ANYPORTAL: Zero-Shot Consistent Video Background Replacement

Wenshuo Gao Xicheng Lan Shuai Yang✉
Wangxuan Institute of Computer Technology, State Key Laboratory of Multimedia Information Processing,
Peking University, Beijing, China
{gaowenshuo, lanxicheng}@stu.pku.edu.cn williamyang@pku.edu.cn

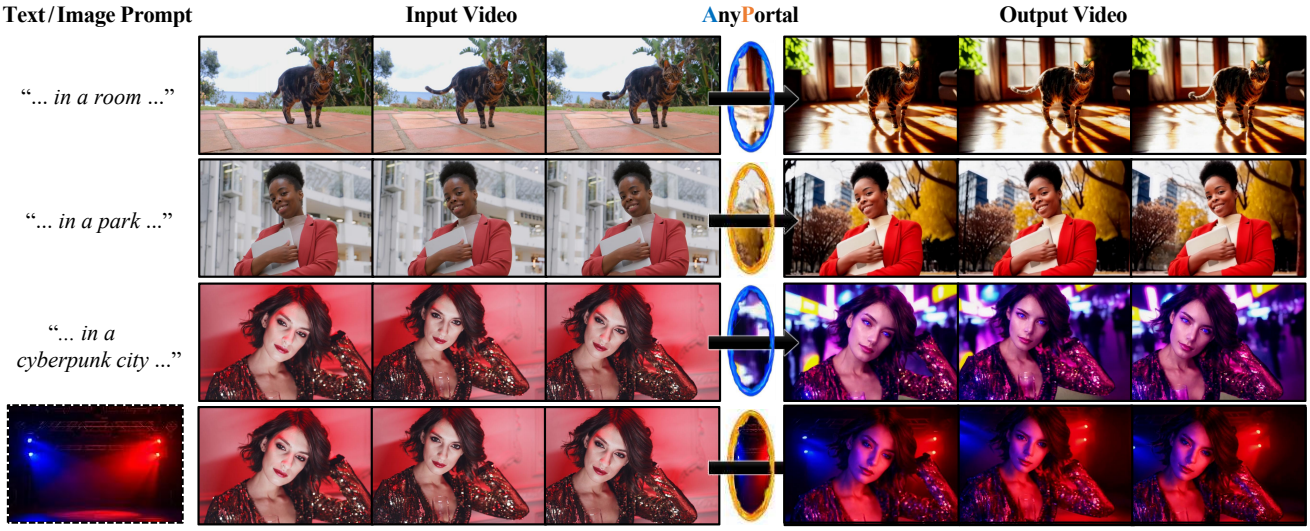


Figure 1. **We propose ANYPORTAL, a training-free framework for high-consistency video background replacement and foreground relighting.** Given an input foreground video and a text or image prompt of the background, our method produces a video with the target background under harmonious illuminations, while maintaining the foreground video details and intrinsic properties.

Abstract

*Despite the rapid advancements in video generation technology, creating high-quality videos that precisely align with user intentions remains a significant challenge. Existing methods often fail to achieve fine-grained control over video details, limiting their practical applicability. We introduce **AnyPortal**, a novel zero-shot framework for video background replacement that leverages pre-trained diffusion models. Our framework collaboratively integrates the temporal prior of video diffusion models with the relighting capabilities of image diffusion models in a zero-shot setting. To address the critical challenge of foreground consistency, we propose a Refinement Projection Algorithm, which enables pixel-level detail manipulation to ensure precise foreground preservation. AnyPortal is training-free and overcomes the challenges of achieving foreground consistency and temporally coherent relighting. Experimental results demonstrate that AnyPortal achieves high-quality results on consumer-grade GPUs, offering a practical and efficient solution for video content creation and editing.*

1. Introduction

Teleportation, often regarded as one of the most popular superpowers, offers the fascinating ability to travel anywhere instantly. While true teleportation remains confined to the realm of science fiction, digital technologies have made its virtual counterpart a reality, particularly in the film and entertainment industry. Through the use of green screens and digital techniques, actors can be seamlessly transported from a studio to virtually any location. However, this process is far from trivial. It involves a complex pipeline that includes constructing green screen environments, generating backgrounds that are geometrically consistent with the camera's perspective, and meticulously replacing the green screen with the synthesized background while ensuring realistic illuminations. Despite its widespread use in professional settings, this workflow remains resource-intensive and labor-expensive beyond normal users.

Recent years have witnessed rapid advancement in AIGC, highlighting the potential to make "virtual teleportation" accessible to the general public. The state-of-the-

art image diffusion model, IC-Light [41], enables users to replace the background of a photo with harmonized illuminations, achieving robust performance through extensive training on paired image datasets. However, collecting large-scale paired video datasets is considerably more difficult compared to paired images, making scaling this approach to video significant challenges. Meanwhile, recent cutting-edge video diffusion models [37, 44] demonstrate impressive capabilities in video generation and editing. Despite their potential, these models still fall short for widespread video background replacement tasks. First, existing video diffusion models exhibit limited controllability over generated content. While some approaches [9, 28, 31] introduce coarse controls for edges, poses and motion, they lack pixel-level precision, often resulting in unintended alterations to the foreground appearance. Second, adapting video models to our specialized task typically requires task-specific training or fine-tuning, which is hindered by the scarcity of paired video data and the substantial computational resources needed to train large video models.

We believe that pre-trained large diffusion models inherently possess rich prior knowledge for video background replacement: IC-Light provides valuable insight into how lighting should be rendered, while video models capture real-world dynamics. Our key insight is to explore *to which extent these pre-trained models can manage tasks that extend beyond their original training task, collaboratively leveraging their inherent priors in a zero-shot setting.*

To this end, we investigate the zero-shot video background replacement problem. While IC-Light excels at illumination harmonization, and video models provide powerful temporal priors, naively combining them fails to address the critical challenge of foreground consistency, which requires precise pixel-level control over the generation process. While mature training-free control schemes exist for image models – such as inference-time optimization [38, 39] or DDIM inversion [30] with latent manipulations [7, 17] – these methods face significant limitations when applied to video models: 1) Optimization on video models incurs prohibitive computational costs; 2) Video models typically operate in a highly compact 3D latent space [37], which degrades inversion quality and hinders detailed manipulations. To address these challenges, we propose a novel **Refinement Projection Algorithm (RPA)** tailored for video models. RPA computes a projection direction in the latent space that simultaneously ensures high consistency with the input foreground details and high-quality background, offering a robust and efficient solution for zero-shot video background replacement.

We introduce **AnyPortal**, a novel training-free framework for video background replacement. AnyPortal first generates a coarse video with illumination harmonized by IC-Light and then enhances its temporal consistency using a pre-trained video diffusion model. To achieve precise control over foreground details, a Refinement Projection Algorithm is proposed to enable pixel-level manipulation. As shown in Fig. 1, AnyPortal seamlessly transfers foreground subjects (*e.g.*, humans or objects) from an input video to a new environment, specified by either a text prompt or a background image, while ensuring natural illuminations, realizing "virtual teleportation" in videos. Remarkably, our framework operates efficiently on a single 24GB GPU. Furthermore, its modular design allows each component to be implemented using the best available pre-trained models, ensuring compatibility with the latest advancements in AIGC. Our contributions are threefold:

- We introduce AnyPortal, an efficient and training-free framework for video background replacement.
- We design a modular pipeline that integrates the latest pre-trained image and video diffusion models, to combine their strengths for realistic and coherent video generation.
- We propose a novel Refinement Projection Algorithm that enables pixel-level detail manipulation in compact latent spaces, ensuring precise foreground preservation.

2. Related Work

Image Diffusion Model. Latent Diffusion Model (LDM) has become a strong method for image generation, notably gaining significant popularity with Stable Diffusion [27]. The main idea is to first compress the image data into latent space using a Variational Autoencoder (VAE), then progressively denoise Gaussian noises with algorithms such as DDPM [13] and DDIM [30], and finally decode the denoised latent back to an image. Traditionally, U-Net architectures were used, followed by DiT [23] that introduces Transformers to improve generated results as in SD3 [5].

To meet user demands for generating images with specific conditions, SDEdit [18] allows for training-free image editing in the LDM framework by denoising the noisy image from an intermediate timestep. Methods like ControlNet [40], T2I-Adapter [20] and ControlNeXt [24] create a learnable branch of the denoising model to offer additional control conditions such as edges and depth maps. Another approach to image editing is using DDIM inversion [30] and Null-Text Inversion [19], which inverts the denoising process of a given image, then re-denoises it with text guidance and attention manipulations [3, 12, 22, 32].

Video Diffusion Model. Many works attempt to extend image diffusion models to video diffusion generation. Early attempts make partial modifications to image models, by redesigning the sampling scheme for zero-shot video generation [15], fine-tuning inflated models for one-shot video generation [33], and training a plug-and-play temporal module to turn image models into animation generators [10]. With increased computational resources, full model training on large-scale video dataset has been pro-

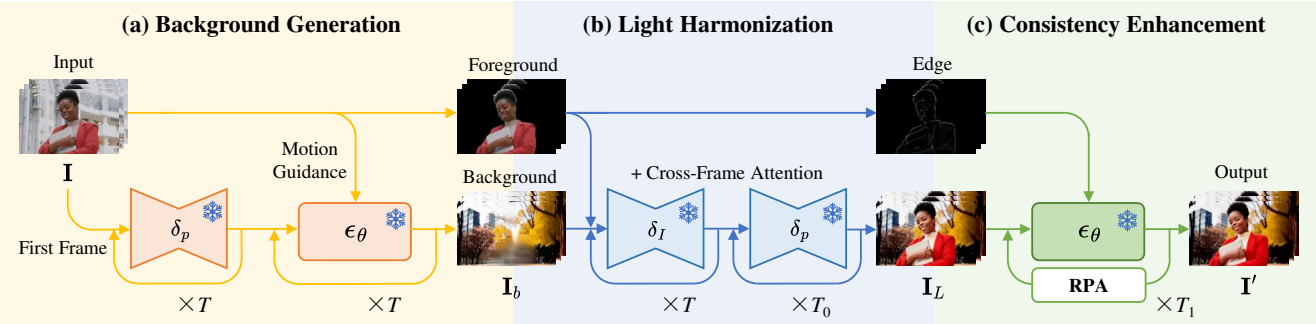


Figure 2. **Framework of AnyPortal.** (a) Background Generation: A video diffusion model ϵ_θ is used to generate a basic background video $\mathbf{I}_b$ following the first frame generated by IC-Light model δ_p and the camera motion of the input video $\mathbf{I}$; (b) Light Harmonization: a two-step pipeline based on IC-Light model δ_I and δ_p is proposed to combine the foreground and background video and harmonize its illumination; (c) Consistency Enhancement: The video diffusion model ϵ_θ is used to improve the temporal consistency of $\mathbf{I}_L$, with a novel Refinement Projection Algorithm (RPA) to further strengthen the foreground consistency with the input $\mathbf{I}$. (❄: All models are frozen)

posed [1, 2, 11, 14, 29, 45]. Modern practices such as OpenSora [44] and CogVideoX [37] have focused on learning in the 3D latent space (2D for spatial + 1D for temporal). These methods typically extend the 2D image VAE to 3D VAE that compresses both spatial and temporal dimensions, mapping videos into a 3D latent space. The denoising model works in this latent space, usually using a DiT architecture [23], which exhibits better temporal consistency and scalability.

Diffusion models can be used for video editing. An intuitive idea is to apply image editing techniques like SDEdit [18] and Prompt-to-Prompt [12] to the image models with cross-frame attention to strengthen temporal consistency [8, 25, 35, 36, 42]. However, image models intrinsically lack modeling of real-world motion, leading to unnatural dynamics. Meanwhile, due to the high complexity and compactness of 3D latent space, the above editing and DDIM inversion techniques [30] are not directly compatible with the DiT-based diffusion models, resulting in motion degradation and appearance distortions. To leverage the latest advancements of video models, we propose an effective Refinement Projection Algorithm to maintain the input video details without harming the generated motions.

Foreground Relighting and Background Replacement. TotalRelighting [21] and SwitchLight [16] train neural networks to predict surface normals and albedo to recompute new lighting. Relightful Harmonization [26] finetunes an image diffusion model conditioned on the background for foreground relighting. IC-Light [41] simultaneously achieves impressive background replacement and illumination harmonization by concatenating the input noise and foreground condition (and optionally background condition) before feeding into the diffusion model and finetuning the model with light transport consistency. Currently, there are few diffusion models specifically for video background replacement and foreground relighting. RelightVid [6] combines IC-Light and AnimateDiff [10] with finetuning. By comparison, our method does not require any training, achieving high compatibility and modularity. Each of the modules can be implemented by the best pretrained models, allowing us to leverage the latest advancements (*e.g.*, CogVideoX [37]) for better video consistency.

3. AnyPortal

3.1. Preliminary

Video Diffusion Model. The latest video diffusion generation models [37, 44] typically includes a 3D VAE $\mathcal{D} \circ \mathcal{E}$ and a denoising DiT model ϵ_θ. The VAE consists of a video encoder $\mathcal{E}$ to encode a video clip $\mathbf{I}$ into the compressed latent feature $x = \mathcal{E}(\mathbf{I})$ and a video decoder to the latent back into a video $\mathbf{I} = \mathcal{D}(x)$. The denoising DiT model ϵ_θ is trained for denoising in the latent space. At the timestep t, ϵ_θ takes as input noisy latent x_t and condition c (typically, the prompt) to output a noise prediction $\epsilon_\theta(x_t, c, t)$. A clean x_0 could be sampled from a Gaussian noise $x_T \sim \mathcal{N}(0, 1)$ by iteratively predicting x_{t-1} from x_t following denoising schemes such as DDIM [30],

$$x_{t-1} = \sqrt{\alpha_{t-1}} x_0^t + \sqrt{1 - \alpha_{t-1}} \epsilon_\theta(x_t, c, t), \tag{1}$$

where αs are a group of parameters related to t and x_0^t is the denoised latent at timestep t,

$$x_0^t = \frac{x_t - \sqrt{1 - \alpha_t} \epsilon_\theta(x_t, c, t)}{\sqrt{\alpha_t}}. \tag{2}$$

Finally, the generated video $\mathbf{I} = \mathcal{D}(x_0)$ is obtained.

IC-Light. IC-Light [41] is an image diffusion model for background replacement and foreground relighting. It has two versions δ_p and δ_I. δ_p is conditioned on prompts p describing the appearance of the background, while δ_I is conditioned on background images I_b. Given a foreground image I_f, it generates $I' = \delta_p(I_f, p)$ or $I' = \delta_I(I_f, I_b)$ with

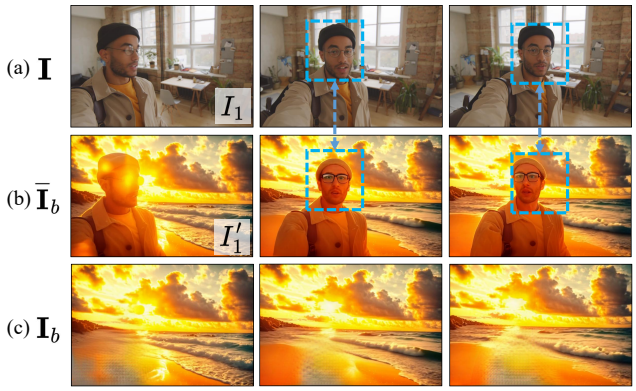


Figure 3. **Motion-aware background generation.** The generated background video $\mathbf{I}_b$ follows the camera motion of the input video. (a) Input video $\mathbf{I}$. (b) Video output $\bar{\mathbf{I}}_b$. (c) Inpainted output $\mathbf{I}_b$. The blue dotted lines indicate the inconsistent foreground areas.

the corresponding foreground and background under harmonized illumination. For simplicity, we omit the iterative sampling operations and transformations between the image space and the 2D latent space. We experimentally find that the text-guided model excels at intensively harmonized illumination, whereas the image-guided model can create more consistent background. We will detail how we combine the two models to leverage their strengths in Sec. 3.2.2 .

3.2. Zero-Shot Video Background Replacement

As illustrated in Fig. 2, our framework is divided into three stages: (1) Background Generation; (2) Light Harmonization; (3) Consistency Enhancement. Our input is a foreground video $\mathbf{I}$ and a prompt p describing the background. In the first stage, we generate a background video $\mathbf{I}_b$ that matches the camera movements of $\mathbf{I}$ with the help of a pre-trained video diffusion model ϵ_θ. The second stage harmonizes the lighting of the foreground object in the new background based on our proposed two-step IC-Light pipeline to produce a coarse video $\mathbf{I}_L$. The third stage introduces a novel Refinement Project Algorithm (RPA) that addresses inconsistencies between frames and refines the foreground details to match those of $\mathbf{I}$, yielding the final video $\mathbf{I}'$.

Note that our method is zero-shot and modular, fully leveraging the powerful generative ability of the pre-trained diffusion models ϵ_θ and δs without any training or inference-time optimization. This allows us to generate impressive videos on a single 24GB-memory GPU. Moreover, our method fully benefits from rapidly growing vision diffusion research, as it can be implemented on the latest pre-trained models once available to boost the performance.

3.2.1. Background Generation

To seamlessly integrate the foreground with the background, the first stage produces a basic background video $\mathbf{I}_b$ that corresponds with the background prompt p and, crucially, matches the camera motion of the $\mathbf{I}$. To this end,

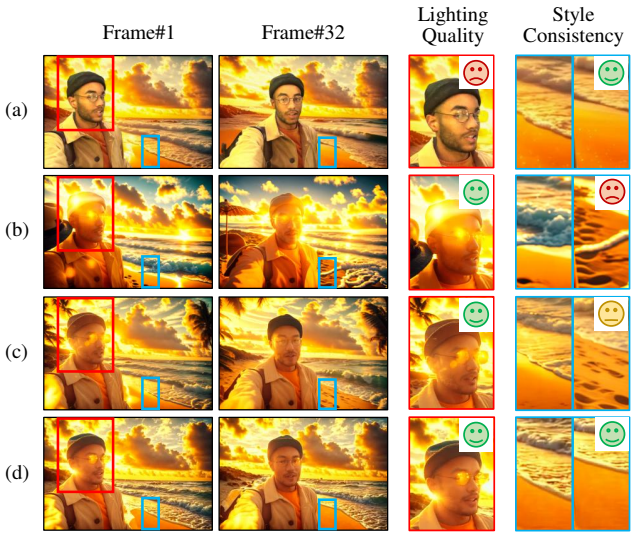


Figure 4. **Two-step light harmonization.** Our two-step harmonization pipeline with cross-frame attention enables high lighting quality (enlarged red region) and inter-frame style consistency (enlarged blue region). Light harmonization results of (a) δ_I, (b) δ_p, (c) $\delta_I + \delta_p$, (d) $\delta_I + \delta_p$+ cross-frame attention.

we follow Diffusion-As-Shader (DAS) [9], a ControlNet-based video generation framework that guides the video diffusion model with the first frame and the motion (tracked 3D points) of a guiding video. Specifically, $\mathbf{I}$ serves as the guiding video. To obtain the first frame, we apply IC-Light to the first frame I_1 of $\mathbf{I}$, resulting in $I_1' = \delta_p(I_1, p)$. Then, we apply DAS to the backend video diffusion model to generate $\bar{\mathbf{I}}_b$ based on I_1' and $\mathbf{I}$. As shown in Fig. 3, $\bar{\mathbf{I}}_b$ has the same camera motion as the input $\mathbf{I}$, but its foreground object may differ significantly from $\mathbf{I}$ and cannont be directly used as our video background replacement result. Finally, we use ProPainter [46] to remove the foreground object, obtaining the basic background video $\mathbf{I}_b$.

3.2.2. Light Harmonization

We first extract foreground objects $\mathbf{I}_f$ from $\mathbf{I}$ with an image segmentation model BiRefNet [43]. Now, for each frame of $\mathbf{I}'$, we have a background prompt p, a background image $I_b \in \mathbf{I}_b$ and a foreground image $I_f \in \mathbf{I}$. We have tried applying both the image-guided and text-guided IC-Light to combine them. However, neither produces reasonable results. As shown in Figs. 4(a)(b), the image-guided result $I' = \delta_I(I_f, I_b)$ has an insufficiently harmonized illumination (missing backlight effect), while the text-guided result $I' = \delta_p(I_f, p)$ has a strong illumination effect and suffers from temporal inconsistency and mismatched camera motions in the background due to the lack of image guidance.

To strike a balance of the illumination effect between δ_I and δ_p while simultaneously utilizing the temporally coherent visual guidance from $\mathbf{I}_b$, we propose a two-step harmonization pipeline, as illustrated in Fig. 2(b). In the first step,

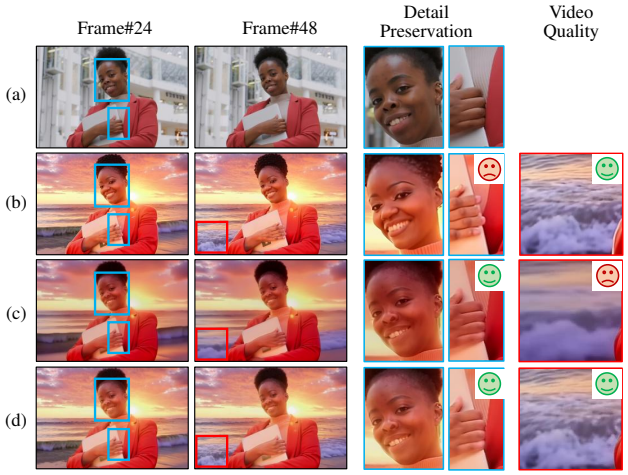


Figure 5. **Consistency enhancement via Refinement Projection Algorithm (RPA).** (a) Input video. (b) Video diffusion model strengthens temporal consistency, but creates inconsistent foreground appearance. (c) Foreground refinement without RPA introduces quality degradation in the background. (d) RPA effectively constrains foreground details while preserving other regions.

we obtain the image-guided result $I'_{img} = \delta_I(I_f, I_b)$. In the second step, we take the idea of SDEdit [18] to refine the illumination of I'_{img} by denoising it using δ_p. In particular, we add noise of T_0 steps ($T_0 < T$) to I'_{img} with DDPM forward process [13], which is then denoised for T_0 steps using δ_p under the conditions of I_f and p. As shown in Fig. 4(c), the foreground lighting is well enhanced. However, such per-frame processing cannot ensure style consistency (*e.g.*, the inconsistent appearance of beaches and the varied position of palm leaves in two frames). To alleviate this issue, we employ cross-frame attention [15, 33, 35] to δ_I and δ_p. We replace δ's self-attention layers with cross-frame attention layers, where all frames aggregate key and value features from the first frame rather than themselves. As a result, the style consistency is strengthened, as shown in Fig. 4(d). Note that our pipeline allows one to adjust the illumination effect via T_0, *i.e.*, a large T_0 produces results with intensive lights and shadows.

3.2.3. Consistency Enhancement

The video $\mathbf{I}_L$ generated in Sec. 3.2.2 still presents two issues: 1) Even introducing cross-frame attention for global style consistency, there are still pixel-level jitters between frames. 2) $\mathbf{I}_L$'s foreground details do not exactly match $\mathbf{I}$'s foreground details. Therefore, we aim to take advantage of the ability of the video diffusion model ϵ_θ to improve both the inter-frame temporal continuity and the foreground detail consistency. As with Sec. 3.2.2, our high-level idea is to use SDEdit to refine the temporal consistency of $\mathbf{I}_L$ by denoising it using ϵ_θ for T_1 steps ($T_1 < T$). We additionally apply edge-based ControlNet to ϵ_θ to preserve the main structure of $\mathbf{I}$, as shown in Fig. 2(c). However, ControlNet only provides coarse structure guidance, failing to maintain the identity in Fig. 5(b). Thus, we propose a **Refinement Projection Algorithm (RPA)** to enforce the consistency of the foreground details at the pixel level.

Algorithm 1: Foreground Refinement

Input: Edited video $\mathbf{I}_0^t$, original input video $\mathbf{I}$
Output: Refined video $\tilde{\mathbf{I}}_0^t = \text{Refine}(\mathbf{I}_0^t, \mathbf{I})$

$\mathbf{I}_{0,\text{LF}}^t = \text{GaussianBlur}(\mathbf{I}_0^t)$;
$\mathbf{I}_{0,\text{HF}}^t = \mathbf{I}_0^t - \mathbf{I}_{0,\text{LF}}^t$;
$\mathbf{I}_{\text{LF}} = \text{GaussianBlur}(\mathbf{I})$;
$\mathbf{I}_{\text{HF}} = \mathbf{I} - \mathbf{I}_{\text{LF}}$;
$\mathbf{M}_0^t = \text{ForegroundSegmentation}(\mathbf{I})$;
$\mathbf{I}_{\text{BG}} = \text{Inpaint}(\mathbf{I}_0^t, \text{ForegroundSegmentation}(\mathbf{I}_0^t))$;
$\tilde{\mathbf{I}}_0^t = \mathbf{M}_0^t \cdot (\mathbf{I}_{\text{HF}} + \mathbf{I}_{0,\text{LF}}^t) + (\mathbf{1} - \mathbf{M}_0^t) \cdot \mathbf{I}_{\text{BG}}$;

The key idea is to transfer the high-frequency details (since high-frequency information of a frame typically depicts its edges and textures, while low-frequency information characterizes its colors and illuminations) from $\mathbf{I}$ to $\mathbf{I}_L$ in the foreground during SDEdit denoising. As analyzed in Sec. 1, the compact 3D latent space impedes direct high-frequency refinement in the pixel domain. Thus, we first decode the latent back to the pixel domain to apply the refinement, and then encode the refined video back to the latent space. To avoid quality degradation from the inherent reconstruction error of the 3D VAE, RPA computes a zero-error projection direction to guide the encoding. Specifically, RPA has two parts: foreground refinement and DDIM denosing with RPA.

Foreground Refinement. To avoid the interference of noises, we follow common practice [35] to operate on noise-free latent x_0^t in Eq. (2). x_0^t is first decoded back to video $\mathbf{I}_0^t = \mathcal{D}(x_0^t)$. Subsequently, $\mathbf{I}_0^t$ and $\mathbf{I}$ are decomposed into their low-frequency (LF) and high-frequency (HF) components. In the foreground region of the refined video $\tilde{\mathbf{I}}_0^t$, we combine $\mathbf{I}_0^t$'s LF component and $\mathbf{I}$'s HF component. The background region of $\tilde{\mathbf{I}}_0^t$ is set to the inpainted $\mathbf{I}$, which removes the foreground object using ProPainter [46]. The refinement details is summarized in Algorithm 1.

DDIM Denosing with RPA. We would like to re-encode the refined video $\tilde{\mathbf{I}}_0^t$ back into latent space as $\hat{x}_0^t$, to replace the original x_0^t during DDIM denoising. Ideally, apart from refined HF details, $\hat{x}_0^t$ should remain unchanged compared to x_0^t. However, there are two places where errors could be introduced. First, encoding and decoding is not strictly reversible; second, the stochastic nature of VAE causes inevitable discrepancies. Actually, VAE outputs the mean and standard deviation of the latent: $\hat{\mu}, \hat{\sigma} = \mathcal{E}(\tilde{\mathbf{I}}_0^t)$, and $\hat{x}_0^t$ is sampled by reparameterization: $\hat{x}_0^t = \hat{\mu} + \epsilon\hat{\sigma}$ with $\epsilon \sim N(0, 1)$, which introduces randomness. As the DDIM denoising iterates, such randomness and errors accumulate, resulting in a blurred background as in Fig. 5(c).

Instead of a random ϵ, our RPA uses a deterministic $\hat{\epsilon}$

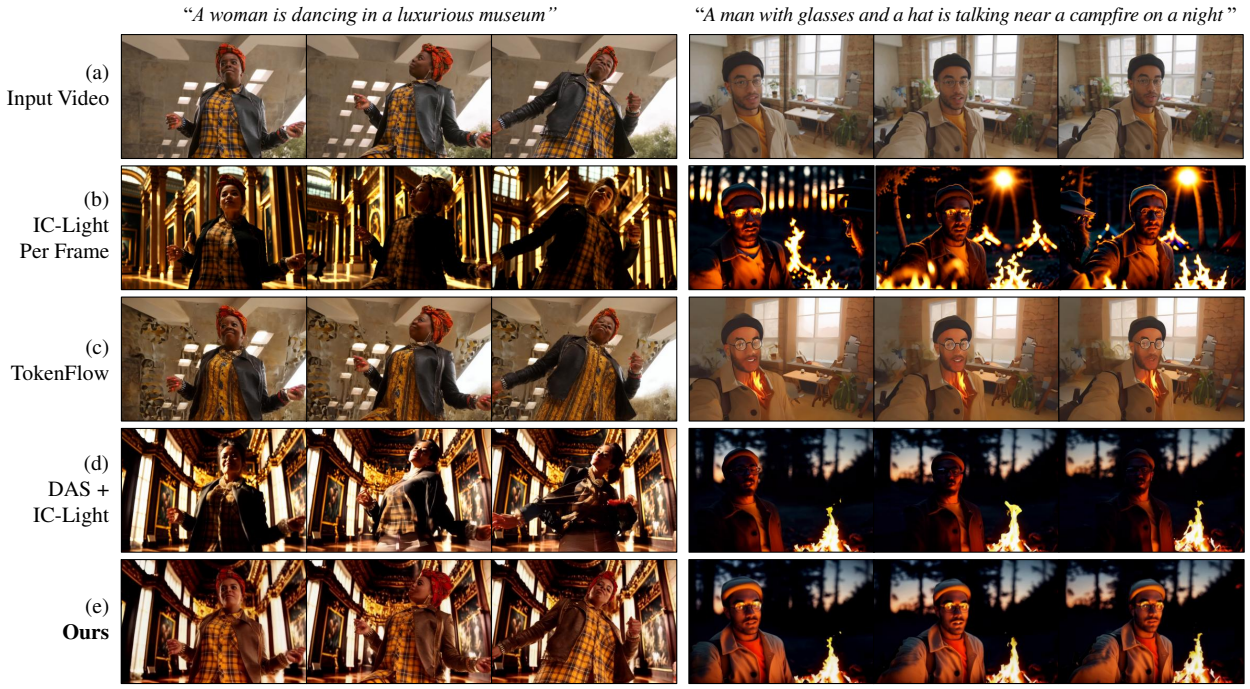


Figure 6. **Visual comparison on zero-shot video background replacement.** Full video results are provided in the supplementary material.

Algorithm 2: DDIM Denoising with RPA

Input: Initial noise x_{T_1}, input video $\mathbf{I}$, condition c
Output: Refined denoised result x_0

for $t = T_1 \dots 1$ **do**
- $x_0^t = \big(x_t - \sqrt{1-\alpha_t}\epsilon_\theta(x_t, c, t)\big)/\sqrt{\alpha_t}$;
- $\mathbf{I}_0^t = \mathcal{D}(x_0^t)$;
- $\mu, \sigma = \mathcal{E}(\mathbf{I}_0^t)$;
- $\tilde{\mathbf{I}}_0^t = \text{Refine}(\mathbf{I}_0^t, \mathbf{I})$;
- $\hat{\mu}, \hat{\sigma} = \mathcal{E}(\tilde{\mathbf{I}}_0^t)$;
- $\hat{\epsilon} = (x_0^t - \mu)/\sigma$;
- $\hat{x}_0^t = \hat{\mu} + \hat{\epsilon}\hat{\sigma}$;
- $x_{t-1} = \sqrt{\alpha_{t-1}}\hat{x}_0^t + \sqrt{1-\alpha_{t-1}}\epsilon_\theta(x_t, c, t)$;

that is computed to ensure a perfect reconstruction on x_0^t. Specifically, we assume a perfect reconstruction $\mu + \hat{\epsilon}\sigma = x_0^t$. Note that x_0^t and $\mu, \sigma = \mathcal{E}(\mathcal{D}(x_0^t))$ are all available, so we can deterministically calculate $\hat{\epsilon}$. Then, for the refined video $\tilde{\mathbf{I}}_0^t = \text{Refine}(\mathbf{I}_0^t, \mathbf{I})$, we obtain $\hat{\mu}, \hat{\sigma} = \mathcal{E}(\tilde{\mathbf{I}}_0^t)$ and the final projection solution is $\hat{x}_0^t = \hat{\mu} + \hat{\epsilon}\hat{\sigma}$. We summarize our proposed RPA in Algorithm 2. Our key insight is that if no refinement is applied (*i.e.*, $\tilde{\mathbf{I}}_0^t = \mathbf{I}_0^t$), this projection will cause $\hat{x}_0^t$ exactly equal to x_0^t. Such an alignment property ensures the resulting video's background area remains almost identical with only foreground details refined, which is also verified by Fig. 5(d).

4. Experiments

Implementation Details. We instantiate AnyPortal with CogVideoX [37] as the video diffusion model ϵ_θ, and IC-Light [41] as the image background replacement model δ_p and δ_I. We set $T = 20$, (T_0, T_1) to $(0.7T, 0.7T)$ and $(0.4T, 0.4T)$ for strong and weak illumination effects, respectively, due to different scenario needs. All experiments are conducted on a single NVIDIA 4090 GPU with CPU offload activated for CogVideoX. The testing videos are uniformly resized to 480×720 and trimmed to 49 frames to comply with CogVideoX specifications. Each video requires approximately 12 minutes for inference (which can be further accelerated with CPU offload off if larger GPU memory is available). The code of this work will be released along with the publication of this paper.

Baseline. Since there are few other works exactly handling our zero-shot video background replacement task, we choose the following most related baselines for comparison.

- IC-Light [41]: A state-of-the-art image background replacement model. We apply it frame-by-frame.
- TokenFlow [8]: A state-of-the-art zero-shot text-guided video editing model.
- Diffusion-As-Shader (DAS) [9]: A versatile video generation control model. We use its motion transfer function, which creates a new video by transferring motion from an input video to a provided image as the first frame. Here, we use IC-Light to generate the first frame.

Note that all above baselines are zero-shot diffusion-based editing methods to ensure a fair comparison.

Evaluation. We construct a test set consisting of 30 samples and prompts for evaluation, and use the following metrics for evaluation: 1) Fram-Acc [25]: The proportion of video frames where the CLIP-based cosine similar-

Table 1. Quantitative comparison and user preference rates

Metric	IC-Light	TokenFlow	DAS	Ours
Fram-Acc ↑	**0.983**	0.541	0.937	0.973
Tem-Con ↑	0.945	0.981	0.986	**0.993**
ID-Psrv ↓	0.578	0.632	0.364	**0.313**
Mtn-Psrv ↑	0.844	0.985	0.878	**0.987**
User-Pmt	1.11%	1.11%	29.72%	**68.06%**
User-Tem	0.56%	5.56%	28.61%	**65.28%**
User-Psrv	2.78%	18.33%	17.22%	**61.67%**
User-Lgt	11.11%	11.11%	30.56%	**47.22%**

ity with the target prompt is higher than that with the source prompt, to measure whether the background is successfully edited. 2) Tem-Con [25]: CLIP-based cosine similarity between consecutive frames to measure temporal consistency; 3) ID-Psrv: Preservation of foreground detail of the generated video, measured by the identity loss [4] between the human face (if applicable) in generated video and the input video; 4) Mtn-Psrv: Preservation of the motion of the generated video, measured by point motion tracking similarity between the generated video and the input video. We use SpatialTracker [34] to track points.

For user study, we invite 24 participants. Participants are asked to select the best results among the four methods based on three criteria: 1) User-Pmt: how well the result aligns with the prompt, 2) User-Tem: the temporal consistency of the result, 3) User-Psrv: how well the foreground details and motions are preserved, and 4) User-Lgt: the quality of relighting on foreground.

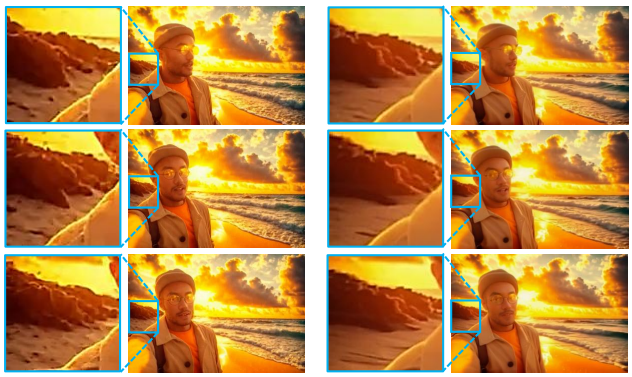
4.1. Comparison to State-of-the-Art Methods

Figure 6 visually compares the proposed method with other baselines. IC-light [41], being fundamentally an image diffusion model, inherently suffers from temporal inconsistency. Moreover, it tends to overly relight the subject, even changing the intrinsic properties like the color of the clothes and the headscarf. TokenFlow [8] demonstrates limited editing capabilities and insufficient foreground detail control, while DAS [9] fails to maintain control over foreground motion dynamics and intrinsic appearance properties. In contrast, our method achieves high-quality background replacement and foreground relighting while ensuring temporal consistency and foreground detail consistency. Full results are provided in the supplementary material.

Table 1 gives quantitative evaluations. IC-Light achieves the best Fram-Acc as it is specifically trained for this background replacement task, without the need to consider temporal consistency. Our method achieves the second-best Fram-Acc, and the best results across all other metrics and user preferences, striking a good balance between single-frame relighting quality and overall video smoothness.

(a) $T_0 = 0$ (b) $T_0 = 0.4T$ (c) $T_0 = 0.7T$

Figure 7. **Effect of T_0 in illumination harmonization.** With increased T_0, δ_p assumes a more prominent role in facilitating intensive light and shadow effect for foreground.



(a) $\mathbf{I}_L$ (b) $\mathbf{I}'$

Figure 8. **Effect of Consistency Enhancement.** The temporal prior of video diffusion model effectively helps improve temporal consistency, *e.g.*, eliminating the inconsistent appearance of rocks as in the enlarged blue region.

4.2. Ablation Study

To validate the contributions of different modules to the overall performance, we systematically deactivate specific modules in our framework. The results are reported in Figs. 4, 5, 7, 8 and Table 2.

- **Two-Step Harmonization**. In the Image Harmonization stage, we employ IC-Light δ_I and δ_p to provide improved illumination for videos, enabling better integration of the foreground into the background. A naive one-step harmonization (*i.e.*, w/o δ_p or $T_0 = 0$) would result in video models generating foregrounds with less natural lighting, as reported in Table 2. As T_0 increases, δ_p gradually plays a stronger role in achieving more natural foreground lighting, as shown in Fig. 7.
- **Cross-Frame Attention**. The effect of cross-frame attention injection is studied in Fig. 4. Disabling cross-frame attention leads to severe inter-frame appearance discrepancy in the generated results (*e.g.*, footprints suddenly appeared on the beach), degrading temporal consistency.
- **Temporal Prior**. The Consistency Enhancement (Cst-Enh) stage optimizes the foreground details and overall temporal consistency of the video $\mathbf{I}_L$ generated in the sec-

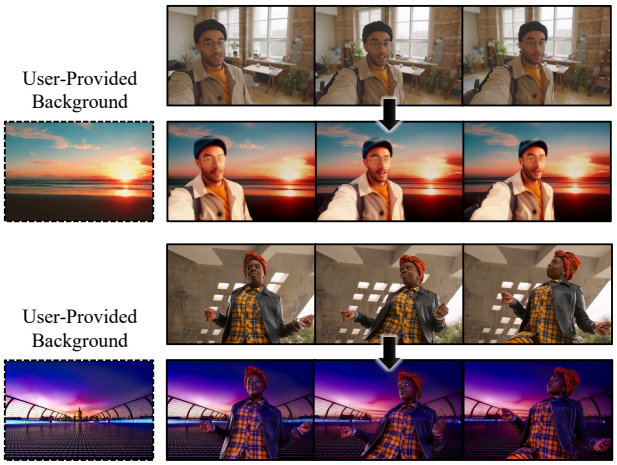


Figure 9. **Image-guided background replacement.** In addition to text prompts, our method also supports replacing the background based on a user-provided image.

ond stage. Without Cst-Enh, Tem-Con drops significantly as in Table 2. As demonstrated in Fig. 8, $\mathbf{I}_L$ suffers from inconsistent rock in the background. By leveraging the strong prior of video diffusion models, this issue is effectively solved in the stage-three result $\mathbf{I}'$.

- **RPA**. RPA performs high-frequency detail refinement on the latent. Without RPA, the identity discrepancy becomes greater as in Table 2 and Fig. 5(b). A naive high-frequency detail refinement through decoding and encoding leads to a blurry background, as in Fig. 5(c). Our designed RPA provides a deterministic sampling scheme that well preserves the non-refined regions like background areas, as in Fig. 5(d).

5. More Results

Image-Guided Video Background Replacement. Our method can be easily adapted to image prompts. In the first stage, we generate the first frame with δ_I and a user-provided background scene image, while the subsequent stages remain identical to those with text prompts. Our image-guided results are shown in Fig. 1 and Fig. 9.

Comparison to Light-A-Video. We further provide a visual comparison with a concurrent work of Light-A-Video [47] in Fig. 10. The two methods, both based on CogVideoX, produce outputs of comparable quality. However, the CogVideoX implementation of Light-A-Video can only relight the existing background, while our method can generate new background content.

6. Limitations

While AnyPortal demonstrates promising results, several limitations remain. Figure 11 gives a typical example. 1) Low-quality inputs (*e.g.*, low-resolution/blurry) reduce high-frequency detail transfer, causing blurred results like

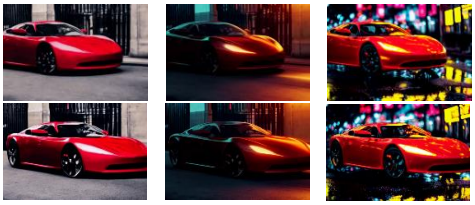


Figure 10. **Comparison to Light-A-Video.** As concurrent work, Light-A-Video only relights backgrounds instead of generating new content.

Table 2. Quantitative ablation study

Metric	w/o δ_p	w/o Cst-Enh	w/o RPA	Full
Fram-Acc ↑	0.966	0.970	0.970	**0.973**
Tem-Con ↑	0.989	0.961	0.987	**0.993**
ID-Psrv ↓	0.329	0.353	0.371	**0.313**
Mtn-Psrv ↑	**0.987**	0.973	0.984	**0.987**

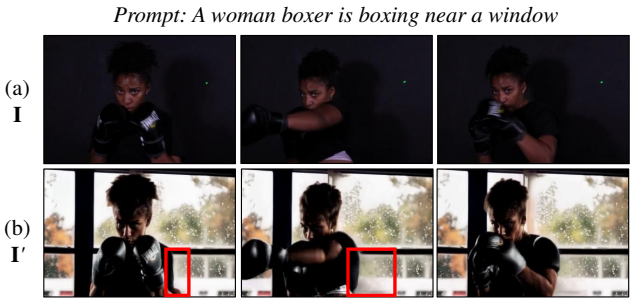


Figure 11. **Limitations.** Low-quality input, poor boundary condition and fast movement degrades the performance of our method. Red boxes indicate the mismatched and blurry boundary region.

Fig. 11's hair; 2) Unclear foreground-background boundaries lead to mismatched inpainting and enlarged blurry regions around subjects; 3) Rapid motion challenges diffusion models, causing artifacts on the left arm.

7. Conclusion and Discussion

In this paper, we propose AnyPortal, a zero-shot framework for video background replacement and foreground relighting that achieves high temporal consistency and detail fidelity without task-specific training. Specifically, by integrating motion-aware video diffusion for background generation, extending image relighting models with cross-frame attention, and introducing the Refinement Projection Algorithm to preserve foreground details in latent space, our method outperforms existing approaches in both lighting harmonization and temporal coherence.

One possible future direction is to investigate the extension of diverse editing tasks (*e.g.*, recolorization, stylization, facial attribute editing, inpainting) to video domains with the temporal prior of large video diffusion models.

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China under Grant 62471009, in part by CCF-Tencent Rhino-Bird Open Research Fund, 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] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 3

[2] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *CVPR*, pages 22563–22575, 2023. 3

[3] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. MasaCtrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *ICCV*, 2023. 2

[4] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, pages 4690–4699, 2019. 7

[5] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. 2024. 2

[6] 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. 3

[7] Xiang Gao and Jiaying Liu. Fbsdiff: Plug-and-play frequency band substitution of diffusion features for highly controllable text-driven image translation. pages 4101–4109, 2024. 2

[8] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv preprint arXiv:2307.10373*, 2023. 3, 6, 7

[9] Zekai Gu, Rui Yan, Jiahao Lu, Peng Li, Zhiyang Dou, Chenyang Si, Zhen Dong, Qifeng Liu, Cheng Lin, Ziwei Liu, et al. Diffusion as shader: 3d-aware video diffusion for versatile video generation control. *arXiv preprint arXiv:2501.03847*, 2025. 2, 4, 6, 7

[10] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 2, 3

[11] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity video generation with arbitrary lengths. *arXiv preprint arXiv:2211.13221*, 2022. 3

[12] 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*, 2022. 2, 3

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, pages 6840–6851, 2020. 2, 5

[14] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 3

[15] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *ICCV*, pages 15954–15964, 2023. 2, 5

[16] 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. 3

[17] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *CVPR*, pages 11461–11471, 2022. 2

[18] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 2, 3, 5

[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. 2

[20] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *AAAI*, pages 4296–4304, 2024. 2

[21] 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. 3

[22] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. In *ACM SIGGRAPH Conference Proceedings*, pages 1–11, 2023. 2

[23] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, pages 4195–4205, 2023. 2, 3

[24] Bohao Peng, Jian Wang, Yuechen Zhang, Wenbo Li, Ming-Chang Yang, and Jiaya Jia. Controlnext: Powerful and efficient control for image and video generation. *arXiv preprint arXiv:2408.06070*, 2024. 2

[25] 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. 3, 6, 7

[26] 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. 3

[27] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 2

[28] Dian Shao, Mingfei Shi, Shengda Xu, Haodong Chen, Yongle Huang, and Binglu Wang. FinePhys: Fine-grained human action generation by explicitly incorporating physical laws for effective skeletal guidance. In *CVPR*, pages 1905–1916, 2025. 2

[29] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-A-Video: Text-to-video generation without text-video data. In *ICLR*, 2023. 3

[30] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2, 3

[31] TheDenk. Cogvideox controlnet extension, 2023. Accessed: 2025-02-12. 2

[32] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *CVPR*, pages 1921–1930, 2023. 2

[33] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *ICCV*, pages 7623–7633, 2023. 2, 5

[34] Yuxi Xiao, Qianqian Wang, Shangzhan Zhang, Nan Xue, Sida Peng, Yujun Shen, and Xiaowei Zhou. Spatialtracker: Tracking any 2d pixels in 3d space. In *CVPR*, pages 20406–20417, 2024. 7

[35] Shuai Yang, Yifan Zhou, Ziwei Liu, and Chen Change Loy. Rerender a video: Zero-shot text-guided video-to-video translation. In *SIGGRAPH Asia Conference*, pages 1–11, 2023. 3, 5

[36] Shuai Yang, Yifan Zhou, Ziwei Liu, and Chen Change Loy. Fresco: Spatial-temporal correspondence for zero-shot video translation. In *CVPR*, pages 8703–8712, 2024. 3

[37] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 2, 3, 6

[38] Haotian Ye, Haowei Lin, Jiaqi Han, Minkai Xu, Sheng Liu, Yitao Liang, Jianzhu Ma, James Zou, and Stefano Ermon. Tfg: Unified training-free guidance for diffusion models. *arXiv preprint arXiv:2409.15761*, 2024. 2

[39] Jiwen Yu, Yinhuai Wang, Chen Zhao, Bernard Ghanem, and Jian Zhang. Freedom: Training-free energy-guided conditional diffusion model. In *ICCV*, pages 23174–23184, 2023. 2

[40] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, pages 3836–3847, 2023. 2

[41] 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*, 2025. 2, 3, 6, 7

[42] Yabo Zhang, Yuxiang Wei, Dongsheng Jiang, Xiaopeng Zhang, Wangmeng Zuo, and Qi Tian. Controlvideo: Training-free controllable text-to-video generation. *arXiv preprint arXiv:2305.13077*, 2023. 3

[43] 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. 4

[44] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*, 2024. 2, 3

[45] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*, 2022. 3

[46] Shangchen Zhou, Chongyi Li, Kelvin CK Chan, and Chen Change Loy. Propainter: Improving propagation and transformer for video inpainting. In *ICCV*, pages 10477–10486, 2023. 4, 5

[47] 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. 8