FlowPortal: Residual-Corrected Flow for Training-Free Video Relighting and Background Replacement

Supplementary Material

Appendix

This supplementary material provides additional qualitative results, methodological details, extended analyses, and user study descriptions. Specifically, we present extra visual results and comparisons in Sec. A, derive Residual-Corrected Flow from FlowEdit in Sec. B, and describe the full algorithmic pipeline in Sec. C. Extensions of our framework, including full-scene relighting and user-provided background control, are provided in Sec. D and Sec. E. We discuss limitations and failure cases in Sec. F, and the user study is demonstrated in Sec. G.

A. Addtional Results

We present additional visual results and comparison results in Fig. A and Fig. B, respectively. For video demonstrations, please refer to our project page: https://gaowenshuo.github.io/LINR-bridge/.

Generalization across Backbones. To validate the generalizability of our approach, we evaluate its performance on different video generation backbones, for example Wan2.2 and HunyuanVideo. Results are presented in Table A. As shown, our method maintains robust performance across different architectures.

Table A. Quantitative comparison across backbones.

Backbone	CLIP-T	Structural Cons.
Wan2.1	0.3326	0.8804
Wan2.2	**0.3346**	**0.8811**
Hunyuan	0.3287	0.8807

Analysis of Residual Reusing. For the residual reusing strategy, we present a time-quality curve with respect to the reuse step r to analyze the trade-off between inference speed and generation quality. The detailed metrics are summarized in Table B. The results indicate that increasing the reuse step r significantly reduces inference time with minimal impact on quality.

Table B. Time-quality curve on reusing.

r	CLIP-T	Detail Cons.	Time (s)
1	**0.3275**	41.2871	307
5	0.3269	41.5094	193
10	0.3271	**41.7632**	181
20	0.3262	41.1985	**177**

Impact of Albedo Maps. We investigate the effectiveness of utilizing albedo maps compared to using original video frames in computing metrics. Table C reports the metrics. Incorporating albedo maps reduces the impact of differences in lighting conditions on evaluation metrics.

Table C. Effect of albedo.

	Structural Cons.	Motion Cons.	Detail Cons.
albedo map	**0.8913**	**0.8949**	**41.2044**
original video	0.8804	0.8727	40.6762

B. From FlowEdit to Residual-Corrected Flow

As shown in Sec. 3.6 of our main paper, our method is partially inspired by FlowEdit [3]. We reformulate it into a more interpretable residual form and show that this residual inherently encodes the detailed information required for reconstruction that makes it particularly suitable for relighting tasks. Moreover, reinterpretation naturally enables two additional benefits: residual reuse for acceleration and background separation for fine-grained control.

Specifically, assuming FlowEdit use the same global noise ϵ at every timestep:

$$V_t^{\text{FE-edit}}(z_t^{\text{FE-edit}}) = V_t^{\text{tar}}(z_t^{\text{FE-pred}}) - V_t^{\text{src}}(z_t), \tag{1}$$

$$z_t = (1-t)z_0 + t\epsilon, \tag{2}$$

$$z_t^{\text{FE-pred}} = z_t + z_t^{\text{FE-edit}} - z_0, \tag{3}$$

where $t \in [0,1]$ and $z_1^{\text{FE-edit}} = z_0$. Then we can derive:

$$z_1^{\text{FE-pred}} = \epsilon, \tag{4}$$

$$\begin{aligned} V_t^{\text{FE-pred}}(z_t^{\text{FE-pred}}) &= V_t(z_t) + V_t^{\text{FE-edit}}(z_t^{\text{FE-edit}}) - 0 \\ &= V_t(z_t) + V_t^{\text{tar}}(z_t^{\text{FE-pred}}) - V_t^{\text{src}}(z_t). \end{aligned} \tag{5}$$

Note that $V_t(z_t)$ is a constant velocity:

$$V_t(z_t) = \frac{z_0 - \epsilon}{1-0} = V_0, \tag{6}$$

so that

$$V_t^{\text{FE-pred}}(z_t^{\text{FE-pred}}) = V_t^{\text{tar}}(z_t^{\text{FE-pred}}) + (V_0 - V_t^{\text{src}}(z_t)), \tag{7}$$

which shows that $z_t^{\text{FE-pred}}$ can be interpreted as a generation process starting from ϵ and evolving along $V_t^{\text{FE-pred}}$. We then rename this variable as z_t^{edit} and track its trajectory. Furthermore, we observe that $V_0 - V_t^{\text{src}}(z_t)$ can be interpreted as

Figure A. Additional Results of FlowPortal. Refer to our project page for video demonstrations.

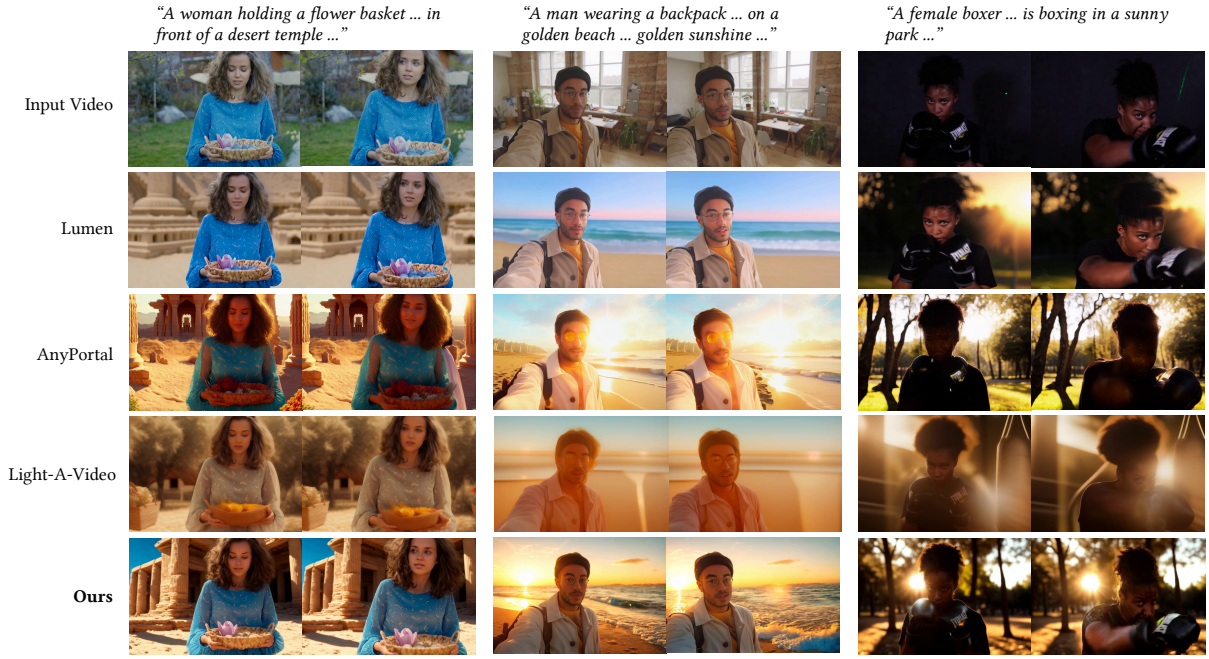

Figure B. Additional Comparison Results. Refer to our project page for video demonstrations.

a residual velocity corresponding to reconstructing z_0 using the source prompt, which we denote as:

$$V_t^{\text{res}}(z_t) = V_0 - V_t^{\text{src}}(z_t), \tag{8}$$

Finally, we obtain:

$$V_t^{\text{edit}}(z_t^{\text{edit}}) = V_t^{\text{tar}}(z_t^{\text{edit}}) + V_t^{\text{res}}(z_t), \tag{9}$$

which corresponds to our method.

Essentially, V_t^{res} is a residual velocity that maintains the structure and content of the original video, ensuring that the editing process preserves structural consistency. Its value is equivalent to the finite difference with respect to t of the difference between the original video z_0 and the result of a single-step denoising applied to z_0 after adding noise at level t. Therefore, V_t^{res} actually carries the precise structural information of the original video, making it particularly suitable for video relighting tasks.

Vanilla FlowEdit actually selects n different noise samples at each step of the generation process and averages the resulting outputs, which leads to stability but blurring in the generated results. To produce a clear background, we discard this multi-noise sampling procedure.

This reinterpretation offers the following advantages over the original FlowEdit: First, our reinterpretation allows the background to remain a purely generative process. By simply multiplying V_t^{res} with a foreground mask M, the background is prevented from being influenced by the structural information of the original video, thereby avoiding artifacts, as shown in Fig. 8 in our main paper. Second, to reduce additional computational overhead, our reinterpretation enables the structural residual information in V_t^{res} to be reused every r steps, reducing the total number of steps from $2T$ to $(1+1/r)T$ without significantly degrading generation quality, as also illustrated in Fig. 8 in our main paper.

C. Full Algorithm

In this section, we present the detailed implementation of our proposed framework. For simplicity, we omit the details regarding the latent space. The videos are mapped into the latent space through a VAE while the masks are mapped into latent space through bilinear downsampling.

C.1. Decoupled Condition Design

We first construct the condition inputs, including text prompts, reference frames, and shared structural information. To obtain the foreground mask of the input video, denoted as M, we apply BiRefNet [10] on the first frame and propagate it across the sequence using MatAnyone [6]. Given the mask, the reference frames and structural conditions are prepared as follows:

- **Reference frame for the source condition.** We directly use the first frame of the input video as the reference. Note that, in Wan2.1 [1], where the first frame is not strictly required, any frame can serve as the reference.
- **Reference frame for the target condition.** We employ IC-Light [9] to relight the selected source frame according to the target textual description. The inputs to IC-Light include the chosen video frame, its corresponding foreground mask, and the target prompt.
- **Shared structural information.** We extract Canny edges (C), HED boundaries (H), and depth maps (D) from the original video [8], and empirically combine them as $(0.25C + 0.25H + 0.5D) \cdot M$. The edge-based components (Canny, HED) preserve fine structural details, while the depth map encourage structural consistency without over-constraining the model. This structural condition is shared between both source and target conditions.

C.2. Inference Algorithm

Starting from a global Gaussian noise ϵ, we perform denoising using Wan2.1 [1, 5] integrated with our proposed method. At each timestep t, the model predicts the velocity fields V_t^{src} and V_t^{tar} under the source and target conditions, respectively. The complete inference procedure is shown in Algorithm 1.

C.3. Implementation of ControlNet of the Video Model

The structural condition component is an adapter that projects input maps to the model's hidden space using a Conv3d layer. These control features are injected via element-wise addition to the video embeddings only at the input level (pre-Transformer blocks).

D. Extension: Full Scene Relighting

In our method, removing the masking mechanism in condition preparation, Residual-Corrected Flow, and High-Frequency Transfer enables full scene relighting, allowing the model to adjust illumination across the entire scene while preserving background structure. We provide a visual comparison of full scene relighting against TC-Light [4] and Light-A-Video [11] with different backbones, in Fig. C. Notably, our approach achieves the most natural illumination while maintaining the highest level of structural consistency across the entire scene.

Figure C. **Full Scene Relighting comparison.** Our method is capable of relighting entire scene while preserving background structural coherence. Compared with TC-Light and Light-A-Video with different backbones, our approach produces the most natural illumination and preserves scene structure most faithfully.

Algorithm 1 Residual-Corrected Flow for Video Relighting

Input: Initial noise ϵ, input video z_0, timestep schedule $0 = t_0 < t_1 < \cdots < t_N = 1$, velocity field V_t^{src} and V_t^{tar}, reusing steps r, *High-Frequency Transfer* intensity λ, foreground mask M
Output: Relit video z_0^{edit}

$V_0 \leftarrow \epsilon - z_0$
$z_{t_N}^{\text{edit}} \leftarrow \epsilon$
for $i = N, N-1, \ldots, 1$ **do**
  $z_{t_i} \leftarrow (1 - t_i)z + t_i\epsilon$
  $z_{t_{i-1}} \leftarrow (1 - t_{i-1})z + t_{i-1}\epsilon$
  if $(N - i) \bmod r == 0$ **then** ▷ calculate new residual for reusing in next r steps
    $V_{t_i}^{\text{res}}(z_{t_i}) \leftarrow V_0 - V_{t_i}^{\text{src}}(z_{t_i})$
    $t_{\text{last}} \leftarrow t_i$
  end if
  $V_{t_i}^{\text{edit}}(z_{t_i}^{\text{edit}}) \leftarrow M \cdot V_{t_{\text{last}}}^{\text{res}}(z_{t_i}) + V_{t_i}^{\text{tar}}(z_{t_i}^{\text{edit}})$
  $z_{t_{i-1}}^{\text{edit}} \leftarrow z_{t_i}^{\text{edit}} + (t_i - t_{i-1})V_{t_i}^{\text{edit}}(z_{t_i}^{\text{edit}})$
  $z_{t_{i-1}}^{\text{edit}} \leftarrow \text{LF}(z_{t_{i-1}}^{\text{edit}}) + \lambda M \cdot \text{HF}(z_{t_{i-1}}) + (1 - \lambda M) \cdot \text{HF}(z_{t_{i-1}}^{\text{edit}})$
end for

E. Extension: User-Provided background

In some user scenarios, the user may wish to specify the background scene of the generated video, and our method naturally supports this type of task. During condition preparation, we input the user-specified scene image into IC-Light when generating the reference frame, enabling our method to produce a relit video whose foreground matches the lighting of the specified scene as iluustrated in Fig. D.

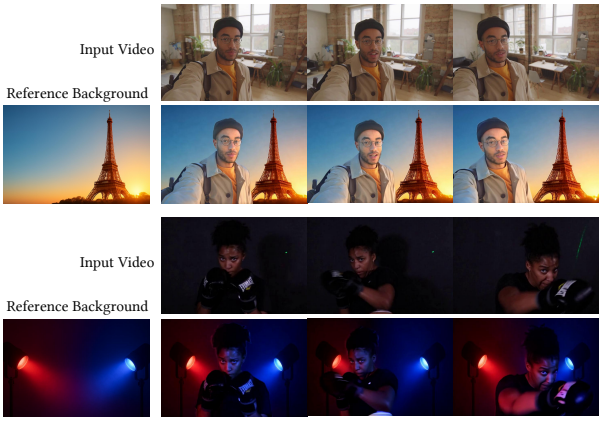

Figure D. **Scene-specified relighting.** Our method supports user-specified background scenes by feeding the target scene image into IC-Light during reference-frame preparation, enabling foreground relighting consistent with the desired scene.

F. Limitation

Although our method can handle a wide range of video relighting and background replacement scenarios, it may still be limited in extremely complex cases by the generation capabilities of the two base models, as shown in Fig. E.

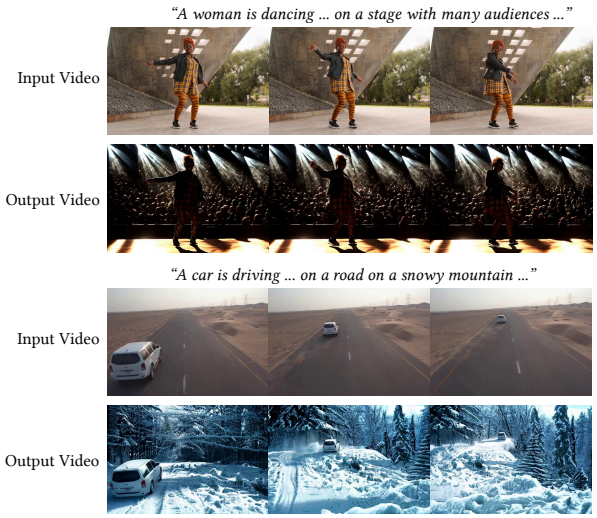

Figure E. **Failure Cases**. Although our method is capable of handling a wide range of video relighting and background replacement scenarios, it can still be constrained by the generation capacity of its two base models in extremely complex situations. In the first example, the IC-Light model fails to generate a complex backlit stage scene containing many audience, . In the second example, IC-Light produces a reference frame whose scene layout does not align with the actual driving trajectory of the vehicle. This mismatch restricting the video model's ability to produce a reasonable and structurally coherent output.

G. User Study Details

The user study in our main paper is conducted with 24 invited participants including 16 male participants and 7 females, with ages ranging from 18 to 39, who were invited to complete a questionnaire comparing our method with AnyPortal [2], Light-A-Video [11], and Lumen [7] on a dataset of randomly-selected 17 video–prompt pairs. A screenshot of the evaluation interface is shown in Fig. F.

References

[1] AIGC-Apps. Videox-fun: Github repository, 2025. GitHub repository: github.com/aigc-apps/VideoX-Fun. 3

[2] Wenshuo Gao, Xicheng Lan, and Shuai Yang. Anyportal: Zero-shot consistent video background replacement. In *ICCV*, pages 18990–18999, 2025. 5

[3] Vladimir Kulikov, Matan Kleiner, Inbar Huberman-Spiegelglas, and Tomer Michaeli. Flowedit: Inversion-free text-based editing using pre-trained flow models. In *ICCV*, pages 19721–19730, 2025. 1

[4] Yang Liu, Chuanchen Luo, Zimo Tang, Yingyan Li, Yuran Yang, Yuanyong Ning, Lue Fan, Junran Peng, and Zhaoxiang Zhang. Tc-light: Temporally consistent relighting for dynamic long videos. *arXiv preprint arXiv:2506.18904*, 2025. 3

[5] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 3

[6] Peiqing Yang, Shangchen Zhou, Jixin Zhao, Qingyi Tao, and Chen Change Loy. Matanyone: Stable video matting with consistent memory propagation. In *CVPR*, pages 7299–7308, 2025. 3

[7] Jianshu Zeng, Yuxuan Liu, Yutong Feng, Chenxuan Miao, Zixiang Gao, Jiwang Qu, Jianzhang Zhang, Bin Wang, and Kun Yuan. Lumen: Consistent video relighting and harmonious background replacement with video generative models. *arXiv preprint arXiv:2508.12945*, 2025. 5

[8] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, pages 3836–3847, 2023. 3

[9] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Scaling in-the-wild training for diffusion-based illumination harmonization and editing by imposing consistent light transport. In *ICLR*, pages 1–12, 2025. 3

[10] Peng Zheng, Dehong Gao, Deng-Ping Fan, Li Liu, Jorma Laaksonen, Wanli Ouyang, and Nicu Sebe. Bilateral reference for high-resolution dichotomous image segmentation. *arXiv preprint arXiv:2401.03407*, 2024. 3

[11] Yujie Zhou, Jiazi Bu, Pengyang Ling, Pan Zhang, Tong Wu, Qidong Huang, Jinsong Li, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, et al. Light-a-video: Training-free video relighting via progressive light fusion. *arXiv preprint arXiv:2502.08590*, 2025. 3, 5

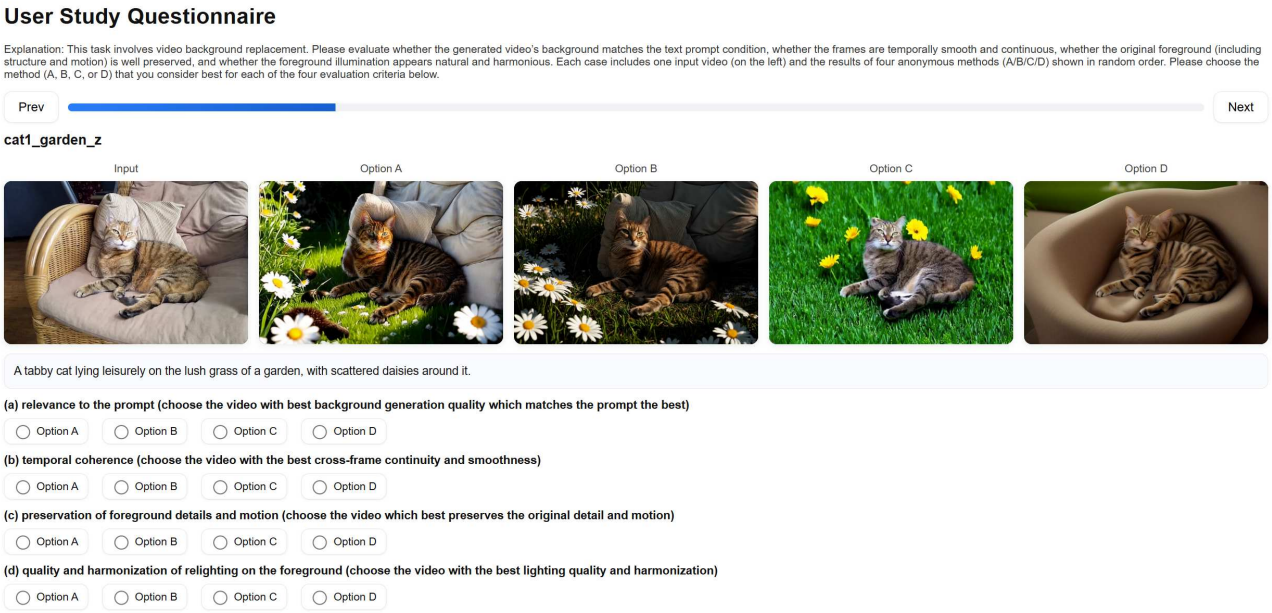

Figure F. A screenshot of the User Study interface.