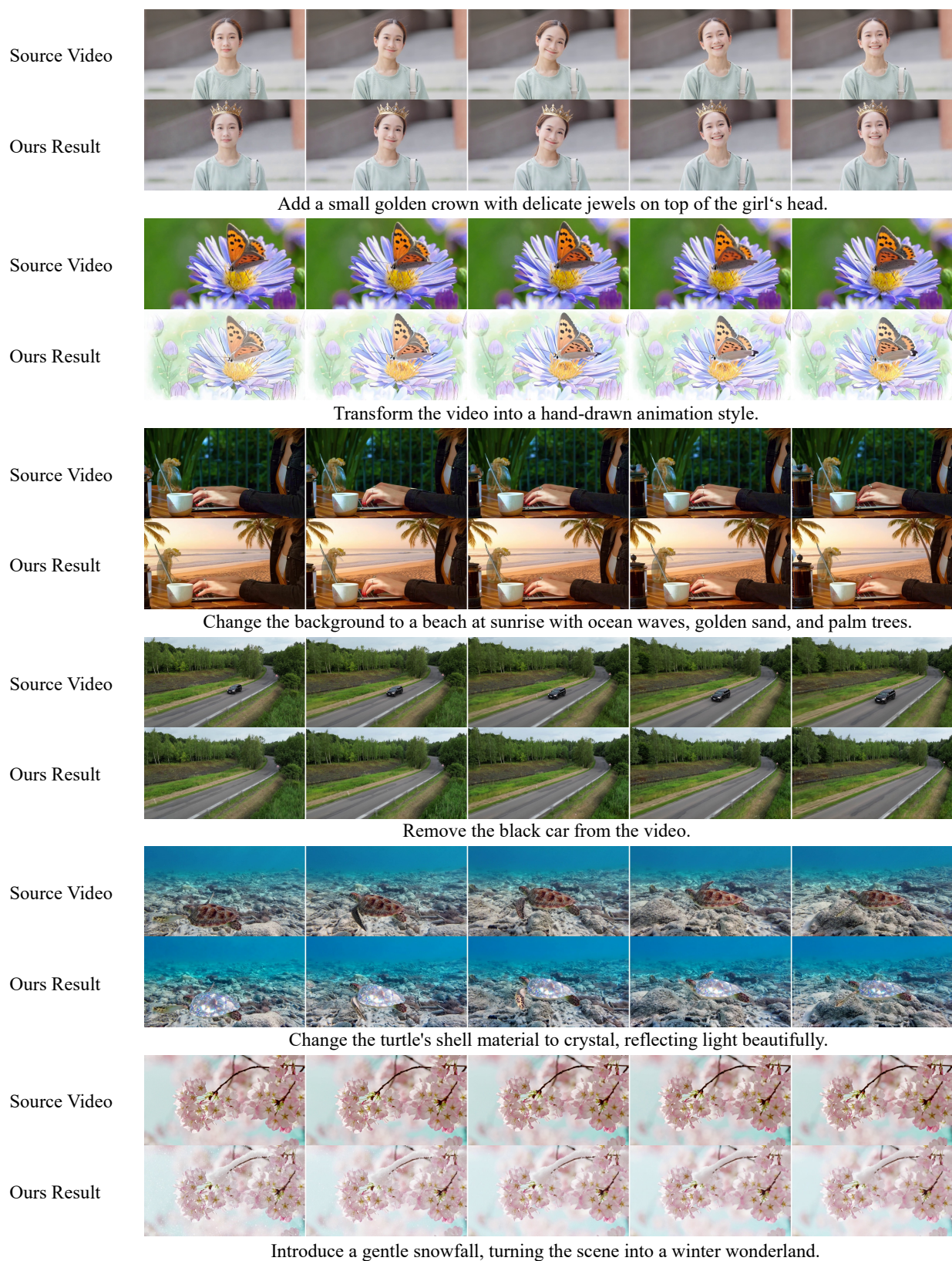


Figure 7. More visual results in the landscape orientation on EditVerseBench.



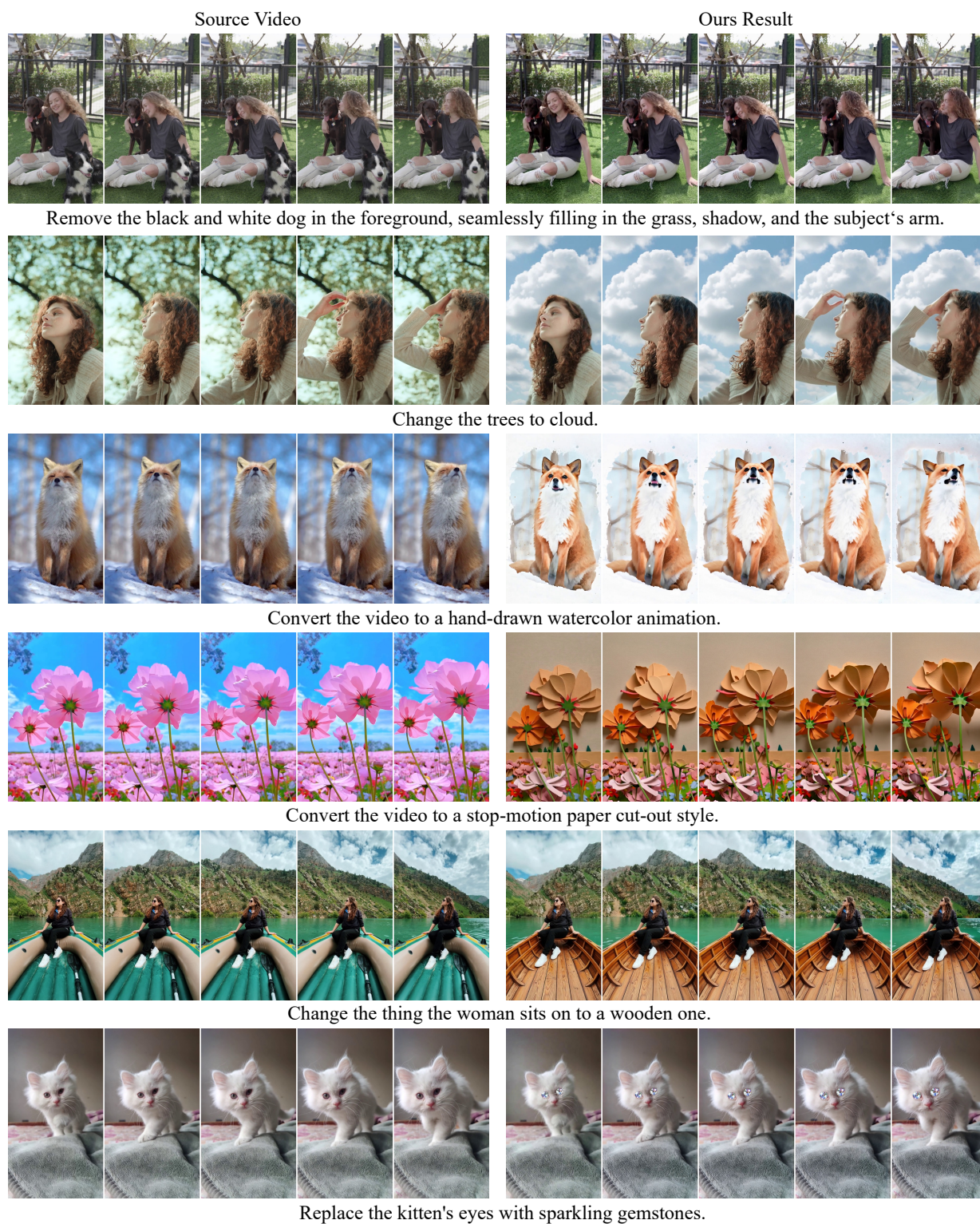
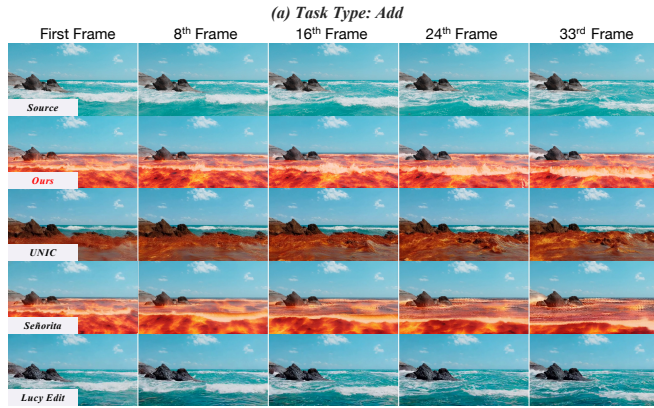
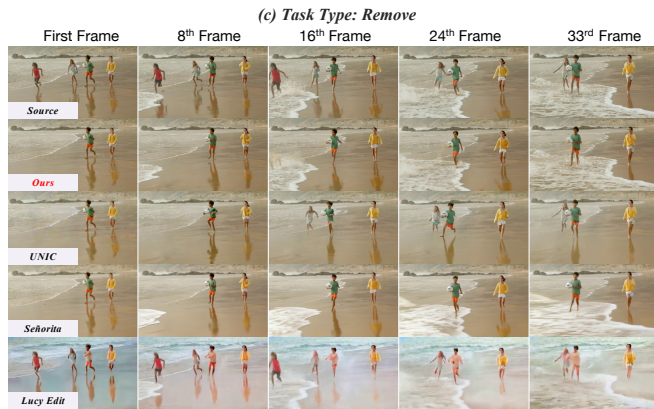


Figure 8. More visual results in the portrait orientation on EditVerseBench.

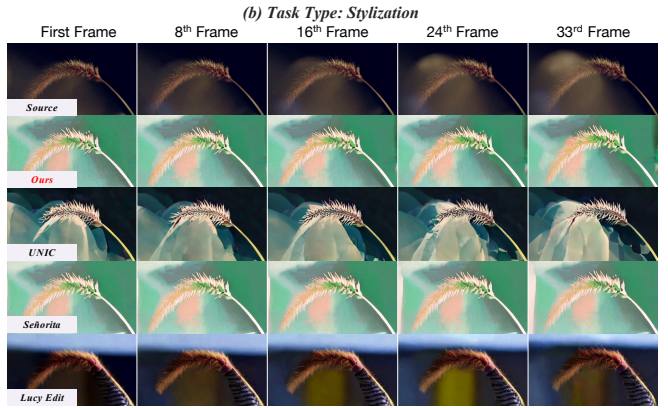




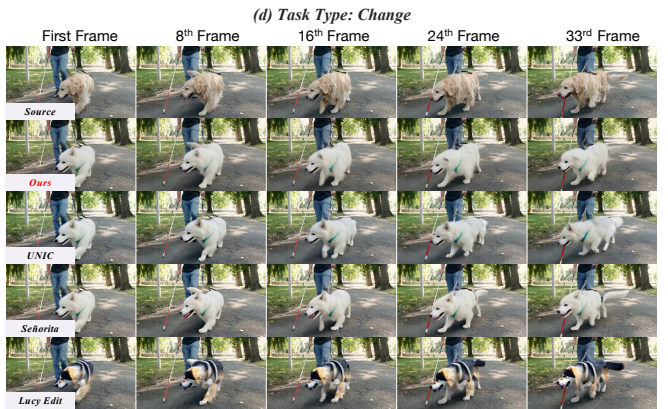
A dramatic seascape where the ocean surface is red and appears to be engulfed in flames.



The video depicts two individuals, a young boy and a girl, running along a sandy beach.



A single blade of grass with the sunlight shining on it.



A man in casual attire, including a black t-shirt, blue jeans, and sneakers, is seen walking on a paved path in a park with his Samoyed guide dog.

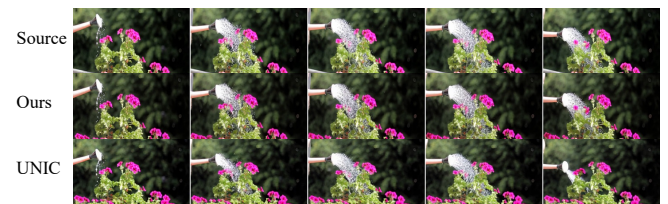
Figure 9. **Qualitative Comparison.** We choose top three methods in quantitative comparison to compare with our-33f visual results across local editing and global stylization tasks.



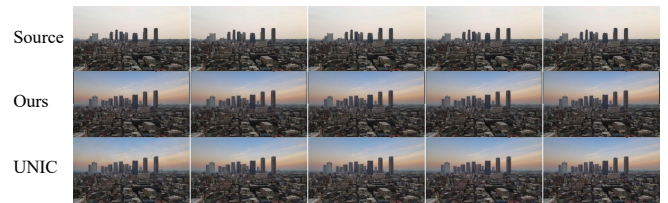
Remove the monitor on the left side.



Remove the dog from the scene.



Remove the fence.



Change the weather and sky to a sunset scene.

Figure 10. Visualization results of our method and UNIC on UNICBench for FFP-based video editing.

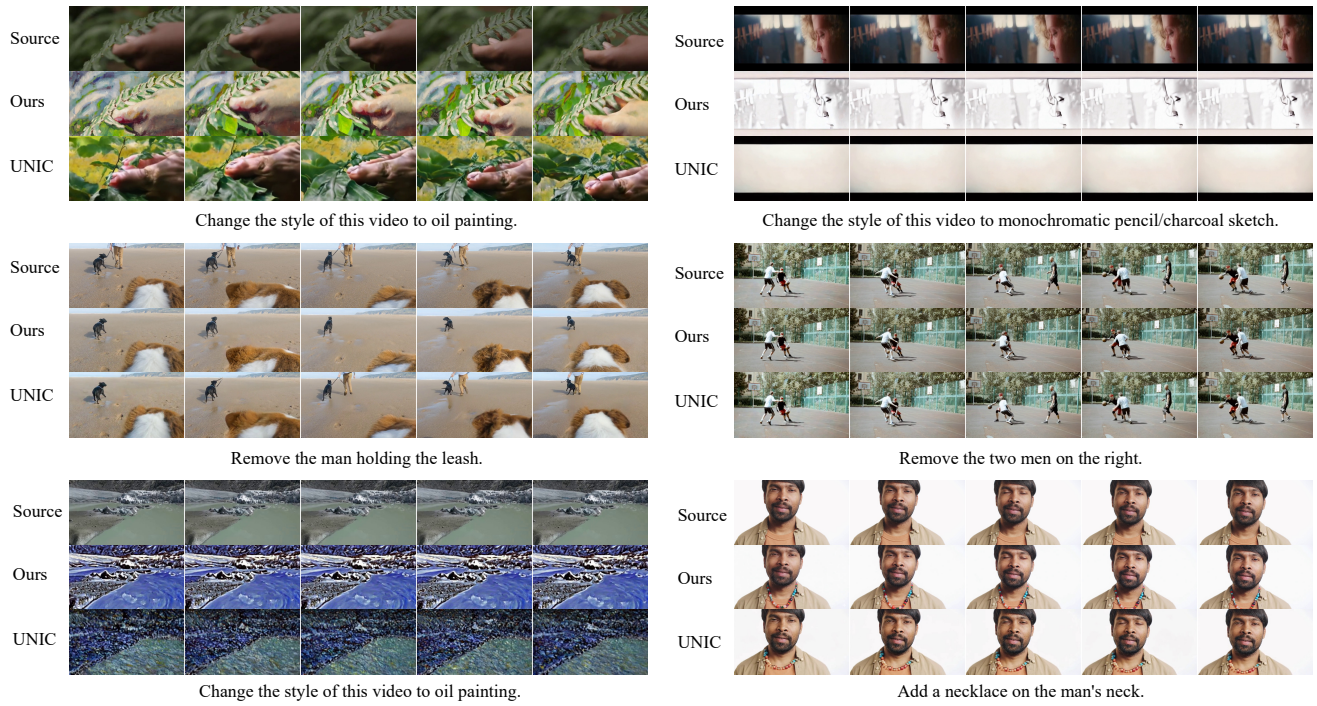


Figure 11. Mixed visualization results of our method and UNIC on UNICBench.

## References

- [1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. [arXiv preprint arXiv:2502.13923](#), 2025. 1
- [2] Max Ku, Cong Wei, Weiming Ren, Huan Yang, and Wenhua Chen. Anyv2v: A plug-and-play framework for any videot-to-video editing tasks. [arXiv preprint arXiv:2403.14468](#), 2(3):5, 2024. 4
- [3] Decart AI Team. Lucy-edit: Open-weight text-guided video editing. Technical report, Lucy-Edit Team, 2024. Accessed: 2025-10-28. 4
- [4] Zixuan Ye, Xuanhua He, Quande Liu, Qiulin Wang, Xintao Wang, Pengfei Wan, Di Zhang, Kun Gai, Qifeng Chen, and Wenhua Luo. Unic: Unified in-context video editing, 2025. 4
- [5] Bojia Zi, Penghui Ruan, Marco Chen, Xianbiao Qi, Shaozhe Hao, Shihao Zhao, Youze Huang, Bin Liang, Rong Xiao, and Kam-Fai Wong. Señorita-2m: A high-quality instruction-based dataset for general video editing by video specialists, 2025. 4