WORLDPORTAL: RECAST RESIDUAL CORRECTION FOR INSTRUCTION-GUIDED VIDEO BACKGROUND REPLACEMENT AND RELIGHTING

Anonymous authors
Paper under double-blind review

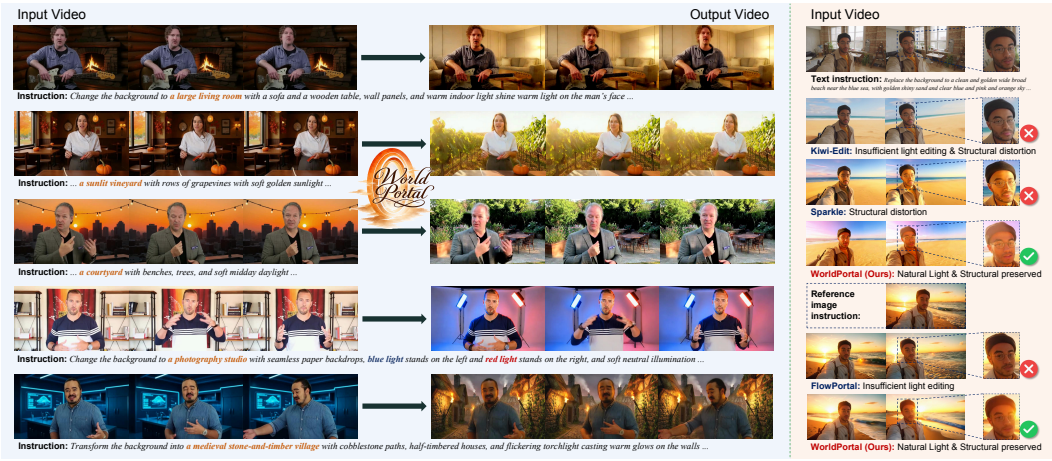

Figure 1: We propose WorldPortal for instruction-guided video background replacement and foreground relighting. **Left:** Given an input video and a text instruction, WorldPortal replaces the background and adapts the foreground illumination while preserving subject identity and temporal consistency. **Right:** The upper example evaluates text-instruction-guided editing against Kiwi-Edit and Sparkle, where WorldPortal produces more natural illumination while better preserving facial structure. The lower example evaluates reference-image-guided editing against FlowPortal, demonstrating more natural light without sacrificing foreground fidelity. *Zoom in for better view.*

ABSTRACT

Instruction-guided video background replacement and relighting require a model to synthesize a new environment and plausible illumination while preserving the identity, structure, texture, and motion of the foreground subject. Existing methods struggle to balance these objectives: training-free pipelines rely on rigid structural conditions that restrict natural lighting changes, whereas directly trained models degrade foreground structures and details. We present WorldPortal, a unified video editing framework based on Recast Residual Correction. During inference, a forward branch first predicts the video under the target background and illumination. A backward recasting branch then places the edited foreground state back into the source background domain, where its discrepancy from the source trajectory primarily captures foreground discrepancies introduced by editing. Transferring this residual to the target-domain prediction restores foreground details. To support both branches, we further introduce a joint VLM-DiT architecture that adaptively parses the instruction and input video into semantic-, foreground-, and background-centric token groups, providing flexible guidance. Experiments demonstrate that WorldPortal achieves superior overall performance in instruction alignment, lighting quality, structural fidelity, and fine-detail preservation, while introducing only modest additional inference cost.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Instruction-guided video background replacement and foreground relighting aim to place a subject in a new environment while adapting its illumination to the target scene and preserving its identity, structure, texture, and motion. This capability is valuable for film post-production, e-commerce advertising, social-media content creation, and immersive virtual reality, where coherent scene changes would otherwise require costly reshooting or manual frame-level refinement. It can also support world-model simulation by composing dynamic agents with diverse environments and illumination conditions, enabling scalable construction of visually consistent interaction scenarios. Despite this broad utility, the task remains challenging: changing the scene and its illumination requires substantial generative freedom, whereas maintaining the foreground requires strict control over spatial details and temporal dynamics.

Existing approaches broadly fall into training-free and training-based paradigms. One category of training-free methods combines image relighting models with video generation or editing priors. For example, AnyPortal (Gao et al., 2025b) and Light-A-Video (Zhou et al., 2025) apply IC-Light-based relighting (Zhang et al., 2025) to individual frames or reference images and subsequently introduce cross-frame attention or temporal modules to improve consistency. Because relighting is not modeled holistically in video space, these pipelines may suffer from temporal discontinuities, structural mismatches, and complex multi-stage inference. Another training-free approach, FlowPortal (Gao et al., 2025a), performs residual-corrected flow with decoupled structural conditions and high-frequency transfer. Although it improves foreground fidelity, its hard-coded designed edge and depth conditions implicitly constrain structure to remain compatible across illumination changes, while high-frequency transfer assumes that fine details can be preserved directly before and after relighting. Such manually imposed constraints reduce generative freedom and may suppress shadows, highlights, color shifts, and other natural lighting responses required by the target environment.

Training-based methods instead learn background replacement and relighting directly from paired editing data. Sparkle (Zeng et al., 2026), for example, trains an instruction-guided video model with decoupled foreground and background guidance. These methods offer a simpler inference pipeline and greater flexibility in generating target scenes. However, their preservation constraints are comparatively implicit: while synthesizing a new background and illumination, the model may repaint the foreground as a side effect. This leads to identity drift, boundary flicker, local structural distortion, and unstable motion details. Existing methods therefore face a fundamental trade-off. Hand-crafted structural constraints can be too strong to permit natural relighting, whereas direct learned constraints can be too weak to preserve fine foreground content.

To address this, we propose **WorldPortal**, an instruction-guided video editing framework based on **Recast Residual Correction**. Rather than relying solely on a direct target prediction, our inference procedure introduces a backward recasting branch that places the foreground state produced by forward editing back into the source background domain. Comparing this recast prediction with the aligned source trajectory largely suppresses the shared background component and primarily captures foreground discrepancies introduced by the edit. We then transfer the resulting residual to the target-domain prediction, restoring identity, structure, texture, and motion details while retaining the desired background and illumination. This formulation explicitly compensates for the weak detail preservation of direct training without forcing the target result to copy the source lighting.

We further avoid rigid hand-crafted control through a joint VLM-DiT architecture with adaptive condition parsing. Learnable queries extract three complementary groups of tokens: semantic-centric tokens describe the editing intent, foreground-centric tokens encode subject appearance and motion, and background-centric tokens capture scene context and illumination. Forward editing and backward recasting activate different token groups and replace unused groups with learnable null queries. Unlike fixed masks, edges, depth maps, or high-frequency, these learned tokens adapt their strength and content to each input, providing sufficient structure awareness without suppressing plausible illumination changes. Experiments show that WorldPortal achieves the best overall performance in instruction alignment, lighting quality, structural fidelity, and fine-detail preservation.

Our contributions are summarized as follows:

- We introduce Recast Residual Correction, an inference algorithm that estimates foreground degradation by recasting the edited state into the source background domain and transfers the resulting residual to correct the target trajectory and repair foreground degradation.
- We develop an adaptive token group condition design that jointly represents editing semantics, foreground content, and background context, providing more flexible control.
- We present a unified VLM-DiT architecture that learns these adaptive conditions and supports both forward editing and backward recasting for instruction-guided video background replacement and relighting.

2 Related Work

Image relighting and background replacement. Image relighting modifies foreground illumination while preserving identity, geometry, albedo, and texture. Earlier portrait and scene methods typically estimate intrinsic or surface properties and then re-render the input under target lighting. Total Relighting (Pandey et al., 2021) predicts portrait normals and albedo for background-aware relighting, while SwitchLight (Kim et al., 2024) combines a physics-driven architecture with a dedicated pre-training strategy. More recent diffusion-based methods exploit generative priors to model richer light transport. Relightful Harmonization (Ren et al., 2024) conditions diffusion on the target background, LumiNet (Xing et al., 2025) transfers indoor lighting through intrinsic and extrinsic representations, and IC-Light (Zhang et al., 2025) learns consistent light transport for illumination harmonization. Related approaches include DiffusionLight, DiLightNet, Neural Gaffer, and SynthLight (Phongthawee et al., 2024; Zeng et al., 2024; Jin et al., 2024; Chaturvedi et al., 2025). These image methods provide strong lighting priors and are often used as building blocks for video relighting, but frame-wise application alone cannot guarantee temporal coherence or motion consistency.

Video relighting. Video relighting introduces temporal constraints in addition to illumination realism. Classical or reconstruction-based approaches such as The Relightables and Neural Video Portrait Relighting (Guo et al., 2019; Zhang et al., 2021) capture or learn relightable dynamic human appearances with explicit temporal modeling. Recent video diffusion methods learn relighting from paired or synthesized video data. RelightVid (Fang et al., 2025) trains a temporal-consistent diffusion model for video relighting under background, text, or environment-map conditions. IllumiCraft (Lin et al., 2025) combines HDR video maps, appearance cues, and 3D point tracks for controllable relit video generation. UniRelight (He et al., 2025b) jointly estimates albedo and synthesizes relit videos in a single diffusion model. Lumen (Zeng et al., 2025) targets consistent video relighting with harmonious background replacement using realistic and synthetic relighting pairs. Lux Post Facto and TC-Light (Mei et al., 2025; Liu et al., 2025) further explore portrait-performance relighting and temporally coherent long-video relighting. These training-based methods can model temporal lighting changes, but they require expensive paired or simulated relighting data and may struggle to balance illumination richness with strict foreground fidelity under diverse instructions.

Training-free video relighting and background replacement. Training-free pipelines can be broadly divided into two categories. The first avoids collecting paired relighting videos by combining image relighting models with video generation or editing priors. AnyPortal and Light-A-Video (Gao et al., 2025b; Zhou et al., 2025) are representative examples: both leverage IC-Light for image-level relighting and then use video models or temporal modules to improve consistency. However, since IC-Light is still applied to individual frames or reference images rather than to the video holistically, such pipelines may suffer from temporal discontinuities, structural mismatches, or complex inference chains. The second category is represented by FlowPortal (Gao et al., 2025a), which addresses training-free video relighting and background replacement more directly through residual-corrected flow, decoupled condition design, and high-frequency transfer. Its design improves structure and detail preservation, but relies on mask prediction, first-frame IC-Light results, textual source and target captions, and hard-coded structural conditions such as depth and edges. All these strong structural constraints and the high-frequency transfer mechanism assume that the overall structure, edges, and fine-scale details remain compatible before and after relighting. Although these assumptions protect fidelity, they constrain illumination changes and result in less natural and diverse lighting effects.

Instruction-guided background replacement. Instruction-guided video editing datasets have recently expanded the scale of paired supervision, including background-change categories in OpenVE-3M and reference-guided control in Kiwi-Edit (He et al., 2025a; Lin et al., 2026). Sparkle (Zeng et al., 2026) focuses more specifically on background replacement, arguing that this task is harder than local editing or style transfer because it requires a new temporally consistent scene and plausible foreground-background interaction; it therefore constructs decoupled foreground/background guidance and filtered background-replacement pairs. These data-driven methods simplify inference and improve background diversity, but they are usually based on direct target prediction. When the model generates a new background and illumination, the foreground may be repainted as a side effect, leading to identity drift, texture softening, boundary flicker, and motion instability. Our method targets this relighting-centered failure mode: rather than relying only on external structure conditions or a single target prediction, we explicitly estimate foreground degradation through backward recasting and use the resulting residual to restore foreground details while retaining the target illumination and background.

3 WORLDPORTAL

3.1 TASK FORMULATION

Given a source video V_s and an editing instruction p, our goal is to generate an edited video

$$V_o = \mathcal{G}(V_s, p), \tag{1}$$

whose background and illumination follow the target semantics specified by p, while the foreground preserves the identity, appearance, structure, and motion of V_s. We denote the latent encoding of the source video by z_0^s and perform the editing process in the latent space.

3.2 PRELIMINARIES: FLOW MATCHING

Flow matching learns a time-dependent velocity field that transports samples between the data distribution and a simple noise distribution (Lipman et al., 2023; Liu et al., 2023). Given a clean video latent z_0 and Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$, we define the linear interpolation path

$$z_t = (1-t)z_0 + t\epsilon, \qquad t \in [0, 1]. \tag{2}$$

The corresponding target velocity is constant along this path:

$$u_t = \frac{dz_t}{dt} = \epsilon - z_0. \tag{3}$$

In flow matching, generation starts from $z_1 \sim \mathcal{N}(0, I)$ and integrates the learned conditional velocity field $v_\theta(z_t, t, c)$ backward from $t = 1$ to $t = 0$, where c denotes the generation conditions. With a step size $\Delta t > 0$, one Euler step is

$$z_{t-\Delta t} = z_t - \Delta t\, v_\theta(z_t, t, c). \tag{4}$$

This one-step prediction provides the basic update on which our method is constructed.

3.3 RECAST RESIDUAL CORRECTION

We propose a **Recast Residual Correction** mechanism, illustrated in Figure 2. The core idea is not to perform only one source-to-target forward edit, but to introduce a backward recasting branch during inference, simulating degradation under the original conditions and thereby repairing the degradation in the edited result. First, the forward editing branch generates a prediction with the target background and illumination according to the text instruction. However, its foreground may be distorted and not fully faithful to the source. Then, the backward recasting branch recasts the foreground state in this editing prediction back to the source background domain, producing a recast prediction. Next, we identify the deviation of the recast prediction from the source video reconstruction. Since the two predictions share the source background, their difference largely suppresses the shared background component and primarily captures foreground discrepancies introduced by forward editing. This difference is the foreground repair recast residual. Finally, we apply this deviation as a residual to the target editing prediction from the forward editing branch, recovering source foreground details while preserving the target background and illumination.

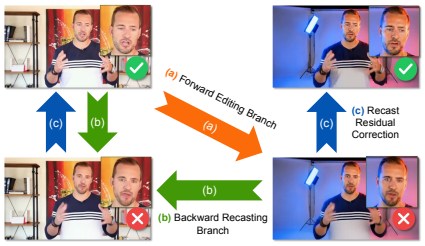

Figure 2: Overview of Recast Residual Correction. Forward editing creates the target scene, backward recasting returns the edited foreground to the source domain, and the recast residual corrects foreground degradation.

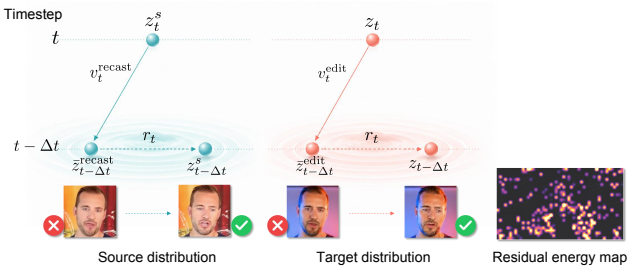

Figure 3: Recast residual correction in latent flow space. The source-domain residual r_t is transferred to the forward editing proposal to compensate for foreground degradation.

At each inference step, we maintain an editing trajectory z_t together with an aligned source trajectory. Let z_0^s be the latent encoding of the source video. Using the same initial noise ϵ as the editing trajectory, we construct the source state at time t as

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

Sharing the noise couples the two trajectories and makes their one-step predictions directly comparable.

We first perform forward editing with forward condition $c_{\text{edit}}(p, z_0^s)$, where p is the target instruction and z_0^s is the source-video latent. The model predicts forward velocity

$$v_t^{\text{edit}} = v_\theta\big(z_t, t, c_{\text{edit}}(p, z_0^s)\big), \tag{6}$$

and produces the one-step editing proposal

$$\bar{z}_{t-\Delta t}^{\text{edit}} = z_t - \Delta t\, v_t^{\text{edit}}. \tag{7}$$

Under the linear flow path, the corresponding one-step clean prediction is

$$\hat{z}_0^{\text{edit}} = z_t - t\, v_t^{\text{edit}}. \tag{8}$$

This prediction contains the foreground state currently produced by the editing branch, including any loss of identity, texture, or structure introduced by the edit.

We then recast this predicted state into the source background domain using $c_{\text{recast}}(p_{\text{null}}, z_0^s, \hat{z}_0^{\text{edit}})$, where p_{null} is the null instruction, z_0^s provides the source background domain, and $\hat{z}_0^{\text{edit}}$ provides the foreground state to be recast. Starting from the aligned source state z_t^s, the recasting branch predicts backward recast velocity

$$v_t^{\text{recast}} = v_\theta\big(z_t^s, t, c_{\text{recast}}(p_{\text{null}}, z_0^s, \hat{z}_0^{\text{edit}})\big), \tag{9}$$

followed by the one-step recast proposal

$$\bar{z}_{t-\Delta t}^{\text{recast}} = z_t^s - \Delta t\, v_t^{\text{recast}}. \tag{10}$$

Meanwhile, the aligned source trajectory at the next time step is known analytically:

$$z_{t-\Delta t}^s = \big(1 - t + \Delta t\big)z_0^s + \big(t - \Delta t\big)\epsilon. \tag{11}$$

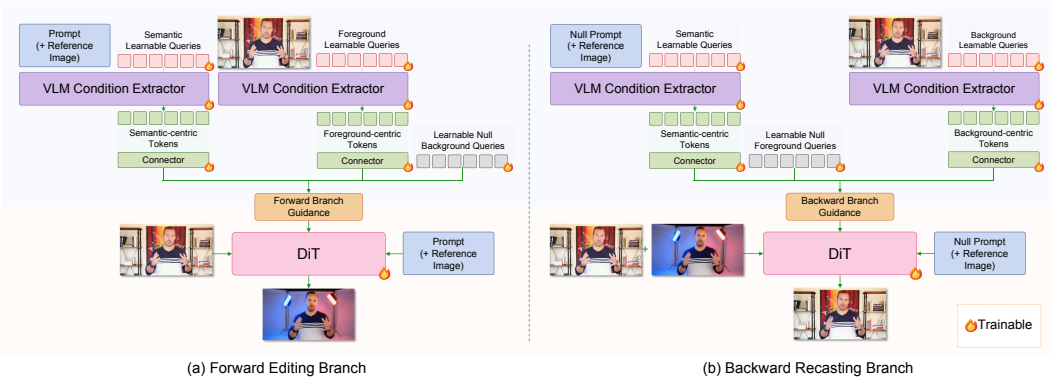

Figure 4: Dual-branch model design. (a) The forward editing branch combines semantic-centric and foreground-centric guidance with null background queries. (b) The backward recasting branch combines null-prompt semantic guidance and background-centric guidance with null foreground queries, while injecting the source and edited states as role-specific video references. Flame icons denote trainable components.

Both $\bar{z}^{\text{recast}}_{t-\Delta t}$ and $z^{s}_{t-\Delta t}$ lie in the source background domain. Their difference therefore largely suppresses the shared background component and primarily captures foreground discrepancies introduced by forward editing. We define the recast residual as

$$r_t = z^{s}_{t-\Delta t} - \bar{z}^{\text{recast}}_{t-\Delta t}. \tag{12}$$

Finally, we transfer this discrepancy to the target-domain proposal $\bar{z}^{\text{edit}}_{t-\Delta t}$, yielding the corrected latent:

$$z_{t-\Delta t} = \bar{z}^{\text{edit}}_{t-\Delta t} + r_t. \tag{13}$$

The forward proposal determines the target background and illumination, whereas the recast residual restores source-consistent foreground information, as summarized in Figure 3. Repeating these steps from $t = 1$ to $t = 0$ yields the final latent, which is decoded into the edited video.

3.4 Dual-Branch Model Design

As shown in Figure 4, we implement forward editing and backward recasting as two operating modes of a shared video DiT. Both modes use a Vision Language Model (VLM) condition extractor to produce fixed-length guidance tokens through 3 sets of learnable queries and lightweight connectors, and expose 3 complementary token groups: semantic-centric tokens for the editing intent, foreground-centric tokens for subject appearance and motion, and background-centric tokens for scene context and illumination. A branch activates only the token groups relevant to its role and replaces the remaining group with learnable null queries.

Forward editing branch. The forward branch realizes $c_{\text{edit}}(p, z^{s}_{0})$. The prompt p (and an optional reference image in the image-conditioned setting) is parsed with semantic learnable queries, producing semantic-centric tokens that describe the target environment and illumination. In parallel, the source video is parsed with foreground learnable queries to obtain foreground-centric tokens that encode subject identity, appearance, and motion. The background token group is replaced with learnable null background queries. After projection by their connectors, the resulting semantic, foreground, and null-background token groups constitute the forward-branch guidance. The DiT additionally receives z^{s}_{0} as its video reference and the original prompt condition (and an optional reference image), and predicts the target-domain velocity v^{edit}_{t}.

Backward recasting branch. The recasting branch realizes $c_{\text{recast}}(p_{\text{null}}, z^{s}_{0}, \hat{z}^{\text{edit}}_{0})$. Its semantic queries receives the null prompt p_{null}. The foreground token group is replaced with learnable null foreground queries, while the source video is parsed with background learnable queries to extract background-centric tokens describing the original scene and illumination. The resulting semantic, null-foreground, and background token groups constitute the backward-branch guidance. At the

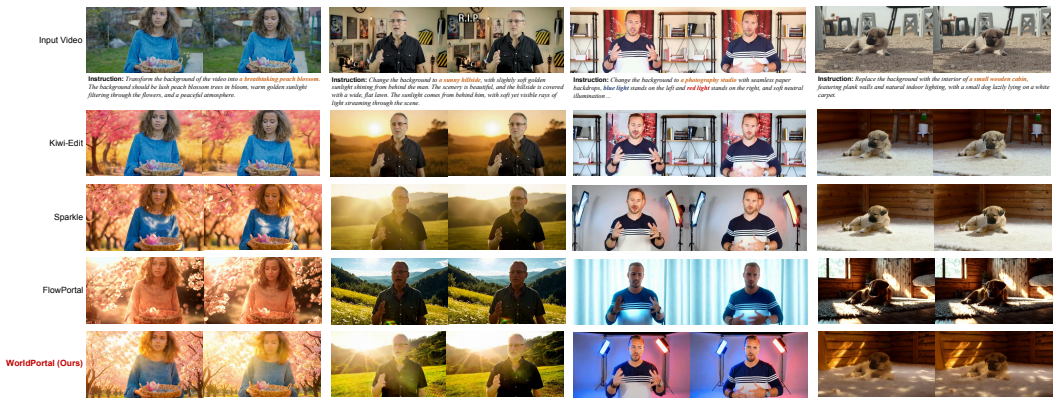

Figure 5: Qualitative comparison with Kiwi-Edit, Sparkle, and FlowPortal on instruction-guided video background replacement and foreground relighting. WorldPortal produces target-consistent environments and illumination while preserving foreground identity and temporal consistency.

DiT input, z_0^s and $\hat{z}_0^{\text{edit}}$ are injected jointly as two video references: the former specifies the source background domain, and the latter specifies the foreground state to be recast. Together with the null prompt condition, the branch predicts v_t^{recast} and produces a source-domain counterpart of the forward prediction.

The asymmetry between the two branches is designed specifically to compensate for degradation introduced by forward editing. The forward branch combines target semantics with source foreground evidence to create the desired scene and illumination, but this transformation may distort identity, texture, boundaries, or motion details. The recasting branch places the resulting foreground state back into the source background domain, making foreground discrepancies estimable against the aligned source trajectory. It is therefore not an inverse editor or a second generator; instead, it provides a diagnostic prediction that primarily captures these discrepancies. The resulting residual compensates for forward-editing degradation while leaving the intended target background and illumination intact.

4 Experiments

4.1 Implementation Details

We initialize the video DiT with the pretrained parameters of Wan2.2-TI2V-5B (Team Wan et al., 2025) and the VLM condition extractor with those of Qwen3VL-Embedding-8B (Li et al., 2026). The newly introduced semantic, foreground, and background learnable queries, their corresponding learnable null queries, and the lightweight connectors are initialized from scratch. We use 1,024 semantic query tokens and 512 query tokens for each of the foreground and background groups. The VLM condition extractor extracts the output representations at these query positions. Each connector is a two-layer MLP that projects the VLM outputs to the context dimension of the video DiT. We jointly fine-tune the video DiT, learnable queries, null queries, and connectors while adapting the VLM condition extractor with LoRA with rank 64. We train both text-to-video and image-to-video variants, with the image-to-video variant receiving an additional reference-image condition.

At inference, we apply classifier-free guidance (CFG) with a guidance scale of 1.8 only to the forward editing branch, while disabling CFG for the recasting branch to reduce computation. Following the caching strategy of FlowPortal (Gao et al., 2025a), we reuse each recasting prediction for 5 consecutive denoising steps. For a standard 50-step sampling process, CFG requires 100 DiT forward passes. Our recasting branch adds only 10 passes, resulting in 110 passes in total and a extra computation of only 10%. We train our model on the background replacement training set introduced by Sparkle (Zeng et al., 2026), which provides instruction-guided source-target video pairs with lively illumination adaptation. Training takes approximately 96h on 8 NVIDIA A100 GPUs.

Method	CLIP-I ↑	CLIP-T ↑	Detail ↑	Identity ↑	Motion ↑	Structure ↑	VLM-Inst. ↑	VLM-Const. ↑	VLM-Quality ↑
Kiwi-Edit	**0.993**	0.210	31.195	0.598	0.764	0.686	2.629	2.714	2.686
Sparkle	0.989	0.249	29.780	0.462	0.756	0.652	3.899	3.544	3.825
Lumen	0.983	0.228	30.052	0.521	0.783	0.634	3.554	3.760	3.718
FlowPortal	0.987	0.235	32.123	0.689	0.785	0.701	3.781	4.059	3.910
WorldPortal (Ours)	0.989	**0.252**	**33.344**	**0.757**	**0.821**	**0.717**	**4.018**	**4.082**	**3.975**

Table 1: Quantitative comparison on Sparkle-Bench. Higher is better for all metrics.

4.2 Baseline Methods

We compare WorldPortal with four state-of-the-art video editing and relighting methods. **Kiwi-Edit** (Lin et al., 2026) is a general video editing model that supports text- and reference-guided control and serves as a strong baseline for instruction-following video editing. **Sparkle** (Zeng et al., 2026) is a training-based method designed specifically for video background replacement trained from paired data. **Lumen** (Zeng et al., 2025) is also a training-based video generative model trained with realistic and synthetic relighting pairs. **FlowPortal** (Gao et al., 2025a) is a training-free video relighting and background replacement method based on residual-corrected flow, decoupled structural conditions, and high-frequency transfer. These baselines cover both direct data-driven editing and training-free structure-preserving relighting paradigms. We evaluate all methods on Sparkle-Bench (Zeng et al., 2026), which comprises 458 curated source videos with 97 distinct scenes.

4.3 Evaluation Metrics

Following the evaluation protocol of FlowPortal (Gao et al., 2025a), we evaluate instruction alignment, temporal smoothness, and foreground consistency. **Video-text alignment** is measured by CLIP-T, the cosine similarity between the generated frames and the target editing prompt (Qi et al., 2023). **Temporal smoothness** is measured by CLIP-I, computed as the CLIP cosine similarity between consecutive generated frames (Qi et al., 2023).

Foreground consistency is evaluated from 4 complementary aspects. We first extract the foreground region from each output video (Zheng et al., 2024) and then compute: (1) **Structure consistency**, measured by SSIM between the Canny edge maps of the source and edited foregrounds (Zhang et al., 2023); (2) **Detail consistency**, measured by PSNR between the LoG responses (Marr & Hildreth, 1980) of their estimated albedo maps (Careaga & Aksoy, 2024); (3) **Motion consistency**, measured by SSIM between their optical-flow maps (Teed & Deng, 2020); and (4) **Identity consistency**, measured by facial identity similarity (Deng et al., 2019) for human-centric videos and CLIP-I similarity for other subjects.

Following OpenVE-3M (He et al., 2025a), we additionally adopt 3 VLM-based quality evaluation metrics: **Instruction Compliance** (VLM-Inst.), **Consistency & Detail Fidelity** (VLM-Const.), and **Visual Quality & Stability** (VLM-Quality). Each metric is scored by Gemini-2.5-Pro (Google DeepMind, 2025) on a scale from 1 to 5. Higher values indicate better performance.

Table 2: Quantitative ablation study on Sparkle-Bench. Higher is better for all metrics.

Metric	w/o Recast	w/o VLM Cond. & Tokens	**Full Model**
CLIP-I ↑	0.986	0.981	**0.989**
CLIP-T ↑	0.240	0.224	**0.252**
Detail ↑	30.862	28.417	**33.344**
Identity ↑	0.551	0.423	**0.757**
Motion ↑	0.779	0.721	**0.821**
Structure ↑	0.608	0.546	**0.717**

4.4 Comparison with Baseline Methods

As shown in Table 1, WorldPortal achieves the best performance on 8 of the 9 reported metrics. The particularly clear gains in detail and identity consistency support the effectiveness of Recast Residual Correction in recovering foreground information that is degraded during direct editing. WorldPortal obtains a temporal smoothness close to the best score achieved by Kiwi-Edit, while providing sub-

stantially stronger lighting naturalness and foreground preservation. These results indicate that our method improves editing fidelity without materially sacrificing temporal smoothness.

Figure 5 presents representative results under diverse target environments. As a directly trained method, Kiwi-Edit does not strictly preserve foreground structure and fine details, and often exhibits weak illumination adaptation. Sparkle produces more evident environment-dependent lighting, but may repaint the foreground and damage its structural details. FlowPortal better preserves the overall structure and details of the subject, however, its explicit structural constraints can limit the flexibility of the lighting response or produce overly strong illumination effects. In contrast, WorldPortal generates backgrounds and foreground illumination that better follow the target instruction while preserving recognizable facial structure, fine-grained texture, and motion across frames.



Figure 6: Qualitative ablation study of Recast Residual Correction and the VLM condition extractor with adaptive token design. Removing Recast Residual Correction degrades foreground fidelity, while removing the VLM condition extractor and token design leads to weaker foreground-background disentanglement and unstable editing results.

4.5 Ablation Study

We conduct ablation studies to examine the contributions of Recast Residual Correction and the proposed adaptive condition parsing design. As shown qualitatively in Figure 6 and quantitatively in Table 2, removing Recast Residual Correction weakens foreground detail and identity preservation. In this variant, we disable the backward recasting branch and update the trajectory using only the forward editing proposal. Without the VLM condition extractor and its adaptive semantic-, foreground-, and background-centric tokens, the model exhibits poorer instruction alignment and less effective foreground–background disentanglement, resulting in degraded editing quality. This variant retains the original prompt and video-reference conditions as well as Recast Residual Correction, thereby isolating the contribution of adaptive condition parsing.

5 Conclusion

We presented WorldPortal, a unified framework for instruction-guided video background replacement and foreground relighting. At its core, Recast Residual Correction uses a backward recasting branch to place the current edited foreground state into the source background domain, where its discrepancy from the aligned source trajectory largely suppresses the shared background component and primarily captures foreground discrepancies introduced by forward editing. Transferring this residual to the target-domain proposal restores identity, structure, texture, and motion details. Our joint VLM-DiT further provides adaptive semantic-, foreground-, and background-centric conditions for both forward editing and backward recasting, avoiding rigid hand-crafted structural controls. Experiments demonstrate strong instruction alignment, natural illumination adaptation, and foreground fidelity, while introducing only modest inference overhead. We hope WorldPortal provides a useful step toward flexible and reliable video editing.

AI USE STATEMENT

In this work, we used OpenAI Codex to assist with code implementation and literature translation. We used Gemini-2.5-Pro to score videos against predefined evaluation criteria. We have not used generative AI tools to develop research ideas or methods, perform mathematical derivations, organize data, or interpret experimental results. Additionally, we used Codex for manuscript polishing, LaTeX formatting, and drafting these statements. We have reviewed all AI-assisted work. AI-assisted code was manually reviewed and tested, translations were checked against the original texts, and edited prose was reviewed to ensure that its technical meaning was preserved. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

WorldPortal supports creative video editing, but realistic background replacement and relighting can also misrepresent events or facilitate deceptive and non-consensual media manipulation. Responsible use should respect the privacy and consent of depicted individuals, clearly identify edited content, and comply with applicable dataset and model licenses. Biases in training data and automated VLM evaluation may affect performance across subjects and scenes. We discuss technical limitations in Appendix F and encourage evaluation in the intended application context before deployment, consistent with the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

Section 3 presents the model and Recast Residual Correction, while Section 4 describes the training setup, data sources, baselines, and evaluation metrics. Appendix C provides branch-wise training details, learnable-query implementation, the sampling procedure, and inference cost, including Algorithm 1. Appendices D and E report the VLM evaluation prompt, scoring criteria, and ablation configurations. These descriptions provide the methodological details needed to guide independent implementation and evaluation.

REFERENCES

Chris Careaga and Yağız Aksoy. Colorful diffuse intrinsic image decomposition in the wild. *ACM Transactions on Graphics*, 43(6):178:1–178:12, 2024.

Sumit Chaturvedi, Mengwei Ren, Yannick Hold-Geoffroy, Jingyuan Liu, Julie Dorsey, and Zhixin Shu. Synthlight: Portrait relighting with diffusion model by learning to re-render synthetic faces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 369–379, 2025.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4690–4699, 2019.

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, 2025.

Wenshuo Gao, Junyi Fan, Jiangyue Zeng, and Shuai Yang. Flowportal: Residual-corrected flow for training-free video relighting and background replacement, 2025a.

Wenshuo Gao, Xicheng Lan, and Shuai Yang. Anyportal: Zero-shot consistent video background replacement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18990–18999, 2025b.

Google DeepMind. Gemini 2.5 pro model card. Google DeepMind model card, 2025. Updated June 27, 2025. Available at `https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-2-5-Pro-Model-Card.pdf`.

Kaiwen Guo, Peter Lincoln, Philip Davidson, Jay Busch, Xueming Yu, Matt Whalen, Geoff Harvey, Sergio Orts-Escolano, Rohit Pandey, Jason Dourgarian, Danhang Tang, Anastasia Tkach, Adarsh Kowdle, Emily Cooper, Mingsong Dou, Sean Ryan Fanello, Graham Fyffe, Christoph Rhemann, Jonathan Taylor, Paul E. Debevec, and Shahram Izadi. The relightables: Volumetric performance capture of humans with realistic relighting. *ACM Transactions on Graphics*, 38(6):217:1–217:19, 2019.

Haoyang He, Jie Wang, Jiangning Zhang, Zhucun Xue, Xingyuan Bu, Qiangpeng Yang, Shilei Wen, and Lei Xie. Openve-3m: A large-scale high-quality dataset for instruction-guided video editing, 2025a.

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. In *Advances in Neural Information Processing Systems*, volume 38, 2025b.

Haian Jin, Yuan Li, Fujun Luan, Yuanbo Xiangli, Sai Bi, Kai Zhang, Zexiang Xu, Jin Sun, and Noah Snavely. Neural gaffer: Relighting any object via diffusion. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25096–25106, 2024.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Qwen3-VL-Embedding and Qwen3-VL-Reranker: A unified framework for state-of-the-art multimodal retrieval and ranking, 2026.

Yiqi Lin, Guoqiang Liang, Ziyun Zeng, Zechen Bai, Yanzhe Chen, and Mike Zheng Shou. Kiwi-edit: Versatile video editing via instruction and reference guidance, 2026.

Yuanze Lin, Yi-Wen Chen, Yi-Hsuan Tsai, Ronald Clark, and Ming-Hsuan Yang. Illumicraft: Unified geometry and illumination diffusion for controllable video generation. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023.

Yang Liu, Chuanchen Luo, Zimo Tang, Yingyan Li, Yuran Yang, Yuanyong Ning, Lue Fan, Junran Peng, and Zhaoxiang Zhang. Tc-light: Temporally coherent generative rendering for realistic world transfer. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

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.

Yiqun Mei, Mingming He, Li Ma, Julien Philip, Wenqi Xian, David M. George, Xueming Yu, Gabriel Dedic, Ahmet Levent Taşel, Ning Yu, Vishal M. Patel, and Paul Debevec. Lux post facto: Learning portrait performance relighting with conditional video diffusion and a hybrid dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5510–5522, 2025.

Rohit Pandey, Sergio Orts-Escolano, Chloe LeGendre, Christian Häne, Sofien Bouaziz, Christoph Rhemann, Paul E. Debevec, and Sean Ryan Fanello. Total relighting: Learning to relight portraits for background replacement. *ACM Transactions on Graphics*, 40(4):43:1–43:21, 2021.

Pakkapon Phongthawee, Worameth Chinchuthakun, Nontaphat Sinsunthithet, Varun Jampani, Amit Raj, Pramook Khungurn, and Supasorn Suwajanakorn. Diffusionlight: Light probes for free by painting a chrome ball. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 98–108, 2024.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15932–15942, 2023.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6452–6462, 2024.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models, 2025.

Zachary Teed and Jia Deng. RAFT: Recurrent all-pairs field transforms for optical flow. In *Proceedings of the European Conference on Computer Vision*, pp. 402–419. Springer, 2020.

Xiaoyan Xing, Konrad Groh, Sezer Karaoglu, Theo Gevers, and Anand Bhattad. Luminet: Latent intrinsics meets diffusion models for indoor scene relighting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 442–452, 2025.

Chong Zeng, Yue Dong, Pieter Peers, Youkang Kong, Hongzhi Wu, and Xin Tong. Dilightnet: Fine-grained lighting control for diffusion-based image generation. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 73:1–73:11, 2024.

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, 2025.

Ziyun Zeng, Yiqi Lin, Guoqiang Liang, and Mike Zheng Shou. Sparkle: Realizing lively instruction-guided video background replacement via decoupled guidance, 2026.

Longwen Zhang, Qixuan Zhang, Minye Wu, Jingyi Yu, and Lan Xu. Neural video portrait relighting in real-time via consistency modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 802–812, 2021.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3836–3847, 2023.

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 *International Conference on Learning Representations*, 2025.

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*, 3:9150038, 2024.

Yujie Zhou, Jiazi Bu, Pengyang Ling, Pan Zhang, Tong Wu, Qidong Huang, Jinsong Li, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Anyi Rao, Jiaqi Wang, and Li Niu. Light-a-video: Training-free video relighting via progressive light fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13315–13325, 2025.

SUPPLEMENTARY MATERIAL

A OVERVIEW

This supplementary material provides additional qualitative comparisons and implementation details for WorldPortal. We first present more results across diverse subjects and target environments. We then describe the branch-wise training strategy, learnable-query implementation, Recast Residual Correction sampling procedure, and inference cost, together with a stage-aligned comparison to FlowPortal. Finally, we report the complete VLM-based evaluation prompt and scoring criteria used in our experiments, followed by the detailed settings of the two ablation variants evaluated in the main paper and a discussion of failure cases in complex scenes.

B ADDITIONAL COMPARISON RESULTS

Figure 7 presents additional qualitative comparisons across diverse subjects and target environments. WorldPortal consistently generates backgrounds that closely follow the target instructions while adapting the foreground illumination to the new scene. In particular, the resulting lighting exhibits natural changes in brightness, color tone, highlights, and shadows, leading to a more coherent interaction between the foreground subject and the generated environment. Meanwhile, our method better preserves subject identity, facial structure, fine-grained textures, and motion details across frames. These results further demonstrate that WorldPortal achieves a favorable balance between flexible, realistic relighting and faithful foreground preservation.

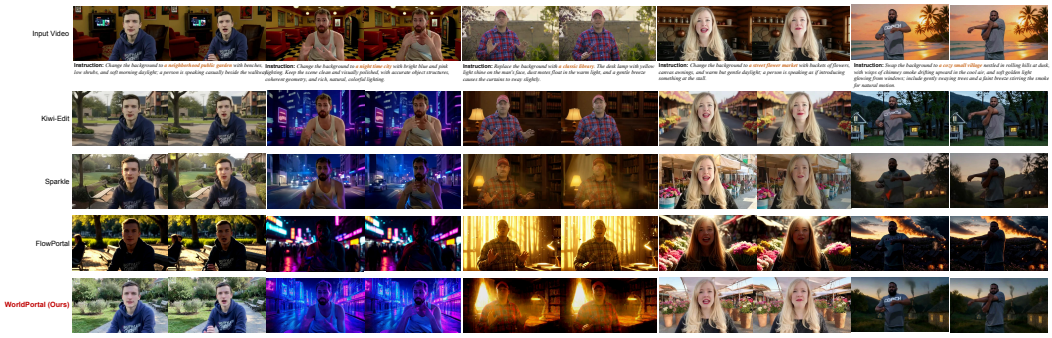

Figure 7: Additional qualitative comparisons with baseline methods on instruction-guided video background replacement and foreground relighting.

C DETAILED IMPLEMENTATION

C.1 BRANCH-WISE TRAINING

We implement the forward editing branch and the backward recasting branch as two operating modes of a shared video DiT. The forward editing branch is trained on instruction-guided source–target video pairs. Given a source video V_s and an instruction p, it predicts a video with the target background and illumination while preserving the source foreground. The backward recasting branch is trained on video triplets. Each triplet contains a target video, a background reference that describes its environment and illumination without an accurate foreground, and a foreground-reference video that shares the target foreground. Specifically, the background reference is a newly generated video that preserves the original background, environment, and illumination of the target video, while its foreground is not required to match the target accurately. The foreground-reference video is taken directly from the corresponding pair in the original video relighting dataset. Because the paired videos share the same foreground structure, this video provides a reliable foreground reference under a different background and illumination. These triplets train the branch to place a given foreground state into a specified background domain, which is the recasting capability used during inference.

We preserve this functional separation through branch-wise conditions. For forward editing, the VLM context consists of semantic-centric tokens parsed from p, foreground-centric tokens parsed from V_s, and learnable null background queries; the DiT additionally receives V_s as its video reference. The semantic tokens describe the requested scene and illumination, the foreground tokens encode the source subject and motion, and learned null-background queries prevent the model from copying the original background through the background-token group. For the backward recasting branch, the VLM context consists of the null prompt, learnable null foreground queries, and background-centric tokens parsed from the background reference. We use a null prompt because the background is already specified by the background reference and therefore does not require an additional textual description. This choice also matches the actual inference setting, in which no textual description of the original background is available. We use learnable null foreground queries because, at each inference step, the foreground information available to the backward recasting branch comes from the dynamically estimated latent $\hat{z}_0^{\text{edit}}$. Unlike an input video, this intermediate latent cannot be directly passed to the VLM for foreground-token extraction. The DiT instead receives the background reference and foreground-reference video as two role-specific video references.

We initialize the video DiT from Wan2.2-TI2V-5B (Team Wan et al., 2025) and the VLM condition extractor from Qwen3VL-Embedding-8B (Li et al., 2026). The extractor uses 1,024 semantic queries, 512 foreground queries, and 512 background queries. A separate two-layer MLP maps each group to the DiT context dimension. The video DiT, learnable queries, null queries, and connectors are jointly optimized, while the VLM is adapted with LoRA using a rank of 64. We use an overall learning rate of 5×10^{-6} for joint optimization. We train both T2V and I2V variants; the I2V variant additionally uses a reference-image condition. Both branches are optimized with the standard diffusion training loss. Training videos retain variable spatial resolutions and temporal lengths, with at most 360,000 pixels per frame and at most 81 frames per video. Training takes approximately 96 hours on eight NVIDIA A100 GPUs.

Learnable query implementation. The learnable queries are trainable vectors inserted into the multimodal input sequence and transformed by the VLM into output feature tokens. For each condition group, we maintain a fixed number of query vectors whose dimension matches the hidden dimension of the VLM. These vectors are optimized jointly with the editing and recasting objectives, and therefore gradually learn how to retrieve the information required by the video generator from the available text, image, and video inputs.

To insert the queries, we append the same number of special placeholder positions to the end of the multimodal conversation sequence. The placeholders determine where the query features reside in the sequence but do not themselves carry the desired conditional semantics. After tokenization and multimodal preprocessing have converted the text, image, and video inputs into a unified sequence, the ordinary vocabulary embeddings at these positions are replaced by the corresponding learnable query vectors. Thus, the Transformer receives directly optimized continuous embeddings at the reserved positions rather than embeddings of ordinary vocabulary tokens.

Through the VLM's self-attention layers, each query can attend to the preceding text, image, and video tokens. The initially semantics-free query vectors are consequently updated into contextualized features that summarize the multimodal information relevant to their assigned role. After the final Transformer layer, we extract the hidden states at the reserved query positions; since the queries are appended to the sequence, these are the final fixed-length block of hidden states. The extracted states constitute the semantic-, foreground-, or background-centric query tokens. Finally, a two-layer nonlinear MLP connector maps each group from the VLM hidden dimension to the context dimension of the video DiT. These projected continuous features, rather than discrete language tokens, are the conditions ultimately supplied to the video generator.

C.2 Recast Residual Correction at Inference

At inference, the forward editing branch and backward recasting branch are executed sequentially at each sampling step. Let z_0^s be the source-video latent and ϵ a global Gaussian noise shared by all trajectories. The aligned source state at noise level t is

$$z_t^s = (1-t)z_0^s + t\epsilon. \tag{14}$$

Algorithm 1 Stage-aligned comparison between WorldPortal and FlowPortal (Gao et al., 2025a). Corresponding rows perform analogous roles in the two sampling procedures.

WorldPortal: Recast Residual Correction	**FlowPortal: Residual Correction**
Input: source latent z_0^s, instruction p, schedule $1 = t_N > \cdots > t_0 = 0$, reuse interval K	**Input:** source latent z_0^s, source condition c_{src}, target condition c_{tar}, mask M, schedule $1 = t_N > \cdots > t_0 = 0$, reuse interval K
1: Sample one $\epsilon \sim \mathcal{N}(0, I)$ and set $z_{t_N} \leftarrow \epsilon$.	**1:** Sample one $\epsilon \sim \mathcal{N}(0, I)$ and set $z_{t_N} \leftarrow \epsilon$.
2: for $i = N, \ldots, 1$ **do**; set $t = t_i$ and $\Delta t = t_i - t_{i-1}$.	**2: for** $i = N, \ldots, 1$ **do**; set $t = t_i$ and $\Delta t = t_i - t_{i-1}$.
3: Construct $z_t^s = (1-t)z_0^s + t\epsilon$ and $z_{t-\Delta t}^s = (1-t+\Delta t)z_0^s + (t-\Delta t)\epsilon$.	**3:** Construct $z_t^s = (1-t)z_0^s + t\epsilon$ and $z_{t-\Delta t}^s = (1-t+\Delta t)z_0^s + (t-\Delta t)\epsilon$.
4: Predict v_t^{edit} from z_t under $c_{\mathrm{edit}}(p, z_0^s)$.	**4:** Predict target velocity $v_t^{\mathrm{tar}}(z_t; c_{\mathrm{tar}})$.
5: Target proposal: $\bar{z}_{t-\Delta t}^{\mathrm{edit}} = z_t - \Delta t\, v_t^{\mathrm{edit}}$.	**5:** Target proposal: $\bar{z}_{t-\Delta t}^{\mathrm{tar}} = z_t - \Delta t\, v_t^{\mathrm{tar}}$.
6: Recover the current edited state $\hat{z}_0^{\mathrm{edit}} = z_t - t v_t^{\mathrm{edit}}$.	**6:** The analytical source state $z_{t-\Delta t}^s$ serves as the source reconstruction target.
7: if $(N-i) \bmod K = 0$ **then**	**7: if** $(N-i) \bmod K = 0$ **then**
8: Cache $v^{\mathrm{recast}} \leftarrow v_t^{\mathrm{recast}}$ predicted under $c_{\mathrm{recast}}(p_{\mathrm{null}}, z_0^s, \hat{z}_0^{\mathrm{edit}})$.	**8:** Cache $v^{\mathrm{src}} \leftarrow v_t^{\mathrm{src}}(z_t^s; c_{\mathrm{src}})$.
9: end if	**9: end if**
10: Form the recast proposal $\bar{z}_{t-\Delta t}^{\mathrm{recast}} = z_t^s - \Delta t\, v^{\mathrm{recast}}$.	**10:** Form the source proposal $\bar{z}_{t-\Delta t}^{\mathrm{src}} = z_t^s - \Delta t\, v^{\mathrm{src}}$.
11: Compute $r_t = z_{t-\Delta t}^s - \bar{z}_{t-\Delta t}^{\mathrm{recast}}$.	**11:** Compute $r_t = z_{t-\Delta t}^s - \bar{z}_{t-\Delta t}^{\mathrm{src}}$.
12: Correct the target proposal: $z_{t-\Delta t} \leftarrow \bar{z}_{t-\Delta t}^{\mathrm{edit}} + r_t$.	**12:** Correct the target proposal: $z_{t-\Delta t} \leftarrow \bar{z}_{t-\Delta t}^{\mathrm{tar}} + M \odot r_t$.
13: No explicit mask or frequency replacement.	**13:** Apply masked high-frequency transfer from $z_{t-\Delta t}^s$ to $z_{t-\Delta t}$.
14: end for; **return** z_0.	**14: end for**; **return** z_0.

Given the current editing trajectory z_t, the forward editing branch predicts v_t^{edit} under $c_{\mathrm{edit}}(p, z_0^s)$ and forms

$$\bar{z}_{t-\Delta t}^{\mathrm{edit}} = z_t - \Delta t\, v_t^{\mathrm{edit}}. \tag{15}$$

Its clean prediction

$$\hat{z}_0^{\mathrm{edit}} = z_t - t\, v_t^{\mathrm{edit}} \tag{16}$$

represents the current edited foreground state together with the target-domain prediction.

The backward recasting branch now changes only the domain in which that state is examined. It uses a null prompt, source-background tokens extracted from V_s, and two video references $[z_0^s, \hat{z}_0^{\mathrm{edit}}]$. Under $c_{\mathrm{recast}}(p_{\mathrm{null}}, z_0^s, \hat{z}_0^{\mathrm{edit}})$, it predicts v_t^{recast} and forms

$$\bar{z}_{t-\Delta t}^{\mathrm{recast}} = z_t^s - \Delta t\, v_t^{\mathrm{recast}}. \tag{17}$$

Both $\bar{z}_{t-\Delta t}^{\mathrm{recast}}$ and $z_{t-\Delta t}^s$ are in the source background domain. Their difference therefore largely suppresses the shared background component and primarily captures foreground discrepancies introduced by forward editing:

$$r_t = z_{t-\Delta t}^s - \bar{z}_{t-\Delta t}^{\mathrm{recast}}. \tag{18}$$

The corrected target update is simply

$$z_{t-\Delta t} = \bar{z}_{t-\Delta t}^{\mathrm{edit}} + r_t. \tag{19}$$

The target proposal supplies the new scene and lighting, whereas the recast residual restores identity, structure, texture, and motion information. Importantly, both terms used to construct the residual lie in the source background domain; the correction is not a direct pixel- or frequency-level copy from the source to the relit target.

C.3 Aligned Comparison with FlowPortal

Algorithm 1 aligns the two inference procedures stage by stage. The shared skeleton is made explicit by matching rows: both methods start from one global noise, construct an analytical source trajectory, produce a target proposal, estimate a source-domain correction, and reuse that correction

for several steps. The decisive difference lies in the conceptual middle of the procedure. FlowPortal measures how far a *source-conditioned generator* deviates from an ideal source reconstruction. WorldPortal instead measures how much foreground information the *current target edit* has lost by recasting that edited state into the source domain.

This alignment makes the relationship precise. Both methods use residual correction to anchor an otherwise unconstrained target generation trajectory, and both reuse additional predictions across sampling steps. FlowPortal's correction, however, is a *source reconstruction residual*: it depends on the ideal source velocity and the source-conditioned velocity, but does not observe the foreground state produced by the current target edit. It therefore relies on an external foreground mask to restrict the correction, structural controls constructed from Canny, HED, and depth, an IC-Light target reference, source/target captions, and masked high-frequency transfer for detail preservation.

This difference also affects the required user input and the robustness of the resulting pipeline. WorldPortal requires only the editing instruction, whereas FlowPortal requires separate textual descriptions of both the source and target videos. Its additional conditions introduce several potential error sources. An inaccurate foreground mask can either leak source content into the generated background or enlarge parts of the foreground area to the background, and applying IC-Light to the first frame can produce an over-saturated target reference that subsequently biases the video generation process. FlowPortal also manually defines the foreground structural condition as a weighted combination of Canny edges, HED boundaries, and depth, while leaving the background region empty. Imposing this hand-crafted condition on a pretrained video model can constrain the foreground geometry too strongly, making it difficult to generate natural lighting responses that alter shadows, highlights, and local contrast. Finally, high-frequency transfer assumes that the fine-scale components of the source foreground remain directly reusable under the target illumination. This assumption is not always valid because specular highlights, cast-shadow boundaries, sharp local contrast, and other high-frequency signals can change substantially with lighting. Directly transferring these components may retain traces of the source illumination, suppress illumination-dependent detail changes, or introduce inconsistencies between the relit foreground and the new environment.

WorldPortal instead uses the backward recasting branch to estimate the foreground discrepancies introduced by forward editing. The branch receives the current clean prediction $\hat{z}_0^{\text{edit}}$ as a video reference and places its foreground state into the source background domain. Comparing the resulting recast prediction with the aligned source trajectory largely suppresses the shared background component and primarily captures foreground discrepancies. Because semantic-, foreground-, and background-centric conditions are learned and activated according to the branch, the algorithm requires neither a foreground mask nor fixed edge/depth controls or Fourier replacement. This removes the assumption that source high frequencies and target lighting must remain directly compatible, leaving the forward editing branch free to generate shadows, highlights, and color shifts appropriate to the new environment.

C.4 Inference Cost

We use 50 sampling steps and a forward CFG scale of 1.8. CFG is disabled for the backward recasting branch to reduce computation. We refresh its prediction every $K = 5$ steps and reuse the resulting residual in between. Standard CFG sampling requires $2 \times 50 = 100$ DiT passes; recasting contributes $50/5 = 10$ passes. The complete sampler therefore uses 110 DiT passes, corresponding to approximately 10% extra computation. Under the same evaluation setting, WorldPortal takes approximately 3 minutes to process one video, compared with 4 minutes for FlowPortal, 4 minutes for Kiwi-Edit, and 6 minutes for Sparkle on a single A100 GPU. WorldPortal therefore achieves the shortest inference time among the compared methods despite performing the additional backward recasting operation.

D VLM-Based Evaluation

Following OpenVE-3M (He et al., 2025a), we use Gemini-2.5-Pro (Google DeepMind, 2025) to evaluate each edited video together with its source video and editing instruction. The evaluation covers three complementary dimensions: **Instruction Compliance** (VLM-Inst.), **Consistency & Detail Fidelity** (VLM-Const.), and **Visual Quality & Stability** (VLM-Quality). Each dimension is

scored independently on a five-point scale, where a higher score indicates better performance. We use the following prompt for all evaluated methods:

> You are a data rater specializing in grading video background editing. You will be given two videos (before and after editing) and the editing instruction. Your task is to evaluate the background change on a 5-point scale from three perspectives:
>
> **Instruction Compliance**
>
> **1.** No change, or background unrelated to prompt, or foreground also replaced/distorted.
>
> **2.** Background partly replaced or wrong style/content; foreground noticeably altered.
>
> **3.** Main background replaced but elements missing/extra, or faint spill onto subject edges.
>
> **4.** Requested background fully present; foreground intact except minute artefacts or small prompt mismatch (e.g., colour tone).
>
> **5.** Background exactly matches prompt (content, style, placement); foreground identity, geometry, texture, and motion are preserved while allowing instruction-consistent illumination changes.
>
> **Consistency & Detail Fidelity**
>
> **1.** Large tearing, posterisation, or significant temporal artifacts like flickering, jittering edges; edit area obvious at a glance.
>
> **2.** Clear cut-out halos, colour-resolution gap, or obvious edge "boiling" (instability) over time.
>
> **3.** Blend acceptable but visible on closer look: slight edge blur, or minor temporal instability (e.g., shimmer).
>
> **4.** Nearly invisible seams; edges are stable across motion, textures aligned, only minor issues when zoomed in.
>
> **5.** Indistinguishable composite: edges, textures, resolution and colour grading are perfectly continuous and stable throughout the video's duration.
>
> **Visual Quality & Stability**
>
> **1.** Severe mismatch: wrong horizon, conflicting light, floating subject, or background remains static during camera movement (no parallax).
>
> **2.** Noticeable inconsistencies in light or scale; incorrect perspective shifts during motion.
>
> **3.** Overall believable; small errors in shadow, perspective, or minor motion tracking flaws.
>
> **4.** Lighting, scale, and depth well matched; background perspective and scale track convincingly with camera motion.
>
> **5.** Physically flawless: foreground and new background share coherent light, shadows, perspective, and atmospheric depth throughout all subject and camera motion, enhancing overall realism.

E Ablation Settings

We evaluate two ablated variants corresponding to the components examined in the main paper. Unless stated otherwise, all variants use the same training data, video DiT initialization, optimization setting, sampling schedule, and forward CFG scale as the full model.

w/o Recast Correction. This variant disables the backward recasting branch during inference and removes the recast residual from every sampling update. Consequently, the editing trajectory is updated using only the proposal predicted by the forward editing branch:

$$z_{t-\Delta t} = \bar{z}^{\text{edit}}_{t-\Delta t}. \tag{20}$$

The forward editing branch, VLM condition extractor, adaptive semantic-/foreground-/background-centric token design, and all other inference settings remain unchanged. Comparing this variant



Figure 8: Failure cases in complex scenes. Challenging target environments with fine-grained spatial structures, reflective surfaces, and substantial illumination changes can lead to distortions in foreground and background details and lower editing quality.

with the full model isolates the contribution of Recast Residual Correction to foreground identity, structure, texture, and motion preservation.

w/o VLM Condition Extractor & Token Design. This variant removes the Qwen3VL-Embedding-8B condition extractor together with the learnable semantic, foreground, and background queries, their learnable null queries, and the associated MLP connectors. The video DiT is trained solely on its original prompt condition and role-specific video references without the additional adaptive token groups. The forward editing branch still receives the source-video reference and editing instruction, while the backward recasting branch still receives the source and current edited-state references required to construct the recast prediction. Recast Residual Correction and the sampling procedure are otherwise retained. This variant isolates the effect of adaptive semantic, foreground, and background condition parsing on instruction alignment, foreground–background disentanglement, and editing stability.

F LIMITATIONS

As shown in Figure 8, WorldPortal may fail in complex scenes that require the synthesis of fine-grained background structures, reflective surfaces, or substantial illumination changes. Under such challenging conditions, the model can introduce distortions in foreground and background details while adapting the subject to the target environment. Improving robustness to complex scene layouts and large appearance changes remains an important direction for future work.