RFDM: Residual Flow Diffusion Models for Video Editing
Supplementary Material

6. Experiments

6.1. Señorita Benchmark Details

Table 3 reports the full quantitative results on the Señorita benchmark for our method, along with Fairy [44] and VidToMe [28], including per-task breakdown. We report the average scores, whereas here we show the complete tables for global style transfer, local style transfer, and object removal.

Overall, both variants of `RFDM` consistently improve over prior work across the majority of metrics, while maintaining orders of magnitude lower computational cost as shown in the paper Table 1. In particular, `RFDM3.5` attains higher or comparable scores in both temporal consistency and faithfulness metrics, which aligns with our design choices of residual flow prediction and autoregressive conditioning. For completeness, we also include the corresponding qualitative videos and error plots in the accompanying HTML visualization page (project page linked in the supplementary material).

7. More qualitative results

In addition to the figures shown in the main paper, we provide a larger set of qualitative examples (149 videos) in the html page provided with this supplementary material. We kindly encourage the reviewer to open the page for more comprehensive qualitative comparison.

7.1. Failure cases

We summarize common failure cases of `RFDM`, with more detailed explanations and video examples provided in the HTML page attached to the supplementary material. We encourage the reviewer to refer to the HTML visualization page for detailed explanations and examples.

7.2. Better than ground-truth

We also observe several cases where `RFDM` produces results that appear more natural than the provided ground truth. Since the Señorita dataset's ground truth is generated through a multi-stage pipeline combining segmentation, inpainting, and tracking modules, it can occasionally contain artifacts, flickering, or imperfect boundaries. Benefiting from the strong spatial understanding of stable diffusion models and the temporal stability introduced by our proposed framework, `RFDM` often yields smoother and more coherent outputs that correct such imperfections. We encourage the reviewer to refer to the HTML visualization page in the supplementary material for side-by-side video comparisons.

7.3. Visualization of ablation studies

We visualize what is measured in our ablation studies through **error accumulation** and **temporal consistency** in the attached HTML page. As shown, when $\Delta = 0$, error accumulation remains low but temporal consistency degrades over time. At $\Delta = 1$, temporal consistency improves, though accumulated error increases. The balance is achieved at $\Delta = 3$, where both metrics remain stable. We encourage the reviewer to refer to the HTML visualization page in the supplementary material for corresponding video examples and plots.

8. MLLM-as-a-Judge

To obtain a human-aligned and instruction-aware evaluation of video editing quality, we employ a multimodal LLM-based scoring mechanism, referred to as *MLLM-as-a-Judge*. This evaluator uses GPT-4o [16] to assess how well an edited video satisfies the user instruction while preserving the structure and semantics of the original content.

Evaluation protocol. For each video, we consider the input sequence x, the predicted edited video $\bar{y}_t^0$, and the ground-truth edit y_t^0. We uniformly sample K frames from each sequence to form triplets

$$\left(x_k,\, \bar{y}_k^0,\, y_k^0\right), \qquad k = 1, \ldots, K.$$

Each triplet is processed independently to obtain frame-level judgments.

Prompting the MLLM. We provide the multimodal LLM with three images—the original frame, the edited frame produced by the method, and the ground-truth edited frame—along with the editing instruction. The model is asked to judge, in a comparative manner, how well the candidate edit adheres to the instruction while maintaining the spatial layout and identity of the original scene.

We use the following fixed prompt template:

```
You are an expert judge for
video editing quality.  You are
given:  (1) the original frame,
(2) a candidate edited frame,
and (3) the ground-truth edited
frame, along with the editing
instruction:  ``{p}''.  Rate how
well the candidate frame follows
the instruction while preserving
the content, structure, and visual
coherence of the original frame.
```

Table 3. Señorita results are averaged across style transfer, local style transfer, and object removal tasks. TempCon denotes temporal consistency. * marks methods using our pretrained SD1.5 UNet as the backbone.

Task	Method	DVS↑	MLLM-JUDGE↑	TEMPCON↓	VIDREAMSIM↓
Style transfer	Fairy [44]	0.49	2.5	0.045	0.77
	Fairy*	0.56	3.7	0.048	0.50
	VidToMe [28]	0.51	3.5	**0.007**	0.60
	VidToMe*	0.57	1.5	0.015	0.51
	`RFDM1.5`	<u>0.59</u>	<u>6.0</u>	**0.007**	<u>0.43</u>
	`RFDM3.5`	**0.63**	**7.5**	<u>0.008</u>	**0.36**
Local editing	Fairy [44]	0.21	3.0	0.037	0.65
	Fairy*	0.29	4.0	0.037	0.17
	VidToMe [28]	0.22	2.8	**0.007**	0.35
	VidToMe*	0.26	1.5	0.015	0.61
	`RFDM1.5`	<u>0.34</u>	<u>6.3</u>	0.012	<u>0.13</u>
	`RFDM3.5`	**0.40**	**6.8**	<u>0.010</u>	**0.12**
Object removal	Fairy [44]	0.26	3.1	0.046	0.46
	Fairy*	0.35	4.2	0.042	0.19
	VidToMe [28]	0.23	3.0	**0.007**	0.43
	VidToMe*	0.29	2.3	0.012	0.65
	`RFDM1.5`	<u>0.37</u>	<u>7.5</u>	0.011	<u>0.12</u>
	`RFDM3.5`	**0.40**	**7.8**	<u>0.010</u>	**0.11**

```
Return a single score from 1 to 10,
where 1 means ``much worse than the
ground truth'' and 10 means ``better
than the ground truth''.  Output
only the number.
```

Score aggregation. For each frame k, the MLLM outputs a scalar score $s_k \in \{1, \ldots, 10\}$. The video-level score is the average across sampled frames:

$$\text{MLLM-Judge}(x, \bar{y}^0, y^0) = \frac{1}{K} \sum_{k=1}^{K} s_k.$$

Finally, the dataset-level score for a method is obtained by averaging over all videos. A higher score indicates that the edited frames are judged to:

- more faithfully follow the user-provided editing instruction,
- better preserve the identity, geometry, and temporal cues of the input,
- exhibit stronger perceptual coherence compared to the ground-truth edit.

Because the evaluation explicitly incorporates instruction text and visual context, it captures aspects of editing quality that traditional pixel- or feature-based metrics fail to measure.

The same protocol is used for all video-editing tasks considered in this work, including global style transfer, local style transfer, and object removal.