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

Wenshuo Gao Junyi Fan Jiangyue Zeng Shuai Yang✉
Wangxuan Institute of Computer Technology, State Key Laboratory of Multimedia Information Processing,
Peking University, Beijing, China
{gaowenshuo, bright_fjy, zengjy}@stu.pku.edu.cn williamyang@pku.edu.cn

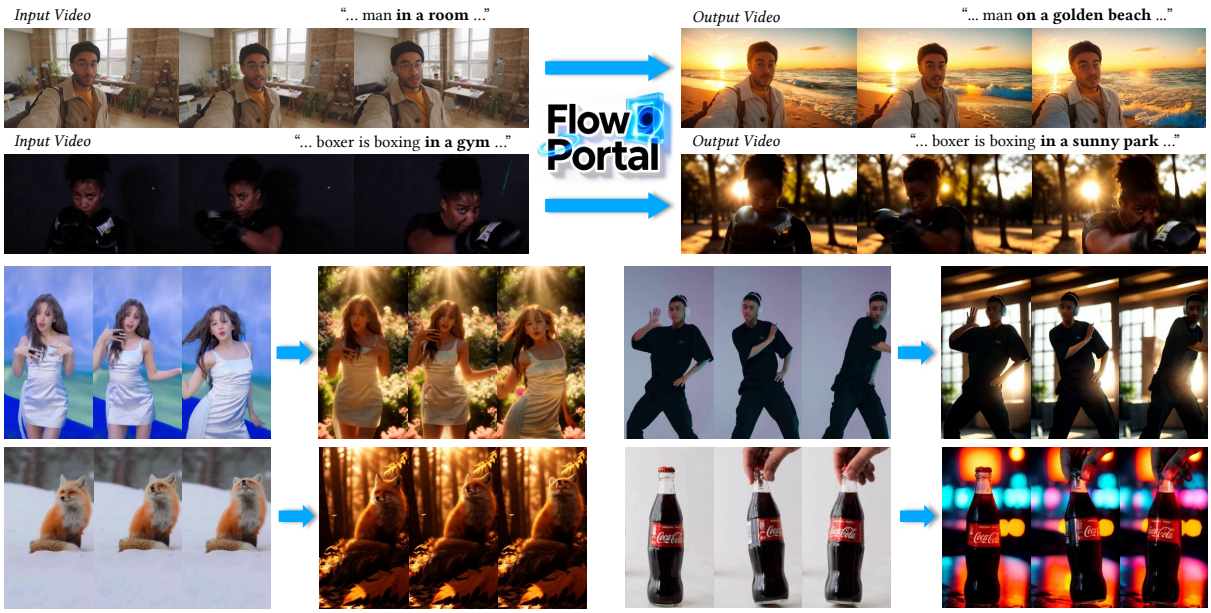


Figure 1. We propose a novel training-free **FlowPortal** framework for efficient video background replacement and foreground relighting.

Abstract

Video relighting with background replacement is a challenging task critical for applications in film production and creative media. Existing methods struggle to balance temporal consistency, spatial fidelity, and illumination naturalness. To address these issues, we introduce FlowPortal, a novel training-free flow-based video relighting framework. Our core innovation is a Residual-Corrected Flow mechanism that transforms a standard flow-based model into an editing model, guaranteeing perfect reconstruction when input conditions are identical and enabling faithful relighting when they differ, resulting in high structural consistency. This is further enhanced by a Decoupled Condition Design for precise lighting control and a High-Frequency Transfer mechanism for detail preservation. Additionally, a masking strategy isolates foreground relighting from background pure generation process. Experiments demonstrate that FlowPortal achieves superior performance in temporal coherence, structural preservation, and lighting realism, while maintaining high efficiency. Project Page: https://gaowenshuo.github.io/FlowPortalProject/.

1. Introduction

Harmonious lighting between the foreground and the background plays a crucial role in determining the realism, visual quality, and aesthetic appeal of a video. The capability to generate a new background that harmonizes with the foreground using adapted lighting, known as the video relighting task with background replacement, has attracted increasing attention due to its broad range of applications in film production, virtual photography, commercial display, and creative media editing. This technology effectively functions as a visual portal, dramatically reducing the need for on-site shooting by enabling flexible scene composition and stylized video generation. Furthermore, this capability holds significant potential for future world-to-world portal applications that seamlessly bridge different visual worlds.

Although promising, video relighting with background replacement poses significant challenges due to the need to maintain video quality, temporal consistency, and subject fidelity under new lighting conditions. Pursuing one goal often compromises others, leading to temporal flickering, loss of detail, unrealistic lighting, or prohibitive

computational cost. For example, training-based methods [4, 7, 12, 16, 18, 29] are resource-intensive and struggle with creating paired dataset and balancing between illumination richness and content fidelity. Meanwhile, training-free methods like AnyPortal [5] and Light-A-Video [36] that combine pretrained image relighting and video generation models suffer from temporal inconsistency caused by per-frame relighting as well as misalignment between input and output video due to weak condition control and limited capability for video models.

We argue that these issues arise from the absence of a cohesive framework that systematically disentangles and controls the core elements of a video: structure, motion, and illumination. To bridge this gap, we introduce a new **FlowPortal** framework built upon a condition-aware editing model through two dedicated mechanisms. First, we propose a **Residual-Corrected Flow** with **High-Frequency Transfer** for structure and detail preservation. Second, **Decoupled Condition Design** is introduced for precise lighting control. Together, these advancements provide a unified solution that addresses the long-standing challenges of video relighting: temporal continuity, spatial fidelity, illumination naturalness, and operational efficiency.

Specifically, our key insight is to exploit the principle of Condition Consistency that every change in the output is driven by a change in the target condition (*i.e.*, illumination). It implies a perfect reconstruction when there is no change in condition. Based on this principle, we introduce a Residual-Corrected Flow mechanism that rephrases a typical flow-based generation model to a novel training-free editing model. It ensures structural consistency by directing the flow towards perfect reconstruction when the target and source conditions are identical, so that subsequent changes in the conditions like illumination variation can be reflected in the video output. Our Residual-Corrected Flow allows for reusing velocity prediction across timesteps to enhance efficiency, and it processes videos holistically, avoiding per-frame processing to ensure temporal continuity.

To further strengthen the model's fidelity and controllability, we introduce three key designs: First, we propose a Decoupled Condition Design with illumination-specific and agnostic visual-textual condition inputs. This decomposition provides a stable, sufficient, and directional guidance that engineeringly enforces Condition Consistency. Second, we introduce High-Frequency Transfer, which adheres to Condition Consistency by injecting illumination-agnostic details to solidify fidelity. Third, we employ masks to isolate and precisely relight the foreground while maintaining a pure background generation process, allowing for a high-quality, contextually natural background replacement. In summary, our contributions are threefold:

- We propose a novel training-free FlowPortal framework for coherent and efficient video relighting and background replacement. By introducing Residual-Corrected Flow and Decoupled Condition Design, our framework ensures structural consistency between source and target videos, as well as illumination naturalness.
- We introduce a High-Frequency Transfer mechanism within the Residual-Corrected Flow, improving fine detail consistency under new lighting conditions.
- We propose Masked Residual-Corrected Flow and Masked High-Frequency Transfer to isolate foreground and background processing, enabling high-quality, contextually natural background replacement.

2. Related Works

Image Relighting and Background Replacement. Recent advances in generation models have led to successful image relighting methods [9, 20, 22, 27, 32]. Total Relighting [20] and SwitchLight [9] train neural networks to predict surface properties, namely normals and albedo, which are then used to recompute illumination. Relightful Harmonization [22] fine-tunes an image diffusion model conditioned on the background to achieve foreground relighting. As the current state-of-the-art method in image relighting, IC-Light [32] simultaneously performs background replacement and illumination harmonization by concatenating the input noise with a foreground condition before feeding it into a diffusion model trained with a light transport consistency objective to ensure a coherent composite.

Relighting in Video. With the maturation of image relighting techniques, researchers have begun to focus on the task of video relighting [4–6, 30, 36]. Training-based methods such as IllumiCraft [12], UniRelight [7], RelightVid [4], Lumen [29], Lux Post Facto [18], TC-Light [16], LumiSculpt [33] and UniLumos [14], perform video relighting by constructing datasets and training video-conditional diffusion models. However, collecting large-scale high-quality paired video relighting datasets is difficult and requires high costs for training. Moreover, they struggle to balance between the lighting conditions and the video foreground conditions. As a result, the trained models tend to produce lighting results with insufficient richness and diversity in illumination, and lacking fidelity under complex conditions.

Some works pursue training-free approaches for efficiency. AnyPortal [5] and Light-A-Video [36] both adopt training-free (zero-shot) strategies, combining the capabilities of IC-Light and video diffusion generation models during inference. AnyPortal introduces cross-frame attention in IC-Light and applies a Refinement Projection Algorithm in the video model to ensure consistency, while Light-A-Video uses Consistent Light Attention in IC-Light and employs a Progressive Light Fusion strategy in the video diffusion process to maintain lighting generation and consistency. However, in these two training-free methods, IC-Light is still used to relight per frame individually, intro-

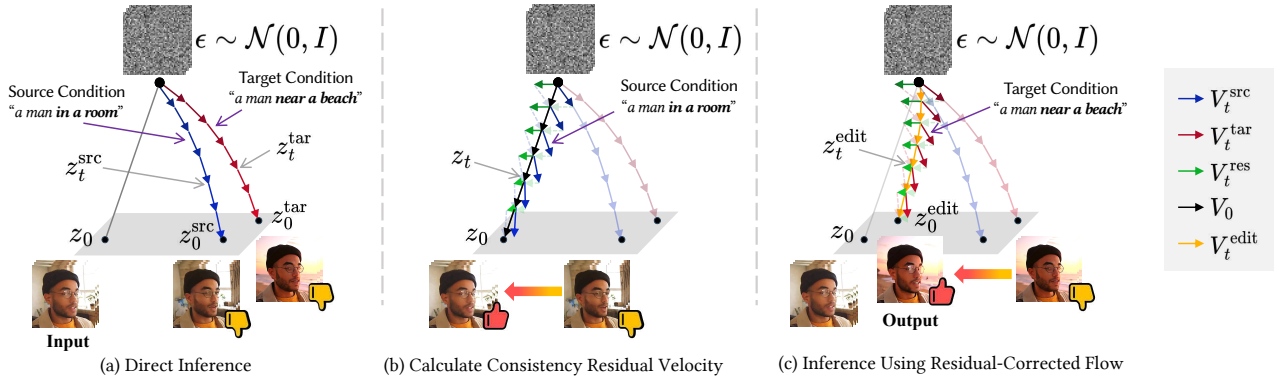


Figure 2. **Illustration of the proposed Residual-Corrected Flow.** (a) The **Naive Edit Flow** builds denoising trajectories under source and target conditions using the same noise ϵ. When applied to a real input video z_0, the mismatch between z_0^{src} and z_0 violates Stability under Identity. (b) **Consistency Residual Velocity** V_t^{res} is constructed as the difference between the ideal restoration path V_0 and the predicted source flow V_t^{src}, aligning the generated z_0^{src} with z_0. (c) **Residual-Corrected Flow** combines V_t^{tar} and V_t^{res} to perform reliable video relighting that preserves identity consistency (*e.g.*, mouth shape, glasses reflection) while enabling directional condition change. The purple arrows indicate the condition to guide the velocity calculation. For simplicity, we omit the reference frame and structural conditions.

ducing frame discontinuities that cannot be fully corrected by the video model. In addition, due to the lack of sufficient control in the video model, the relighting results may exhibit inconsistencies in structure and motion compared to the original video. Overcomplicated pipelines also make these methods less inference-efficient.

Our method is also training-free. Rather than per-frame IC-Light processing, we adopt holistic relighting to achieve better temporal consistency. Moreover, by introducing the proposed Residual-Corrected Flow and High-Frequency Transfer, we better address structural and detail consistency. To achieve stronger conditional control, we propose a Decoupled Condition Design to enhance the model's ability to respond to given conditions. We further achieve higher efficiency by reusing residual predictions for acceleration.

3. Method

3.1. Preliminary: Flow Models

In flow-based generative models [10, 13, 15, 25], data generation process is modeled as a learned continuous flow that gradually transforms a random noise into a data sample.

Formally, let z_t denote the latent variable at timestep (or noise level) $t \in [0, 1]$, where $t = 1$ corresponds to the starting point of the generation process with $z_1 \sim \mathcal{N}(0, I)$, and $t = 0$ indicates the final clean data sample z_0. Let c represent the conditional information. The model F_θ predicts a velocity (or flow) vector field V_t^c at each noise level t:

$$V_t^c(z_t) = F_\theta(z_t, t, c). \tag{1}$$

V_t^c therefore defines how the latent variable evolves continuously from noise to data over time. This generation process follows an ordinary differential equation (ODE) form:

$$\frac{dz_t}{dt} = V_t^c(z_t). \tag{2}$$

In practice, this continuous process is discretized into $N+1$ steps $\{t_0, t_1, \ldots, t_N\}$, with $0 = t_0 < t_1 < \cdots < t_N = 1$. Starting from an initial latent $z_{t_N} = z_1 \sim \mathcal{N}(0, I)$, the model iteratively applies the discrete update rule

$$z_{t_{i-1}} = z_{t_i} + (t_i - t_{i-1})\, V_{t_i}^c(z_{t_i}), \tag{3}$$

for N steps from t_N to t_0, progressively refining the latent until it reaches $z_{t_0} = z_0$ that lies in the data manifold.

3.2. Condition Consistency

Condition Consistency describes a fundamental property of an ideal editing model: changes in the output are driven by changes in the input conditions. It suggests two core principles: **1)** When the editing condition changes, a corresponding and perceivable change should manifest in the output; **2)** When the editing condition remains identical, the output should be a faithful replica of the input. This principle is critically important in the specific context of video relighting. Specifically, it requires that

- Directional Change: The generated video should differ from the input only in lighting if the conditions differ only in lighting. All other aspects, including object structure and motion dynamics, must be perfectly preserved.
- Stability under Identity: In the extreme case where the target and source conditions are identical, an ideal model must reproduce the source video exactly.

It guarantees a faithful mapping of the condition signal, which neither omits required changes nor introduces extraneous ones, thereby ensuring predictable outcomes and high visual fidelity. To this end, we aim to design a reliable video relighting model enforced with Condition Consistency.

3.3. Residual-Corrected Flow with Decoupled Condition for Reliable Video Relighting

Naive Edit Flow. We begin with a simple edit flow. We first sample a fixed Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$ and build two denoising trajectories over ϵ under the source condition "src" (*e.g.*, a prompt describing the original video) and the target condition "tar" (*e.g.*, a prompt describing the target edited video) following Eq. (1)-(3). This results in two distinct flows denoted as V_t^{src} and V_t^{tar}, respectively. The noise variable evolves along the flows from $t = 1$ to 0, producing two outputs: z_0^{src} and z_0^{tar}, as shown in Fig. 2(a). Since two flows share the same ϵ, if "tar"="src", then $z_0^{\text{tar}} = z_0^{\text{src}}$. Clearly, the editing from a synthetic video z_0^{src} to z_0^{tar} satisfies the property of "Stability under Identity". However, issues arise when applying this edit flow to a real input video z_0. Due to inherent randomness, model capability limitations, and insufficient information in condition "src", the generated video z_0^{src} often does not exactly match z_0. This violates Stability under Identity: when "tar"="src", the output $z_0^{\text{tar}} \neq z_0$.

Residual-Corrected Flow. Based on the above analysis, our key idea is to simultaneously adjust both z_0^{src} and z_0^{tar}, pulling z_0^{src} to be identical to z_0, so that the adjusted z_0^{tar}, which we denote as z_0^{edit}, can reflect directional change to z_0. First, we construct a restoration path from the noise ϵ to the source video z_0. This path corresponds to a velocity defined as:

$$V_0 = \frac{z_0 - \epsilon}{1 - 0}. \tag{4}$$

If the denoising process strictly follows this velocity, the noise ϵ will be exactly transformed back into z_0, and the intermediate result z_t at timestep t satisfies

$$z_t = (1 - t)z_0 + t\epsilon. \tag{5}$$

However, in practice, the model predicts a velocity $V_t^{\text{src}}(z_t) \neq V_0$ at z_t. To align this predicted flow with the ideal restoration path, we build a **Consistency Residual Velocity** V_t^{res} at each timestep, defined as:

$$V_t^{\text{res}}(z_t) = V_0 - V_t^{\text{src}}(z_t). \tag{6}$$

After obtaining V_t^{res} shown in Fig. 2(b), we can perform denoising starting from ϵ using the combined flow $V_t^{\text{src}}(z_t) + V_t^{\text{res}}(z_t) = V_0$. Following this adjusted flow trajectory ensures an accurate reconstruction of z_0 from ϵ.

To achieve video transfer under the target condition, we introduce a **Residual-Corrected Flow** based on the Consistency Residual Velocity. Specifically, given the model-predicted flow V_t^{tar} under the target condition and V_t^{res}, we define the final Residual-Corrected Flow as:

$$V_t^{\text{edit}}(z_t^{\text{edit}}) = V_t^{\text{tar}}(z_t^{\text{edit}}) + V_t^{\text{res}}(z_t). \tag{7}$$

Starting from the same initial Gaussian noise $z_1^{\text{edit}} = \epsilon$, we then perform the denoising process following V_t^{edit} from $t =$ 1 to 0, shown in Fig. 2(c). The final result of this process is our target relit video z_0^{edit}.

It is not hard to prove that when "tar"="src", then $V_t^{\text{edit}} = V_0$ and $z_0^{\text{tar}} = z_0$, thereby satisfying Stability under Identity. Our next step is to enforce "Directional Change" property to achieve an overall Condition Consistency.

Decoupled Condition Design. To strengthen Directional Change, we introduce a Decoupled Condition Design. Our condition input is carefully crafted using both illumination-specific and agnostic features to enable the model to more effectively identify what and where to make edits. As shown in Fig. 3, it comprises:

- **Reference frame condition.** We use an I2V model as F_θ to accept a reference frame. For the source condition, the reference frame is the first frame of the input video. For the target condition, it is the first frame edited by the powerful image relighting model, IC-Light [32], to provide a robust visual anchor that guarantees both illumination naturalness and spatial fidelity.
- **Structural condition.** We employ a pretrained ControlNet [1, 31] on a weighted fusion of the depth and edge maps extracted from the input video as the illumination-agnostic structural condition. It is shared by both source and target conditions, ensuring the preservation of structural content during editing.
- **Textual condition.** We use illumination-specific textual prompts, where the source and target prompts differ only in their descriptions of background and lighting, while sharing the same foreground content.

This decomposition provides a stable, sufficient, and disentangled guidance signal to enforce Condition Consistency.

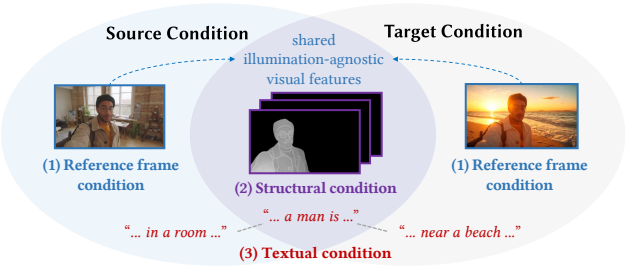


Figure 3. **Decoupled Condition Design.** The source and target conditions share identical illumination-agnostic information, differing only in their illumination-specific information.

3.4. High-Frequency Transfer

To further enhance the detail consistency between the generated target video and the original source video, we aim to inject a portion of high-frequency information from the source video into the generation process of the target video.

Specifically, let $\text{HF}(X)$ and $\text{LF}(X)$ denote the high-frequency and low-frequency components of the video X obtained by Fourier decomposition satisfying $X = \text{HF}(X) + \text{LF}(X)$. Then, during each generation step of the

target video z_t^{edit}, we perform the following replacement:

$$z_t^{\text{edit}} \leftarrow \text{LF}(z_t^{\text{edit}}) + \lambda \cdot \text{HF}(z_t) + (1-\lambda) \cdot \text{HF}(z_t^{\text{edit}}) \quad (8)$$

where z_t is the source video at step t and λ controls the proportion of high-frequency information injected. A larger λ strengthens the preservation of fine details from the source video but may limit the model's ability to fully adapt to the target illumination condition; conversely, a smaller λ allows more flexibility in relighting but may reduce structural and textural consistency.

Note that the injection maintains Stability under Identity since when reconstructing z_0, transferring high-frequency from z_t to z_t itself does not alter itself.

3.5. High-Quality Background Generation with Masking Mechanism

Residual-Corrected Flow and High-Frequency Transfer may introduce unwanted background details from source video into the result. To mitigate this conflict during new background generation, we use the mask to disentangle the foreground and background regions. Let M denote the mask of the foreground region extracted from the original video. Then, the Masked Residual-Corrected Flow is

$$V_t^{\text{edit}}(z_t^{\text{edit}}) = V_t^{\text{tar}}(z_t^{\text{edit}}) + M \cdot V_t^{\text{res}}(z_t). \quad (9)$$

Similarly, we apply the Masked High-Frequency Transfer only to the foreground region:

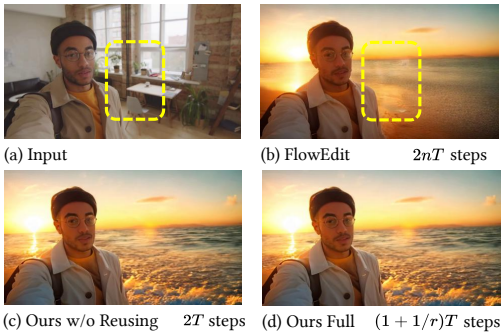
$$z_t^{\text{edit}} \leftarrow \text{LF}(z_t^{\text{edit}}) + \lambda M \cdot \text{HF}(z_t) + (1-\lambda M) \cdot \text{HF}(z_t^{\text{edit}}). \quad (10)$$

The structural condition of our Decoupled Condition Design is modified accordingly. We utilize a combination of the previously extracted features (*i.e.*, depth and edge), but only for the foreground region, leaving the background unconstrained to allow new content generation.

Using this mask mechanism, we can easily preserve the foreground details while generating the background unperturbedly without any interference from the source video. Note that this mask mechanism permits certain regions to violate the Stability under Identity property, allowing for more flexible editing.

3.6. Comparison to FlowEdit in Video Relighting

Inversion and FlowEdit also aim for editing based on reconstruction. Although inversion [8, 19, 23, 26] is commonly used in image generation models for editing, it remains both time-consuming and imprecise in video models. The imprecision also violates the Stability under Identity, resulting in poor Condition Consistency. By comparison, our method is inversion-free, which is partially inspired by the recent approach, FlowEdit [11]. It offers a novel path for image editing that satisfies Stability under Indentity without inversion. However, it has obvious limitations when applied to the video relighting task.



(a) Input (b) FlowEdit $2nT$ steps (c) Ours w/o Reusing $2T$ steps (d) Ours Full $(1+1/r)T$ steps

Figure 4. **Comparison with FlowEdit.** (a) Input video. (b) FlowEdit produces blurry outputs, and suffers from ghosting artifacts due to the interference from the original background (yellow region). (c)(d) Our residual reusing strategy effectively reduces the number of prediction steps with negligible quality degradation.

In FlowEdit, starting with the source image $z_1^{\text{edit}} = z_0$, the target image z_t^{edit} evolves along the flow V_t^{edit}:

$$V_t^{\text{edit}}(z_t^{\text{edit}}) = V_t^{\text{tar}}(z_t^{\text{pred}}) - V_t^{\text{src}}(z_t), \quad (11)$$

$$z_t = (1-t)z_0 + t\epsilon_t, \quad (12)$$

$$z_t^{\text{pred}} = z_t + z_t^{\text{edit}} - z_0, \quad (13)$$

from $t = 1$ to 0, where $\epsilon_t \sim \mathcal{N}(0, I)$. Our design differs from FlowEdit in two aspects: 1) FlowEdit samples n different Gaussian noises at each timestep while we propose to use a single Gaussian noise ϵ throughout the whole editing process. 2) FlowEdit constructs an edit flow from the source video $z_1^{\text{edit}} = z_0$ to target result z_0^{edit}. We rephrase this process and build a new generation flow from noise $z_1^{\text{edit}} = \epsilon$ to the target z_0^{edit} by tracking z_t^{pred} in FlowEdit, just as the standard diffusion denoising process.

It can be proved that if FlowEdit samples a fixed noise across steps, it is theoretically equivalent to our method without masking mechanism. However, it is the above two straightforward changes that make a big difference, establishing key advantages of our approach over FlowEdit:

- Clear Generation Ability. To ensure stable editing, FlowEdit independently samples multiple Gaussian noises and averages V_t^{edit} at each timestep, which inevitably causes blur in the generated background as in Fig. 4 (b). Our fixed noise prevents blur, significantly improving the generation quality.
- Acceleration. Our method and FlowEdit require computing two velocities, doubling the denoising steps from T to $2T$ ($2nT$ for FlowEdit if averaging $n \geq 1$ velocities per step). The benefit we offer is that the proposed consistency residual velocity relies solely on the fixed noise and source condition, thus it is stable across steps and can be reused. Our experiments have shown that applying the same residual velocity every r steps has little impact on quality, while it decreases the total steps to $(1+1/r)T$, as shown in Fig. 4 (c)(d). In contrast, FlowEdit requires new

random noises and target-condition denoising for each edit flow calculation, which prevents reusability.

- Spatial Controllability. The edit flow of FlowEdit starts from the source video, which cannot handle pure generation within specific regions. The background from the source video will interfere with the new background generation as in Fig. 4 (b). By comparison, our method, starting from pure noises, enables foreground masking for editing while leaving the background for pure generation.
- Decoupled Condition. Additionally, we propose a Decoupled Condition Design to combine the reference frame and structural conditions, which helps the model achieve better Condition Consistency.

4. Experiments

4.1. Experimental Setup

Implementation details. We implement our method based on Wan2.1 [1, 25], a widely used state-of-the-art open-sourced video diffusion model. For the foreground mask M, we use BiRefNet [35] to generate the mask for the first frame and MatAnyone [28] to propagate it throughout the entire video. We downsample M to match the video shape in the latent space for masking operation. For the Decoupled Condition Design, the reference frame is generated by IC-Light [32], and the structural information is obtained by combining the HED, depth, and Canny maps [31], followed by masking with M. For Residual-Corrected Flow, we adopt $T = 50$ timesteps for the generation process and reuse the Consistency Residual Velocity every $r = 10$ steps, resulting in a total of 55 steps. For High-Frequency Transfer, we set the frequency decomposition threshold to 0.8 and the injection intensity to $\lambda = 0.5$. The experiments are conducted on a single 80 GB NVIDIA A100 GPU. Representative results are illustrated in Fig. 1 and Fig. 7.

Baseline methods. We compare our method with recent video relighting approaches. Training-free methods such as AnyPortal [5] and Light-A-Video [36] apply the image relighting model IC-Light frame by frame and integrate the results into video generation models. Training-based methods like Lumen [29] and TC-Light [16] are trained on paired video datasets constructed for the relighting task. The basic information for these methods is summarized in Tab. 1. For fairness, we compare only with methods that support new background generation, excluding RelightVid [4] and Lux Post Facto [18] since they are not open-sourced.

Metrics. To comprehensively evaluate the quality of relighted videos, we assess our results from four perspectives.

- **Video–text alignment.** We employ CLIP-T cosine similarity [21] to measure the alignment between generated frames and the target relighting prompts.
- **Temporal smoothness.** We use CLIP-I [21], the CLIP-based cosine similarity between consecutive frames, to evaluate temporal consistency in the generated videos.
- **Foreground consistency.** As existing metrics rarely address the consistency of primary subjects in relighting tasks, we propose a new evaluation protocol. Specifically, we first extract the foreground region [34] of each output video, then compute (a) **Structural consistency** using SSIM between Canny edge maps [31], (b) **Detail consistency** using PSNR between LoG responses [17] of albedo predictions [2], (c) **Motion consistency** using SSIM between optical flow maps [24], and (d) **Identity consistency** by calculating facial ID similarity [3] for human-centric clips or CLIP-I similarity for others.
- **User preference.** We invite 23 participants to select the best result among four methods based on four criteria: (a) User-Pmt (relevance to the prompt), (b) User-Tmp (temporal coherence), (c) User-Fg (preservation of foreground details and motion), and (d) User-Lit (quality and harmonization of relighting on the foreground).

Table 1. Summarization of video relighting methods.

Method	Training-Free	Video Base Model	Support Diverse Resolution	Support Background Replacement
AnyPortal	√	CogVideoX	×	√
Light-A-Video (A)	√	AnimateDiff	√	√
Light-A-Video (C)	√	CogVideoX	×	×
Light-A-Video (W)	√	Wan2.1	√	×
Lumen	×	Wan2.1	√	√
TC-Light	×	VidToMe	√	×
FlowPortal (Ours)	√	Wan2.1	√	√

4.2. Comparison to State-of-the-Art Methods

Qualitative results. Figure 5 shows the qualitative comparison between FlowPortal and other methods. The training-based method Lumen fails to accurately preserve detail fidelity and also exhibits poor lighting alteration and richness. The training-free methods AnyPortal and Light-A-Video produce poor foreground consistency, background quality, and lighting harmonization.

Quantitative results. We construct a test set consisting of 69 pairs of real-world video clips and corresponding relighting prompts with broad diversity, including 54 human-centric videos, 8 animal scenes, and 7 other objects for objective evaluation. Table 2 shows the results among different methods. Our method achieves the best video-text alignment, temporal smoothness, motion consistency, and detail consistency. The training-based method Lumen introduces less variation in lighting conditions with foreground nearly unaltered, which achieves the best detail consistency and identity consistency. Our method ranks the second with negligible consistency value drops, but achieves significant lighting adjustment, indicating our method's high Condition Consistency. For subjective evaluation, Table 2 shows the user preference scores averaged over 17 randomly selected

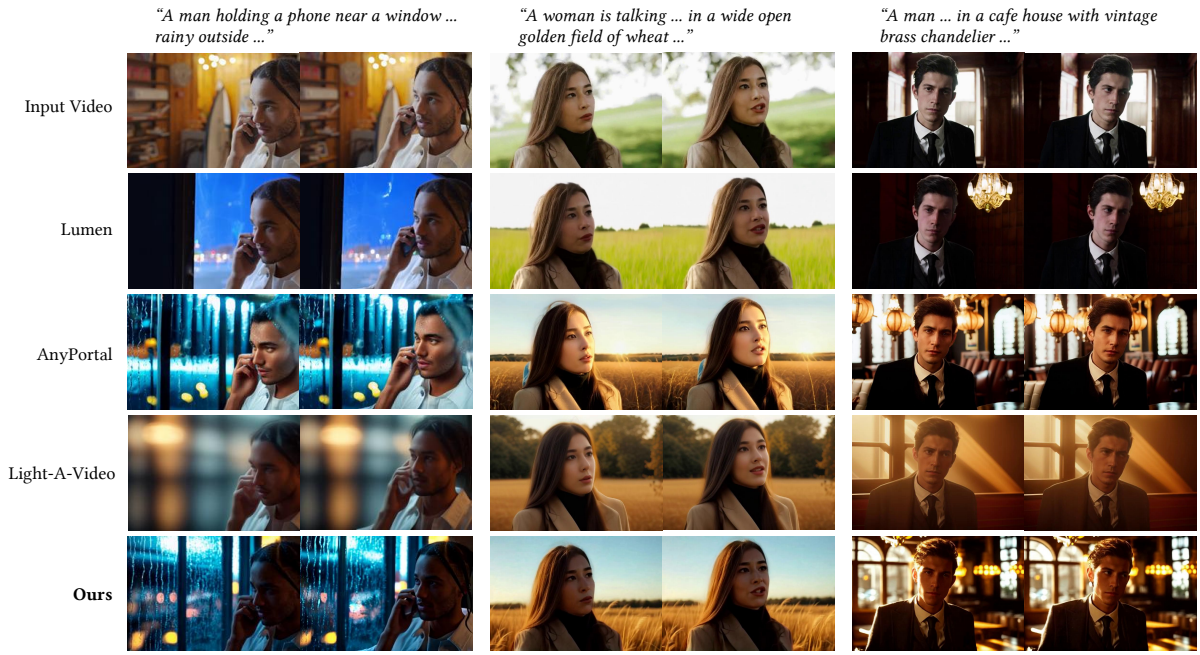


Figure 5. **Qualitative comparison.** The training-based Lumen exhibits insufficient lighting richness and diversity, with foreground nearly unaltered. The training-free AnyPortal and Light-A-Video show poor structural fidelity and lighting quality. Our method not only maintains structural and detail consistency but also demonstrates high-quality background generation and rich relighting effects.

Table 2. Quantitative and user study results.

Method	CLIP-T	CLIP-I	Structural Consistency	Motion Consistency	Detail Consistency	Identity Consistency	User-Pmt	User-Tmp	User-Fg	User-Lit
AnyPortal	0.3196	0.9817	0.8530	0.8876	40.4853	0.4310	15.3	9.9	8.9	12.1
Light-A-Video (A)	0.2956	0.9684	0.8580	0.8869	40.8727	0.5076	4.7	3.7	3.4	9.4
Lumen	0.3055	0.9746	**0.8809**	0.8914	40.4193	**0.7392**	22.7	28.1	35.5	24.4
FlowPortal (Ours)	**0.3271**	**0.9828**	0.8804	**0.8944**	**41.2044**	0.7328	**57.4**	**58.4**	**52.2**	**54.2**

results and 24 participants. Our method achieves the best overall user preference.

Running time. We report the running time of training-free methods on an 80 GB NVIDIA A100 GPU. AnyPortal and Light-A-Video require 20–30 minutes per video due to complex pipelines. In contrast, our method only takes 3–5 minutes across different resolutions roughly equivalent to the direct inference time of a single video diffusion model.

4.3. Ablation Study

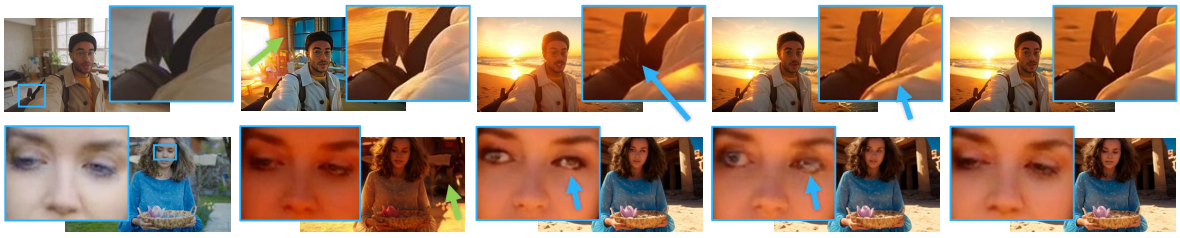
We conduct ablation studies to evaluate the effectiveness of Residual-Corrected Flow, High-Frequency Transfer, and the masking mechanism. The qualitative results are shown in Fig. 6, and the quantitative results are presented in Tab. 3.

- **Masking mechanism.** Inference without masking the Consistency Residual Velocity and High-Frequency Transfer causes interference with the generation of the new background, resulting in an unchanged background structure and poor prompt relevance.
- **Residual-Corrected Flow.** Performing direct inference instead of using the Residual-Corrected Flow leads to severe structural incoherence.
- **High-Frequency Transfer.** The texture-level details fail to be preserved without High-Frequency Transfer due to inaccuracy in structural information and the model's imperfect controllability under these conditions.

Table 3. Quantitative ablation study.

Metric	w/o Mask	w/o Residual	w/o High-Frequency	Full
CLIP-T	0.2809	**0.3310**	0.3290	0.3271
CLIP-I	0.9792	0.9825	0.9798	**0.9828**
Structural Cons.	0.8649	0.8516	0.8688	**0.8804**
Motion Cons.	0.8923	0.8933	0.8882	**0.8944**
Detail Cons.	40.4353	38.5027	40.4852	**41.2044**
Identity Cons.	0.7123	0.4153	0.5527	**0.7328**

Decoupled Condition Design. We study the effect of our Decoupled Condition Design in Fig. 8. It can be seen that when using only text conditions, the foreground character fails to be reconstructed properly. Adding structural conditions maintains consistent character structure but fails to



(a) Input (b) w/o Mask (c) w/o Residual (d) w/o High-Frequency **(e) Full**

Figure 6. **Ablation Study.** Without mask, background cannot be properly generated in (b). Enlarged local blue regions illustrate the structural inconsistency if (c) removing Residual-Corrected Flow and detail inconsistency if (d) removing High-Frequency Transfer.

Figure 7. **More visual results.** Our method can generate realistic backlighting effects that simulate light penetrating through clothing, as well as clear shadows cast by objects onto the background under strong lighting conditions. See our project page: https://gaowenshuo.github.io/FlowPortalProject/ for video demonstration.

produce natural golden lighting. Incorporating the reference frame results in stronger and prompt-consistent lighting, but leads to incomplete character structure. Only when all conditions are utilized, does our method achieve the most natural results in both structure and lighting.

5. Conclusion and Discussion

In this paper, we propose FlowPortal, a novel training-free framework for efficient video relighting and background replacement. We introduce a novel Residual-Corrected Flow mechanism with Decoupled Condition Design that enforces Condition Consistency. A High-Frequency Transfer module is designed to enhance the detail fidelity and a masking mechanism is applied to isolate background regions for high-quality generation. Both qualitative and quantitative experiments demonstrate that our method achieves superior background generation quality, foreground consistency, and lighting realism.

Although our method performs well in most cases, it may still be constrained by the generative capability of the underlying base models (*i.e.*, IC-Light for relighting the reference frame and Wan2.1 for background generation).

Figure 8. **Ablation on Decoupled Condition Design.** The absence of structural information or the reference frame leads to degraded structural fidelity and unnatural illumination.

It is evident that our Residual-Corrected Flow framework is not exclusively designed for relighting. A potential future direction could be the exploration of extending it to various video editing tasks, including video stylization, colorization, object manipulation, human editing, and even more general editing tasks. Further algorithmic improvements on the Residual-Corrected Flow, reconstruction rather than relying on direct directional addition, would be another promising future direction.

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China under Grant 62471009, in part by the Key Laboratory of Science, Technology and Standard in Press Industry (Key Laboratory of Intelligent Press Media Technology), and in part by The Fundamental Research Funds for the Central Universities, Peking University.

References

[1] AIGC-Apps. Videox-fun: Github repository, 2025. GitHub repository: github.com/aigc-apps/VideoX-Fun. 4, 6

[2] Chris Careaga and Yağız Aksoy. Colorful diffuse intrinsic image decomposition in the wild. *ACM TOG*, 43(6):1–12, 2024. 6

[3] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, pages 4690–4699, 2019. 6

[4] Ye Fang, Zeyi Sun, Shangzhan Zhang, Tong Wu, Yinghao Xu, Pan Zhang, Jiaqi Wang, Gordon Wetzstein, and Dahua Lin. Relightvid: Temporal-consistent diffusion model for video relighting. *arXiv preprint arXiv:2501.16330*, 2025. 2, 6

[5] Wenshuo Gao, Xicheng Lan, and Shuai Yang. Anyportal: Zero-shot consistent video background replacement. In *ICCV*, pages 18990–18999, 2025. 2, 6

[6] Kaiwen Guo, Peter Lincoln, Philip Davidson, Jay Busch, Xueming Yu, Matt Whalen, Geoff Harvey, Sergio Orts-Escolano, Rohit Pandey, Jason Dourgarian, et al. The relightables: Volumetric performance capture of humans with realistic relighting. *ACM TOG*, 38(6):1–19, 2019. 2

[7] Kai He, Ruofan Liang, Jacob Munkberg, Jon Hasselgren, Nandita Vijaykumar, Alexander Keller, Sanja Fidler, Igor Gilitschenski, Zan Gojcic, and Zian Wang. Unirelight: Learning joint decomposition and synthesis for video relighting. *arXiv preprint arXiv:2506.15673*, 2025. 2

[8] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross-attention control. In *ICLR*, pages 1–12, 2022. 5

[9] Hoon Kim, Minje Jang, Wonjun Yoon, Jisoo Lee, Donghyun Na, and Sanghyun Woo. Switchlight: Co-design of physics-driven architecture and pre-training framework for human portrait relighting. In *CVPR*, pages 25096–25106, 2024. 2

[10] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024. 3

[11] 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. 5

[12] Yuanze Lin, Yi-Wen Chen, Yi-Hsuan Tsai, Ronald Clark, and Ming-Hsuan Yang. Illumicraft: Unified geometry and illumination diffusion for controllable video generation. *arXiv preprint arXiv:2506.03150*, 2025. 2

[13] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 3

[14] Ropeway Liu, Hangjie Yuan, Bo Dong, Jiazheng Xing, Jinwang Wang, Rui Zhao, Yan Xing, Weihua Chen, and Fan Wang. Unilumos: Fast and unified image and video relighting with physics-plausible feedback. *arXiv preprint arXiv:2511.01678*, 2025. 2

[15] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022. 3

[16] 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. 2, 6

[17] David Marr and Ellen Hildreth. Theory of edge detection. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 207(1167):187–217, 1980. 6

[18] Yiqun Mei, Mingming He, Li Ma, Julien Philip, Wenqi Xian, David M George, Xueming Yu, Gabriel Dedic, Ahmet Levent Taşel, Ning Yu, et al. Lux post facto: Learning portrait performance relighting with conditional video diffusion and a hybrid dataset. In *CVPR*, pages 5510–5522, 2025. 2, 6

[19] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *CVPR*, pages 6038–6047, 2023. 5

[20] Rohit Pandey, Sergio Orts-Escolano, Chloe Legendre, Christian Haene, Sofien Bouaziz, Christoph Rhemann, Paul E Debevec, and Sean Ryan Fanello. Total relighting: learning to relight portraits for background replacement. *ACM TOG*, 40 (4):43–1, 2021. 2

[21] Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. Fatezero: Fusing attentions for zero-shot text-based video editing. In *ICCV*, pages 15932–15942, 2023. 6

[22] Mengwei Ren, Wei Xiong, Jae Shin Yoon, Zhixin Shu, Jianming Zhang, HyunJoon Jung, Guido Gerig, and He Zhang. Relightful harmonization: Lighting-aware portrait background replacement. In *CVPR*, pages 6452–6462, 2024. 2

[23] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 5

[24] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *ECCV*, pages 402–419. Springer, 2020. 6

[25] 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

[26] Chen Henry Wu and Fernando De la Torre. A latent space of stochastic diffusion models for zero-shot image editing and guidance. In *ICCV*, pages 7378–7387, 2023. 5

[27] Xiaoyan Xing, Konrad Groh, Sezer Karaoglu, Theo Gevers, and Anand Bhattad. Luminet: Latent intrinsics meets diffusion models for indoor scene relighting. In *CVPR*, pages 442–452, 2025. 2

[28] 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. 6

[29] 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. 2, 6

[30] Longwen Zhang, Qixuan Zhang, Minye Wu, Jingyi Yu, and Lan Xu. Neural video portrait relighting in real-time via consistency modeling. In *ICCV*, pages 802–812, 2021. 2

[31] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, pages 3836–3847, 2023. 4, 6

[32] 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. 2, 4, 6

[33] Yuxin Zhang, Dandan Zheng, Biao Gong, Shiwen Wang, Jingdong Chen, Ming Yang, Weiming Dong, and Changsheng Xu. Lumisculpt: enabling consistent portrait lighting in video generation. *arXiv preprint arXiv:2410.22979*, 2024. 2

[34] Peng Zheng, Dehong Gao, Deng-Ping Fan, Li Liu, Jorma Laaksonen, Wanli Ouyang, and Nicu Sebe. Bilateral reference for high-resolution dichotomous image segmentation. *CAAI Artificial Intelligence Research*, 2024. 6

[35] 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. 6

[36] 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. 2, 6