This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Relightful Video Portrait Harmonization

Jun Myeong Choi[1] Jae Shin Yoon[2] Luchao Qi[1] Roni Sengupta[1] Joon-Young Lee[2]

[1]University of North Carolina at Chapel Hill [2]Adobe Research
https://chedgekorea.github.io/HarmoVid

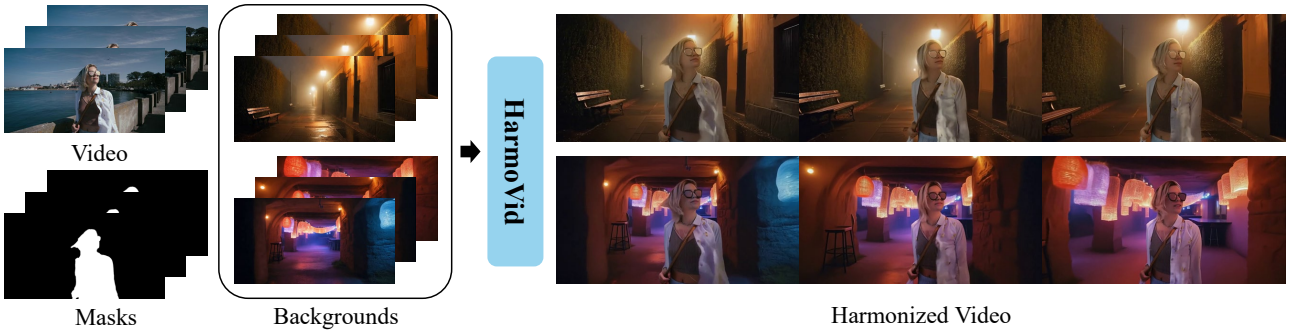


Figure 1. We present HarmoVid, a relightable video harmonization framework that aligns lighting between foreground and background. Given a foreground video with masks and a target background video, HarmoVid leverages a video diffusion model to directly synthesize a harmonized output. It produces realistic, temporally consistent results that accurately match foreground color, intensity, and shadows to the background while preserving both foreground and background content.

Abstract

We present a method for harmonizing the lighting of a foreground video to match a target background scene, adjusting shadows, color tone, and illumination intensity (relightful harmonization). Unlike images, acquiring labeled data for videos, where identical motions are recorded under different lighting conditions, is practically infeasible and non-scalable. While one way to create such paired data is to apply existing image-based harmonization models frame by frame to a video, the resulting outputs often suffer from significant temporal jitters. We overcome this problem by introducing a novel lighting deflickering model that can stabilize the global and local lighting flickering artifacts. Our video diffusion model learns from these upgraded deflickered data with a volume of real and synthetic videos to generate high-quality video harmonization results. We further propose an asymmetric alpha mask conditioning technique to learn the clean boundaries from real videos. Experiments demonstrate that our model achieves strong temporal coherence, naturalness, cleaner boundaries, and physically meaningful lighting behavior, while maintaining strong relighting expressiveness compared to prior image-based and video-based harmonization methods.

1. Introduction

Video harmonization, the task of adjusting a foreground object's appearance to seamlessly blend with a new background, has attracted increasing attention for realistic video composition in various applications such as filmmaking, video editing, and augmented reality. Despite significant progress in image harmonization, video harmonization remains a challenging task due to the additional complexity of the temporal dimension. Image harmonization techniques [4, 35, 47, 50] applied independently to each frame of a video often lead to temporal inconsistencies in lighting, causing noticeable flickering and making these approaches unsuitable for video harmonization.

The primary obstacle in directly extending image harmonization models to videos is the scarcity of paired training data; it is practically impossible to capture an identical performance under varied lighting and background conditions, especially when recording human subjects whose motions, expressions, and poses cannot be precisely reproduced across multiple takes. Although synthetic data can be generated at scale, it still fails to represent subtle lighting–appearance interactions and realistic temporal behaviors observed in real scenes, which limits its effectiveness as a substitute. This lack of high-quality, real-world data

severely constrains the generalization capabilities of existing image harmonization models, particularly for harmonizing unconstrained, in-the-wild videos.

Existing approaches [4, 9, 35, 47, 50, 51] have attempted to circumvent the data scarcity problem but often introduce their own limitations. Methods that extend frame-by-frame image harmonization techniques to video [7, 49] often fail to maintain temporal coherence, resulting in distracting flickering artifacts. Others that rely on synthetic datasets [14, 37, 55] struggle to generalize to the complex lighting dynamics and natural variations of real-world scenes. Although recent generative networks offer more flexibility [9, 23, 48, 51, 56], they often inadvertently alter the core characteristics of the input, *i.e.,* inadvertently alter the foreground or background contents, a problem known as identity shift. Furthermore, most methods [9, 50, 56] are highly sensitive to the quality of the input foreground mask, leading to unnatural boundary artifacts when using imperfect masks, which is common in practical workflows. These limitations highlight the key properties that a 'good' video harmonization model must satisfy: (a) preserve the identity and texture of the foreground and the background of the input video, (b) ensure temporal consistency, (c) exhibit robustness to imperfect masks, and (d) demonstrate high-fidelity expressiveness in lighting across diverse scenes. To date, no single framework has comprehensively addressed all these requirements.

To address these multifaceted challenges, we introduce a framework that achieves robust and temporally coherent video harmonization without requiring real paired data. Our approach is centered on a two-stage training strategy that combines data refinement with a robust harmonization model. We first composite real foreground videos onto various background videos, and generate pseudo-synthetic training pairs by applying an off-the-shelf image harmonization model to individual frames. Recognizing that this process introduces temporal flickering, we then train a lighting de-flickering network to refine these sequences, creating a high-quality and temporally consistent dataset. By jointly learning with real videos and refined synthetic videos, our video harmonization model, HarmoVid, could combine the strength of the videos from each domain (*e.g.,* physical plausibility from real videos, and lighting expressiveness from synthetic videos as shown in Fig. 2).

We further improve the harmonization results on the boundary by asymmetric conditioning of the foreground mask: we use binary mask conditioning when learning from synthetic videos, while pseudo-alpha maps from real videos. By learning from real videos with such pseudo-alpha maps, HarmoVid could produce smooth foreground-background transitions, robust to imperfect segmentation during compositing and harmonizing.

We conducted extensive experiments on challenging real-world video datasets to evaluate our video harmonizaiton model. The results demonstrate that our method significantly outperforms state-of-the-art approaches in both quantitative metrics and qualitative assessments. Notably, HarmoVid produces videos with superior temporal stability, better identity preservation, cleaner boundary blending, and naturalness.

Figure 2. Samples from our paired video harmonization dataset. Real videos serve as high-quality ground truth, offering strong supervision for temporal coherence, physical plausibility, natural appearance, and clean boundaries. Synthetic videos, while highly scalable and cover a wide range of expressive and dynamic lighting conditions, inherently lack full realism and temporal fidelity. Our video harmonization model, HarmoVid, leverages the complementary strengths of both real and synthetic data to achieve robust and high-quality harmonization.

2. Related Work

2.1. Image Relighting and Harmonization

Image relighting aims to modify the illumination of a scene while preserving its underlying geometry and appearance. Most recent approaches are based on deep learning, capturing high-level relationships between scene structure and lighting and enabling controllable adjustments.

Deep learning–based relighting can be categorized into several subgroups. 3D-aware relighting approaches [2, 16, 26, 27, 33, 34] utilize neural 3D representations to jointly model geometry and illumination, allowing complex lighting effects including viewpoint-dependent shadows and reflections. Inverse rendering-based methods [30, 31, 41, 46] explicitly decompose an image into reflectance, shading, and illumination components, providing physically interpretable results, though they often struggle to generalize to complex real-world scenes due to simplified assumptions about materials and lighting. Image-based relighting [5, 8, 15, 18, 25, 38, 54] with explicit lighting control leverages input images while providing additional cues such as light direction or intensity, enabling controllable and realistic relighting without requiring full 3D reconstruction.

In parallel, image harmonization focuses on adjusting the appearance of composite or inserted regions to match the lighting, tone, and color characteristics of the surrounding background [35, 47]. Recent diffusion-based methods further enhance realism and robustness using large-scale synthetic or augmented datasets and enable controllable adjustments via image- or text-guided cues [4, 50].

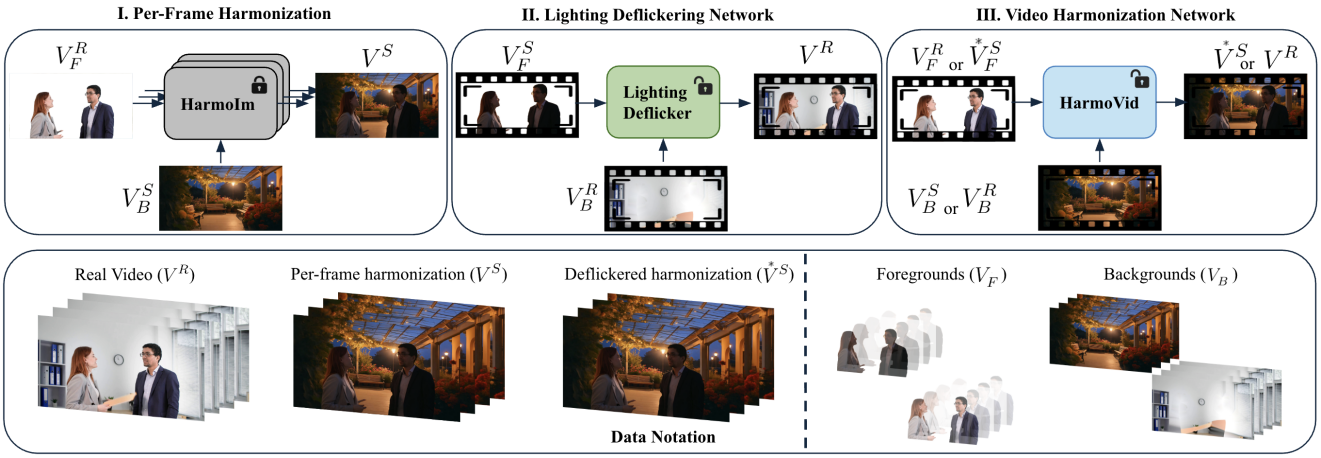


Figure 3. Our three-step pipeline for data synthesis and training for video harmonization. In the first state, we apply off-the-shelf image-based harmonization model to a video frame by frame to reconstruct the intermediate synthetic data. In the second stage, we upgrade the quality of the synthetic data by applying our lighting deflickering model to the synthetic videos. In the third step, HarmoVid jointly learns the real and synthetic data to generate temporally coherent, realistic, and expressive video harmonization results.

2.2. Video Relighting and Harmonization

Video relighting extends image relighting to sequences of frames, introducing the challenge of maintaining temporal consistency while ensuring spatially coherent illumination. Early works [7, 49] used optical flow or temporal regularization to stabilize results across frames, whereas more recent approaches employ neural rendering, 3D-aware representations [3, 10, 13, 20, 21, 43, 44, 53]—such as NeRF-based models and 3D Gaussian representations—or video diffusion models [11, 12, 22, 28] to jointly model lighting, geometry, and temporal dynamics. While those methods allow controlling the lighting properties of the videos, they do not support the harmonization that automatically matches the foreground and background lighting distribution.

Video harmonization shares the objective of maintaining consistent lighting, tone, and shading over time, particularly in composite or edited videos, while supporting the automatic lighting matching process. Recent methods further explore user-controllable or text-guided video harmonization. Some approaches [23, 48] enable text-to-video harmonization using video diffusion models. Light-A-Video [56] converts image-based harmonization models into video harmonization frameworks, whereas LumiSculpt[51] employ light maps and light trajectories to allow user-directed manipulation of lighting within a single video, with LightCtrl[1] additionally supporting user-controlled background lighting. However, all of these methods rely primarily on text prompts and cannot explicitly condition on the input background, effectively relighting the background along with the foreground. In contrast, RelightVid [9] leverages in-the-wild videos together with 3D-rendered data to achieve temporally consistent relighting under precise HDR illumination and explicitly conditions on the background. Nevertheless, when applied to real videos, it does not use paired data; instead, it employs foreground augmentation to adapt the model for real-world harmonization, which limits its ability to fully capture realistic lighting interactions and complex foreground–background relationships.

3. Method

Our goal is to harmonize a given foreground video with a new background video by adjusting shadows, color tones, and illumination intensity in a visually coherent manner (Fig. 1). The primary challenge lies in the scarcity of temporally consistent paired training data. To overcome this, we construct a large-scale pseudo-paired video dataset with diverse lighting conditions. Another challenge is the domain gap between real and synthetic videos. While real videos exhibit natural lighting and shadows, they lack sufficient diversity for learning expressive relighting effects. In contrast, synthetic videos can incorporate varied lighting conditions but often suffer from temporal artifacts.

To address these challenges, we employ a two-stage process for dataset generation: (1) per-frame image harmonization (Sec. 3.1) and (2) lighting deflickering (Sec. 3.2). After constructing the dataset, we train a video harmonization model, HarmoVid, that learns a bidirectional mapping between real and synthetic domains (Sec. 3.3). Fig. 3 provides an overview of the proposed framework.

3.1. Expressive Paired Video Dataset Generation

We first describe the process of generating an expressive paired video dataset. To create data containing diverse lighting effects, we prepare real foreground videos V_F^R containing target objects and synthetically generated background

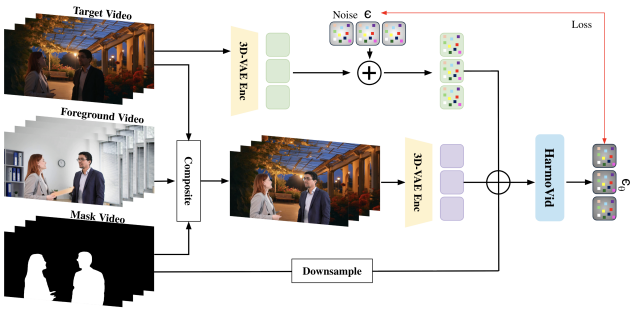


Figure 4. Our relightful video harmonization framework. HarmoVid is 3D transformer-based architecture that takes the input triplets of the latent space of the foregrounds with composited target backgrounds and noisy targets, and foreground masks; and predicts the noises that generates clean harmonized videos.

videos V_B^S. Foreground masks V^M are obtained using Grounded-SAM-2 [36] with the prompt "the main character (human)." The background videos can be generated by variants of many video generators such as [6, 17, 42] and are designed to include various lighting conditions. Similar to Text2Relight [4], we employ a range of lighting-related text prompts to control illumination characteristics.

Next, the foreground and background videos are composited, and a pretrained image harmonization model, HarmoIm [35], is applied to each frame individually to adapt lighting and color tone accordingly. The process can be formulated as:

$$V^S = \text{HarmoIm}(V_F^R, V_B^S, V^M), \tag{1}$$

where HarmoIm denotes the pretrained image harmonization model [35].

Although this stage enables the construction of a large-scale pseudo-paired dataset $\{(V^S, V^R)\}$, the independent per-frame processing inevitably introduces temporal inconsistencies such as flickering and abrupt illumination transitions. While one could apply existing video deflickering models [19] to mitigate such issues, these methods are generally limited to stabilizing global frame flicker and struggle to handle local lighting variations (e.g., rapid local shadow changes between consecutive frames). To resolve this limitation, we develop a lighting deflickering module that enhances temporal coherence across both global and local lighting components.

3.2. Lighting Deflickering Network

Our deflickering network is designed to stabilize temporal lighting coherence, addressing global color transitions and local shadow changes while maintaining spatial fidelity. To capture spatio-temporal dependencies effectively, we adopt a 3D latent diffusion transformer based on CogVideoX [45], which extends the DiT architecture to model spatial and temporal dimensions jointly.

Our lighting deflickering video diffusion model learns to predict the noises given an input flickering video and the corresponding foreground masks, as illustrated in Fig. 3:

$$\mathcal{L}_{\text{deflicker}} = \mathbb{E}_{t,\epsilon}\left[\left\|\epsilon - \epsilon_\theta(z^I, z_t^T, V^M, t)\right\|_2^2\right], \tag{2}$$

where z^I and z^T denote the latent representations of the synthetic composite V_{comp}^S and the real target video V^R, respectively. The representations are encoded with a 3D-VAE encoder $\mathcal{E}(\cdot)$:

$$z^I = \mathcal{E}(V_{\text{comp}}^S), \quad z^T = \mathcal{E}(V^R). \tag{3}$$

Gaussian noise $\epsilon \sim \mathcal{N}(0, 1)$ is added to z^T according to diffusion timestep t. The synthetic composite V_{comp}^S is produced by compositing the real background V_B^R and the synthetic foreground V_F^S using the foreground mask V^M:

$$V_{\text{comp}}^S = \mathcal{C}(V_F^S, V_B^R, V^M), \tag{4}$$

where $\mathcal{C}$ is the element-wise composite operation using the mask V^M.

After training this model, the frame-wise harmonized videos (generated in Sec. 3.1) V^S and their corresponding masks V^M are provided as inputs to the network. The model then produces temporally coherent videos $\hat{V}^S$, which exhibit reduced flickering and temporally coherent transitions under the background scenes.

3.3. Video Harmonization Network

We train our final video harmonization model, HarmoVid, using the paired video data constructed in the previous stage. The goal of HarmoVid is to adapt the lighting distribution of foregrounds to novel backgrounds, ensuring consistency in color and illumination while preserving temporal stability. As illustrated in Fig. 4, the foreground videos are first composited with the background videos using the triplet $\{V_F, V_B, V^M\}$, then the model produces harmonized results.

HarmoVid adopts the same 3D latent diffusion architecture as the deflickering network. It learns to predict noise in the latent space in order to reconstruct temporally coherent and spatially consistent harmonized videos:

$$\mathcal{L}_{\text{harm}} = \mathbb{E}_{t,\epsilon}\left[\left\|\epsilon - \epsilon_\theta(z^I, z^B, z_t^T, V^M, t)\right\|_2^2\right], \tag{5}$$

where z^F and z^B are the latent representations of the composited foreground and background, respectively, and z^T is the noisy latent space of the ground-truth target videos. The associated latent representations are obtained by:

$$z^I = \mathcal{E}(\mathcal{C}_{\text{in}}), \quad z^B = \mathcal{E}(V_B), \quad z^T = \mathcal{E}(V_{\text{target}}), \tag{6}$$

where $\mathcal{E}$ is the 3DVAE encoder [45].

Depending on how the training data are organized within the dataset, HarmoVid learns from two distinct training paths. The first path takes a real foreground and a synthetically generated background as input to produce a harmonized synthetic video (Real → Synthetic). The second path employs processed foreground—adjusted to align with a new background condition—together with its corresponding real background to reconstruct a harmonized real video (Synthetic → Real). This dual-path training strategy enables the model to effectively bridge the distribution gap between real and synthetic domains, thereby enhancing its robustness and generalization capability.

Real → Synthetic path. This training path takes real foregrounds V_F^R composited with synthetic backgrounds $\hat{V}_B^S$ and binary mask V^M, to generate the synthetic harmonization video $\hat{V}^S$. Learning from this real-to-synthetic path is crucial for preserving the expressive relighting and shadow representation capabilities originally captured by the image-based harmonization model.

Synthetic → Real path. In this path, the foreground from a synthetic video $\hat{V}_F^S$ is composited with a real background V_B^R, where the foreground region is first removed using inpainting [40]. A pseudo alpha mask $V^{\tilde{\alpha}}$ is used as a foreground mask during compositing instead of binary masks V^M. The pseudo alpha mask smoothly decays the foreground boundaries. Empirically, using the alpha mask in this synthetic-to-real path is critical for achieving natural and high-quality harmonization around object edges, as real videos provide perfect ground-truth blending supervision. In addition, this synthetic-to-real mapping enables the model to learn temporal coherence and physically consistent adjustments of shadows and illumination, inherited from real videos.

4. Experiments

We conduct comprehensive experiments to evaluate the effectiveness of our **HarmoVid** in terms of both spatial harmonization quality and temporal stability. Our experiments include quantitative comparisons with implementation details, existing baselines, temporal smoothing analysis, ablation studies, and user studies.

4.1. Experimental Details

Implementation Details. We curated videos from stock video sources and applied Grounded-SAM-2 [36] to both identify portrait content and extract masks, resulting in a total of 10,000 portrait videos. For our experiments, we sampled 100 frames per video, and reserved a subset of 200 videos for testing.

We train HarmoVid with a standard L2 diffusion loss on 8 NVIDIA A100 GPUs for 8 hours, totaling 1,200 iterations. During inference, we employ temporal MultiDiffusion [52] for videos longer than 85 frames, enabling high-quality video harmonization on extended sequences.

Baselines. We compare our approach against recent state-of-the-art harmonization methods, including IC-Light [50], Relightful Harmonization [35], RelightVid [9], and Light-A-Video [56]. For a fair comparison, all methods are evaluated on the same composited foreground–background pairs. Among these, IC-Light and Relightful Harmonization perform frame-wise harmonization, whereas RelightVid and Light-A-Video incorporate temporal consistency modeling. As the original Light-A-Video does not support background conditioning, we incorporated its training-free module into the IC-Light background conditioning version. Additionally, to identify the effectiveness of our Video Deflickering Network, we compare our lighting deflickering model with a general deflickering baseline, BVD [19].

Evaluation datasets. We evaluate our model both quantitatively and qualitatively, focusing on spatial harmonization quality and temporal consistency.

For qualitative evaluation, we utilize a real portrait dataset, where each foreground subject is composited with various arbitrary background scenes to demonstrate the visual harmonization capability of our method.

For quantitative evaluation, we construct two synthetic test sets derived from the same real portrait dataset by applying a color look-up-table (LUT) [24] to the foreground videos. In the first setting, a *single LUT is consistently applied to all frames within each video*, producing uniformly relit sequences. This setup provides paired videos that enable direct numerical comparison with the original unmodified sequences, allowing us to evaluate the video harmonization quality (Section 4.2). In the second setting, *different LUTs are randomly applied on a per-frame basis* to introduce temporal color and illumination variations, simulating flickering artifacts. This dataset is used to evaluate the effectiveness of our Video Deflickering Network in restoring temporal consistency while maintaining appearance fidelity (Section 4.3). Together, these two complementary datasets allow a comprehensive evaluation of both the harmonization and deflickering performance of our proposed method.

Evaluation metrics. *Temporal consistency* is measured using two complementary metrics used in Light-A-Video [56]. First, we calculate the average CLIP [32] similarity between consecutive frames to capture semantic stability over time (CLIP Score). Second, we estimate optical flow with RAFT [39] and compute the flow discrepancy between the harmonized and reference videos, reflecting how well motion dynamics are preserved (Motion Preservation).

For *frame-wise harmonization quality*, we use PSNR, SSIM, LPIPS, and RMSE by comparing the generated frames with corresponding reference videos. Higher PSNR and SSIM, along with lower LPIPS and RMSE, indicate

Table 1. We compare HarmoVid with state-of-the-art image and video harmonization techniques to quantitatively evaluate harmonization quality (matching foreground and background lighting) and temporal consistency, along with qualitative evaluation with a user study.

	Harmonization Quality				Temporal Consistency		User Study (%)		
	PSNR ↑	SSIM ↑	LPIPS ↓	RMSE ↓	CLIP Score ↑	Motion Preservation ↓	Temporal ↑	ID ↑	Harmonization ↑
IC-Light [50]	14.77	0.8889	0.0828	0.1881	0.9895	1.2928	56 (%)	57 (%)	27 (%)
Relightful Harmonization [35]	15.89	0.9301	0.0581	0.1643	0.9907	1.0021	36 (%)	50 (%)	36 (%)
RelightVid [9]	15.70	0.9214	0.0707	0.1711	0.9946	0.7096	35 (%)	51 (%)	27 (%)
Light-A-Video [56]	15.64	0.8900	0.0791	0.1716	0.9955	0.5775	56 (%)	58 (%)	51 (%)
Ours (HarmoVid)	**17.91**	**0.9306**	**0.0554**	**0.1325**	**0.9963**	**0.5264**	**82 (%)**	**78 (%)**	**72 (%)**

better perceptual fidelity and realism, while higher CLIP Score and lower Motion Preservation indicate better spatial and temporal consistency.

4.2. Video Harmonization Comparison

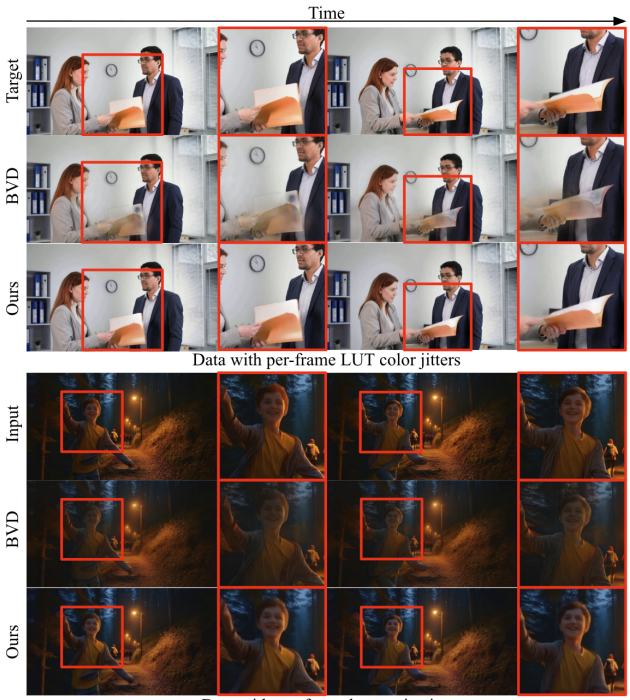
We comprehensively evaluate the proposed **HarmoVid** framework against state-of-the-art video harmonization methods. As shown in Table 1, HarmoVid consistently achieves superior spatial and temporal harmonization performance compared to existing approaches. Qualitative comparisons in Figure 5 further demonstrate that our method seamlessly blends the foreground and background while maintaining realistic illumination and temporal stability. In particular, in the left examples, IC-Light tends to alter global colors excessively, Relightful Harmonization exhibits reflection flickering, RelightVid struggles with ID preservation, and Light-A-Video introduces artifacts in the background. In contrast, our results remain temporally consistent and visually realistic. The right example also highlights that other methods often produce noticeable artifacts around the portrait region, whereas our approach effectively suppresses them, thanks to the use of a pseudo alpha mask that enables more precise boundary blending between the foreground and background.

User Study. We further validate the perceptual quality through a user study involving 33 participants. Each participant views 10 video clips randomly selected from a total of 30 samples. For each clip, participants evaluate results from five competing methods by selecting *all* methods they consider *appropriate* in terms of: (1) *Temporal Consistency* — smoothness and stability of motion, (2) *ID Preservation* — consistency of subject identity and appearance, and (3) *Overall Harmonization* — naturalness and lighting realism. The selections are then aggregated across participants to derive subjective preference scores, complementing the quantitative evaluation metrics.

The aggregated responses in Table 1 show that HarmoVid is the most frequently preferred, confirming its advantage in producing visually coherent and temporally stable results.

4.3. Deflickering Comparison

We evaluate the effectiveness of the proposed Video Deflickering Network in reducing temporal artifacts in flickering videos. Specifically, we compare our deflickering results with BVD [19] on the video data applied with per-frame LUT color jitters described in Section 4.1, which mainly exhibit typical frame-to-frame flickering. We further apply our deflickering method to the video data applied with per-frame harmonization outputs that exhibit more complex temporal inconsistencies of local lighting, such as cast-shadow and highlight flickering. For quantitative evaluation, we employ the same motion preservation metric; however, due to the absence of ground truth videos, the met-



	Per-frame LUT color jitters		Per-frame harmonization	
	CLIP Score ↑	Motion Pres. ↓	CLIP Score ↑	Motion Pres. ↓
BVD [19]	0.9950	0.5114	0.9920	1.3439
HarmoVid	**0.9967**	**0.3630**	**0.9936**	**0.5395**

Table 2. We demonstrate the effectiveness of our deflickering network. In the upper synthetic example, BVD suffers from noticeable loss of spatial detail, whereas in the lower per-frame harmonization example, it fails to preserve lighting effects. Our lighting deflickering method effectively stabilize the global and local lighting flickering artifacts with minimum loss of details and lightings.

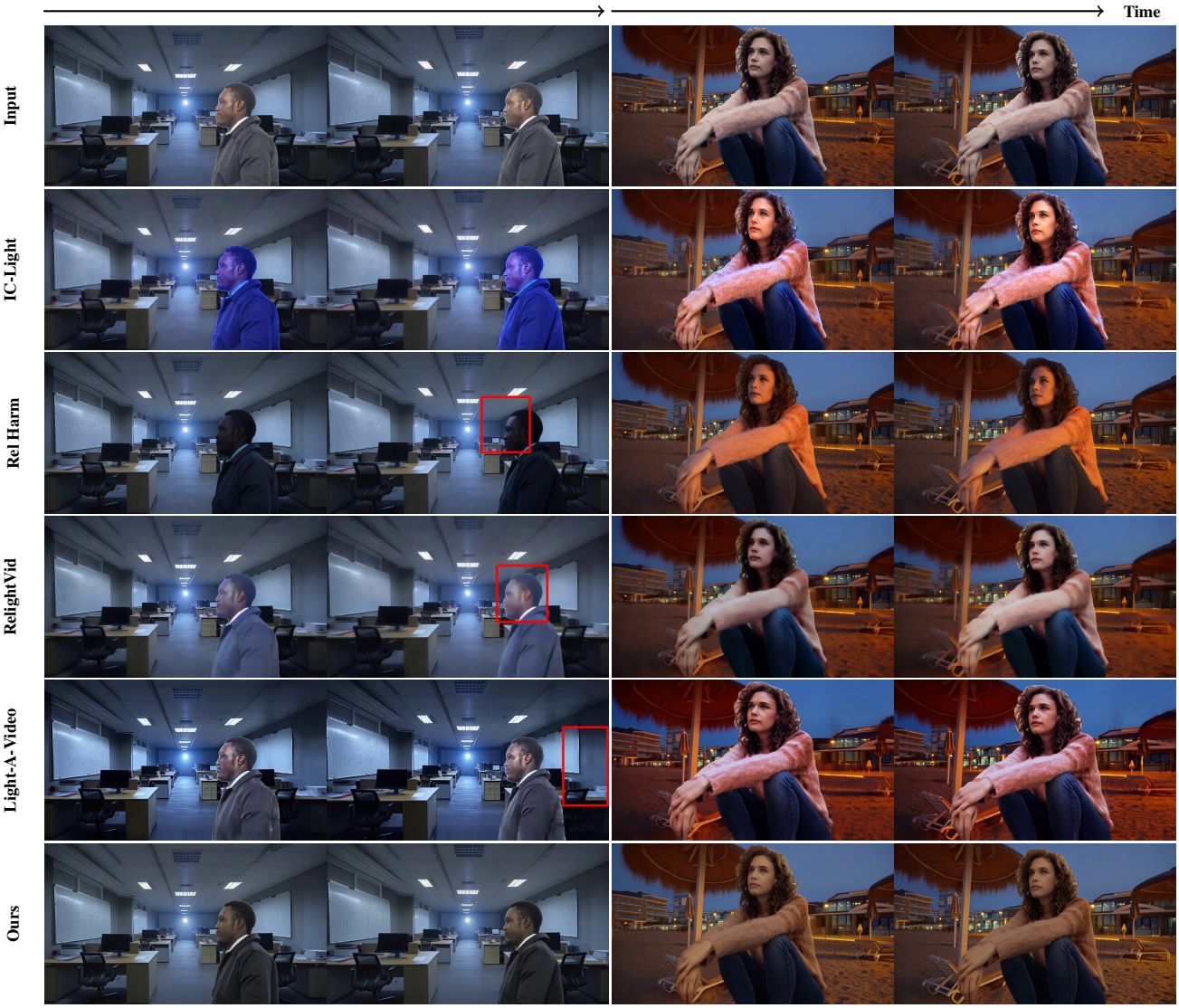


Figure 5. Qualitative comparison of our HarmoVid to other image- or video-based video harmonization methods.

ric is computed using the foreground regions of the original frames. Evaluations based on CLIP Score and Motion Preservation indicate that our method improves temporal consistency in both settings (Table 2). Moreover, qualitative comparisons on consecutive frames demonstrate that our method effectively suppresses flickering artifacts while preserving spatial details, including lighting effects, thereby improving temporal coherence without compromising visual fidelity.

4.4. More Analysis

Ablation study on two-stage training. To investigate the contribution of each stage, we run ablations focusing on Stage 2 (lighting deflickering; Sec. 3.2) and Stage 3 (dual-path training; Sec. 3.3), with results summarized in Tab. 3. Skipping Stage 2 degrades both temporal stability and harmonization quality: when one side of a training pair flickers, a DiT that jointly models space–time struggles to learn stable temporal representations and high-quality harmonization. In particular, Tab. 3 2nd-row, shows that harmonization is almost entirely absent. Using only the Stage 2 deflickering network for harmonization (bypassing Stage 3) suppresses short-term flicker but yields long-horizon inconsistencies (Tab. 3, 3rd-row), such as gradual illumination and tone drift; absent joint learning on real videos, outputs also appear less natural. Moreover, during inference, Stage 2 cannot be used in isolation, as it requires per-frame processing via Stage 1 (image harmonization) beforehand. Overall, the deflickering process (Stage 2) and the dual-path training (Stage 3) are complementary: together they refine fine-grained spatial details and preserve both short- and long-term temporal coherence.

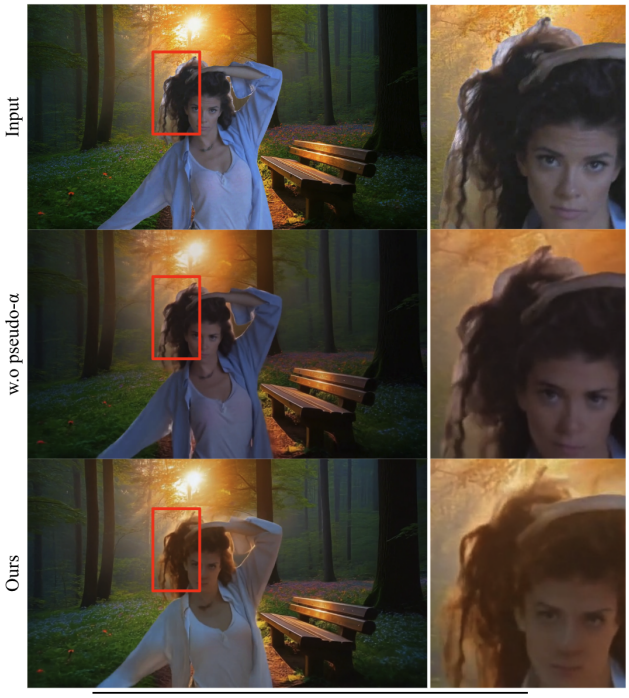
Stage 2	Stage 3	SSIM ↑	LPIPS ↓	CLIP Score ↑	Motion Preservation ↓
–	✓	0.9187	0.0613	0.9911	0.9376
✓	–	0.9217	0.0594	0.9937	0.5490
✓	✓	**0.9306**	**0.0554**	**0.9963**	**0.5264**

Table 3. **Ablation study**: We demonstrate the necessity of both Stage 2 and Stage 3. The results in the table are obtained using the same dataset used in 4.2, while the figures show examples applied to real-world data.

Ablation study on pseudo-alpha mask. Next, we evaluate the effect of incorporating the pseudo alpha mask during training by comparing the outputs generated with and without it. Integrating the pseudo alpha mask enhances foreground-background blending, particularly along object boundaries, leading to more natural and visually coherent harmonization.

To quantitatively assess the improvements without relying on ground-truth references, we compute reference-free image quality metrics—Laplacian Variance [29] and Tenengrad—focused specifically on the foreground-background boundaries. For each frame, we extract a narrow boundary region around the pseudo alpha mask using morphological erosion and collision operations, and calculate the metrics only within these boundary pixels. Laplacian Variance evaluates edge intensity, capturing how well fine details are preserved along the boundary, while Tenengrad assesses gradient magnitudes, reflecting the clarity of structural details in complex regions. Using these boundary-focused metrics, we can systematically verify that integrating the pseudo alpha mask preserves subtle edges and maintains high-quality foreground-background blending. As illustrated in Table 4, the benefits are most pronounced in challenging regions such as hair, where intricate details and subtle boundaries are retained more faithfully.

Generalization. Figure 6 shows that our model, trained exclusively on portrait foreground videos, works effectively on non-portrait foreground objects.



pseudo-α	LaplacianVar ↑	Tenengrad ↑
–	66.106	2611.501
✓	**79.098**	**3324.573**

Table 4. **Ablation study**: Effect of incorporating the pseudo alpha mask. We evaluate the outputs using reference-free, boundary-focused image quality metrics.

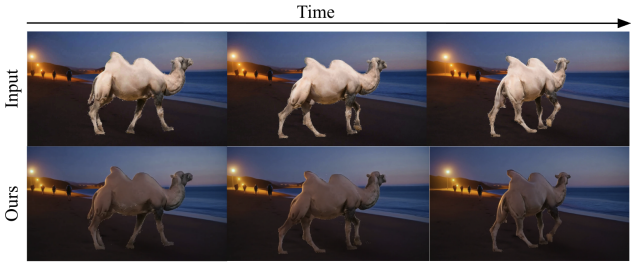


Figure 6. Result on non-portrait foreground object.

5. Conclusion

We have presented **HarmoVid**, a novel framework for high-quality video harmonization that effectively addresses the challenges of temporal consistency, identity preservation, and realistic lighting adaptation. By introducing a two-stage training strategy—comprising per-frame image harmonization followed by a dedicated Video Deflickering Network—our method is able to generate temporally coherent and visually pleasing harmonized videos even in the absence of paired real-world data. By combining real and synthetic video data, our framework achieves expressive relighting effects while maintaining physical plausibility and temporal smoothness, bridging the gap between controllable synthetic scenarios and complex real-world videos.

Acknowledgement. This work was conducted during an internship at Adobe Research. Some of the video assets used in this work are sourced from Adobe Stock and are used in accordance with their licensing terms. These assets are used solely for research and visualization purposes.

This work is partly supported by a National Institute of Health (NIH) project #R21EB037440.

We thank to Mengwei Ren for sharing the image harmonization model.

References

[1] Anonymous. Lightctrl: Training-free controllable video relighting, 2025. under review in ICLR2026. 3

[2] Ziqi Cai, Kaiwen Jiang, Shu-Yu Chen, Yu-Kun Lai, Hongbo Fu, Boxin Shi, and Lin Gao. Real-time 3d-aware portrait video relighting, 2024. 2

[3] Ziqi Cai, Kaiwen Jiang, Shu-Yu Chen, Yu-Kun Lai, Hongbo Fu, Boxin Shi, and Lin Gao. Real-time 3d-aware portrait video relighting, 2024. 3

[4] Junuk Cha, Mengwei Ren, Krishna Kumar Singh, He Zhang, Yannick Hold-Geoffroy, Seunghyun Yoon, HyunJoon Jung, Jae Shin Yoon, and Seungryul Baek. Text2relight: Creative portrait relighting with text guidance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1980–1988, 2025. 1, 2, 4

[5] 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, 2025. 2

[6] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models, 2024. 4

[7] Jun Myeong Choi, Max Christman, and Roni Sengupta. Personalized video relighting with an at-home light stage. In *European Conference on Computer Vision*, pages 394–410. Springer, 2024. 2, 3

[8] Jun Myeong Choi, Annie Wang, Pieter Peers, Anand Bhattad, and Roni Sengupta. Scribblelight: Single image indoor relighting with scribbles. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5720–5731, 2025. 2

[9] Ye Fang, Zeyi Sun, Shangzhan Zhang, Tong Wu, Yinghao Xu, Pan Zhang, Jiaqi Wang, Gordon Wetzstein, and Dahua Lin. Relightvid: Temporal-consistent diffusion model for video relighting. *arXiv preprint arXiv:2501.16330*, 2025. 2, 3, 5, 6

[10] Xuan Gao, Haiyao Xiao, Chenglai Zhong, Shimin Hu, Yudong Guo, and Juyong Zhang. Portrait video editing empowered by multimodal generative priors, 2024. 3

[11] Mingtao Guo, Guanyu Xing, and Yanli Liu. High-fidelity relightable monocular portrait animation with lighting-controllable video diffusion model, 2025. 3

[12] 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, 2025. 3

[13] Yu Hong, Yize Wu, Zhehao Shen, Chengcheng Guo, Yuheng Jiang, Yingliang Zhang, Jingyi Yu, and Lan Xu. Beam: Bridging physically-based rendering and gaussian modeling for relightable volumetric video, 2025. 3

[14] Andrew Hou, Ze Zhang, Michel Sarkis, Ning Bi, Yiying Tong, and Xiaoming Liu. Towards high fidelity face relighting with realistic shadows, 2021. 2

[15] Andrew Hou, Zhixin Shu, Xuaner Zhang, He Zhang, Yannick Hold-Geoffroy, Jae Shin Yoon, and Xiaoming Liu. Compose: Comprehensive portrait shadow editing, 2024. 2

[16] 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. *Advances in Neural Information Processing Systems*, 37:141129–141152, 2024. 2

[17] 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, 2023. 4

[18] 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, 2024. 2

[19] Chenyang Lei, Xuanchi Ren, Zhaoxiang Zhang, and Qifeng Chen. Blind video deflickering by neural filtering with a flawed atlas. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10439–10448, 2023. 4, 5, 6

[20] Ruofan Liang, Zan Gojcic, Huan Ling, Jacob Munkberg, Jon Hasselgren, Zhi-Hao Lin, Jun Gao, Alexander Keller, Nandita Vijaykumar, Sanja Fidler, and Zian Wang. Diffusion-renderer: Neural inverse and forward rendering with video diffusion models, 2025. 3

[21] Wenbin Lin, Chengwei Zheng, Jun-Hai Yong, and Feng Xu. Relightable and animatable neural avatars from videos, 2023. 3

[22] Yuanze Lin, Yi-Wen Chen, Yi-Hsuan Tsai, Ronald Clark, and Ming-Hsuan Yang. Illumicraft: Unified geometry and illumination diffusion for controllable video generation, 2025. 3

[23] Yang Liu, Chuanchen Luo, Zimo Tang, Yingyan Li, Yuran Yang, Yuanyong Ning, Lue Fan, Zhaoxiang Zhang, and Junran Peng. Tc-light: Temporally coherent generative rendering for realistic world transfer, 2025. 2, 3

[24] Xinyuan Lu, Shengyuan Huang, Li Niu, Wenyan Cong, and Liqing Zhang. Deep video harmonization with color mapping consistency. *arXiv preprint arXiv:2205.00687*, 2022. 5

[25] Yiqun Mei, He Zhang, Xuaner Zhang, Jianming Zhang, Zhixin Shu, Yilin Wang, Zijun Wei, Shi Yan, HyunJoon Jung, and Vishal M. Patel. Lightpainter: Interactive portrait relighting with freehand scribble, 2023. 2

[26] Yiqun Mei, Yu Zeng, He Zhang, Zhixin Shu, Xuaner Zhang, Sai Bi, Jianming Zhang, HyunJoon Jung, and Vishal M. Patel. Holo-relighting: Controllable volumetric portrait relighting from a single image, 2024. 2

[27] Yiqun Mei, Yu Zeng, He Zhang, Zhixin Shu, Xuaner Zhang, Sai Bi, Jianming Zhang, HyunJoon Jung, and Vishal M. Pa-

tel. Holo-relighting: Controllable volumetric portrait relighting from a single image, 2024. 2

[28] 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, 2025. 3

[29] Farida Memon, Mukhtiar Ali Unar, and Sheeraz Memon. Image quality assessment for performance evaluation of focus measure operators, 2016. 8

[30] Rohit Pandey, Sergio Orts Escolano, Chloe Legendre, Christian Häne, Sofien Bouaziz, Christoph Rhemann, Paul Debevec, and Sean Fanello. Total relighting: learning to relight portraits for background replacement. *ACM Trans. Graph.*, 40(4), 2021. 2

[31] Puntawat Ponglertnapakorn, Nontawat Tritrong, and Supasorn Suwajanakorn. Difareli++: Diffusion face relighting with consistent cast shadows. *arXiv e-prints*, pages arXiv–2304, 2023. 2

[32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 5

[33] Pramod Rao, Gereon Fox, Abhimitra Meka, Mallikarjun B R, Fangneng Zhan, Tim Weyrich, Bernd Bickel, Hanspeter Pfister, Wojciech Matusik, Mohamed Elgharib, and Christian Theobalt. Lite2relight: 3d-aware single image portrait relighting. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers 24*, page 1–12. ACM, 2024. 2

[34] Pramod Rao, Gereon Fox, Abhimitra Meka, Mallikarjun B R, Fangneng Zhan, Tim Weyrich, Bernd Bickel, Hanspeter Pfister, Wojciech Matusik, Mohamed Elgharib, and Christian Theobalt. Lite2relight: 3d-aware single image portrait relighting. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers 24*, page 1–12. ACM, 2024. 2

[35] 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*, pages 6452–6462, 2024. 1, 2, 4, 5, 6

[36] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024. 4, 5

[37] Guoxian Song, Tat-Jen Cham, Jianfei Cai, and Jianmin Zheng. Half-body portrait relighting with overcomplete lighting representation. *Computer Graphics Forum*, 40(6): 371–381, 2021. 2

[38] Tiancheng Sun, Jonathan T. Barron, Yun-Ta Tsai, Zexiang Xu, Xueming Yu, Graham Fyffe, Christoph Rhemann, Jay Busch, Paul Debevec, and Ravi Ramamoorthi. Single image portrait relighting. *ACM Transactions on Graphics*, 38(4): 1–12, 2019. 2

[39] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pages 402–419. Springer, 2020. 5

[40] Alexandru Telea. An image inpainting technique based on the fast marching method. *J. Graphics, GPU, Game Tools*, 9(1):23–34, 2004. 5

[41] Zhibo Wang, Xin Yu, Ming Lu, Quan Wang, Chen Qian, and Feng Xu. Single image portrait relighting via explicit multiple reflectance channel modeling. *ACM Trans. Graph.*, 39 (6), 2020. 2

[42] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7623–7633, 2023. 4

[43] Zhen Xu, Sida Peng, Chen Geng, Linzhan Mou, Zihan Yan, Jiaming Sun, Hujun Bao, and Xiaowei Zhou. Relightable and animatable neural avatar from sparse-view video, 2023. 3

[44] Haijie Yang, Zhenyu Zhang, Hao Tang, Jianjun Qian, and Jian Yang. Follow your motion: A generic temporal consistency portrait editing framework with trajectory guidance, 2025. 3

[45] 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. 4

[46] Yu-Ying Yeh, Koki Nagano, Sameh Khamis, Jan Kautz, Ming-Yu Liu, and Ting-Chun Wang. Learning to relight portrait images via a virtual light stage and synthetic-to-real adaptation. *ACM Transactions on Graphics*, 41(6):1–21, 2022. 2

[47] Jae Shin Yoon, Zhixin Shu, Mengwei Ren, Xuaner Zhang, Yannick Hold-Geoffroy, Krishna Kumar Singh, and He Zhang. Generative portrait shadow removal, 2024. 1, 2

[48] 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. 2, 3

[49] Longwen Zhang, Qixuan Zhang, Minye Wu, Jingyi Yu, and Lan Xu. Neural video portrait relighting in real-time via consistency modeling, 2021. 2, 3

[50] 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 *The Thirteenth International Conference on Learning Representations*, 2025. 1, 2, 5, 6

[51] Yuxin Zhang, Dandan Zheng, Biao Gong, Shiwen Wang, Jingdong Chen, Ming Yang, Weiming Dong, and Changsheng Xu. Lumisculpt: Enabling consistent portrait lighting in video generation, 2025. 2, 3

[52] Zhixing Zhang, Bichen Wu, Xiaoyan Wang, Yaqiao Luo, Luxin Zhang, Yinan Zhao, Peter Vajda, Dimitris Metaxas,

and Licheng Yu. Avid: Any-length video inpainting with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7162–7172, 2024. 5

[53] Sixiao Zheng, Zimian Peng, Yanpeng Zhou, Yi Zhu, Hang Xu, Xiangru Huang, and Yanwei Fu. Vidcraft3: Camera, object, and lighting control for image-to-video generation, 2025. 3

[54] Hao Zhou, Sunil Hadap, Kalyan Sunkavalli, and David W. Jacobs. Deep single-image portrait relighting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 2

[55] Hao Zhou, Sunil Hadap, Kalyan Sunkavalli, and David W Jacobs. Deep single-image portrait relighting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7194–7202, 2019. 2

[56] Yujie Zhou, Jiazi Bu, Pengyang Ling, Pan Zhang, Tong Wu, Qidong Huang, Jinsong Li, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, et al. Light-a-video: Training-free video relighting via progressive light fusion. *arXiv preprint arXiv:2502.08590*, 2025. 2, 3, 5, 6