Relightful Video Portrait Harmonization

Supplementary Material

Along with this supplemental PDF, we provide additional visual materials (e.g., videos) in an website, accessible via https://chedgekorea.github.io/HarmoVid. We highly recommend readers to refer to the accompanying videos for a comprehensive examination of the visual outcomes.

A. Overview of Website

We present the video results corresponding to the figures shown in the main paper. Here, we explain the order and types of results displayed in our website.

- We first present video results corresponding to the qualitative comparison in Figure 5, showcasing how our **HarmoVid** performs relative to other image- or video-based harmonization methods.
- Next, we show the video results for the deflickering comparison reported in Table 2, which evaluates performance on data processed with per-frame harmonization outputs.
- We then provide videos demonstrating the three-stage pipeline in Table 3, highlighting the necessity of both Stage 2 and Stage 3.
- We also include results for the pseudo-α mask in Table 4, showing that using the pseudo-alpha mask produces smoother boundaries and more realistic harmonization of fine details such as hair.
- Finally, we present harmonization results on the DAVIS dataset [30], illustrating the effectiveness of our method on non-portrait foreground objects.

B. Necessity of Two-Stage Approach

To train a video harmonization model, paired videos are crucial. Since it is impossible to collect the same motion under different lighting, synthetic data is indispensable, which we generate using Stage 1.

Single Stage	SSIM ↑	LPIPS ↓	CLIP Score ↑	Motion Preservation ↓
Real → Synthetic	0.9027	0.0783	0.9903	0.9516
Real ← Synthetic	0.6439	0.1258	0.9961	0.5582
Real ↔ Synthetic	0.9187	0.0613	0.9911	0.9376
Two Stage	**0.9306**	**0.0554**	**0.9963**	**0.5264**

Table 5. Single-Stage Comparison using Table 1 dataset.

Training a single-stage on both real video data and flickering synthetic data presents problems. When training with real input and synthetic target, the model successfully applies some harmonization, but temporal consistency is disrupted, resulting in flickering artifacts in the output. Conversely, when training with synthetic input and real target, the model tends to learn that the input should contain flickering. As a result, if an input without flickering is provided during inference, the output closely resembles the input.

Using a dual-path approach (Table 3, Stage 3) allows both types of supervision to be applied simultaneously. However, this does not reduce the domain gap to produce temporally consistent and harmonization; the two paths operate independently. As a result, when using real video input during inference, harmonization occurs to some degree, but the output exhibits flickering (similar to Table 5 1st row). Therefore, we first apply a deflickering network to reduce flickering in the synthetic data, and use it in Stage 3 to achieve both harmonization and temporal consistency.

C. Effectiveness of Dual-path Training

We demonstrate that, during Stage 3 training using the deflickered videos obtained from Stage 2, the dual-path design achieves better performance than the single-path design, as discussed in the main paper.

Stage 3	SSIM ↑	LPIPS ↓	CLIP Score ↑	Motion Preservation ↓
Real → Synthetic	<u>0.9287</u>	<u>0.0577</u>	0.9951	0.5510
Real ← Synthetic	0.9202	0.0607	<u>0.9961</u>	<u>0.5271</u>
Real ↔ Synthetic	**0.9306**	**0.0554**	**0.9963**	**0.5264**

Table 6. The real→synthetic path modifies lighting expressively, while the real←synthetic path ensures natural harmonization and temporal coherence; together, the model captures both benefits.

D. Computational Cost Analysis

	Inference Time (s) ↓	GPU (GB) ↓
Light-A-Video	1822	42.483
RelightVid	199	27.234
Ours	**168**	**17.385**

Table 7. Our method shows better practical efficiency compared to other baselines.

E. User Study Structure Details

Introduction. This study was designed to gather perceptual feedback on three key aspects of video harmonization: overall visual quality, temporal consistency, and identity preservation. By focusing on these aspects, we aim to evaluate not only the realism of individual frames but also the stability and coherence of the foreground object across the video sequence.

We then presented explanations for each video example as shown in Figure 7, where five different results are displayed in random order.

Introduction

In this study, you will be asked to evaluate the visual quality of harmonized videos. Video harmonization aims to make a composited foreground object look naturally integrated with the background scene — for example, by adjusting lighting, color, and shading to match the background lighting.
The harmonization process modifies only the white mask region to make it visually consistent with the background scene.
Each video is organized as follows:
the first row contains the input (composited video) and mask (white = foreground, black = background), the second row contains 2 different harmonized results (1–2), and the third row contains 3 different harmonized results (3–5).

Figure 7. Illustration of an example video from the user study.

Section 1: Overall Harmonization Quality. Participants watched 10 harmonized video clips and rated how natural and realistic the harmonized foreground appeared as a whole. They were instructed:

Section 1

Please ignore small holes or missing areas inside the mask (due to the objects), and focus solely on how natural and consistent the harmonization looks.
Please base your ratings only on the harmonization quality, not on the content of the scene or the actions performed.

The evaluation question was:

Questions 1-10

How natural and realistic does the harmonized foreground look in the scene? (in terms of harmonization and relighting)
Choose all the realistic videos.

This section specifically measures whether the harmonized foreground appears convincingly integrated into the background, reflecting proper lighting, color matching, and shading adjustments.

Section 2: Temporal Consistency and Identity Preservation. The same 10 video clips were evaluated as follows. First they were instructed:

Temporal Consistency & Identity Preservation

You will evaluate same 10 harmonized videos based on two aspects:
Temporal Consistency — smoothness and stability across frames (i.e., no flickering or unnatural changes).
Identity Preservation — how consistent the subject's identity and appearance remain throughout the sequence. (how similar the harmonized foreground is to the input foreground.)

The evaluation question was:

Questions 11-20

How temporally consistent is the harmonized foreground across the video? (Focus on flickering, and smooth transitions between frames)
Choose all the temporal consistent videos.

This evaluates whether the harmonization is stable over time, which is crucial for video applications, as inconsistent lighting or color shifts can break realism.

Questions 21-30

How consistent does the subject's (foreground) identity and appearance remain throughout the sequence?
Choose all the videos where the identity is well preserved.

This ensures that harmonization does not alter key attributes of the foreground subject, such as facial features, hairstyle, or object shape, which is particularly important for portraits or recognizable objects.

Per-Video Results. In addition to the table results in the main paper (Tab. 1), we also present per-video results as bar charts in Figs. 8 and 9. The user study was conducted using three different templates, with each row corresponding to one template. The first, second, and third templates were completed by 8, 9, and 16 participants, respectively.

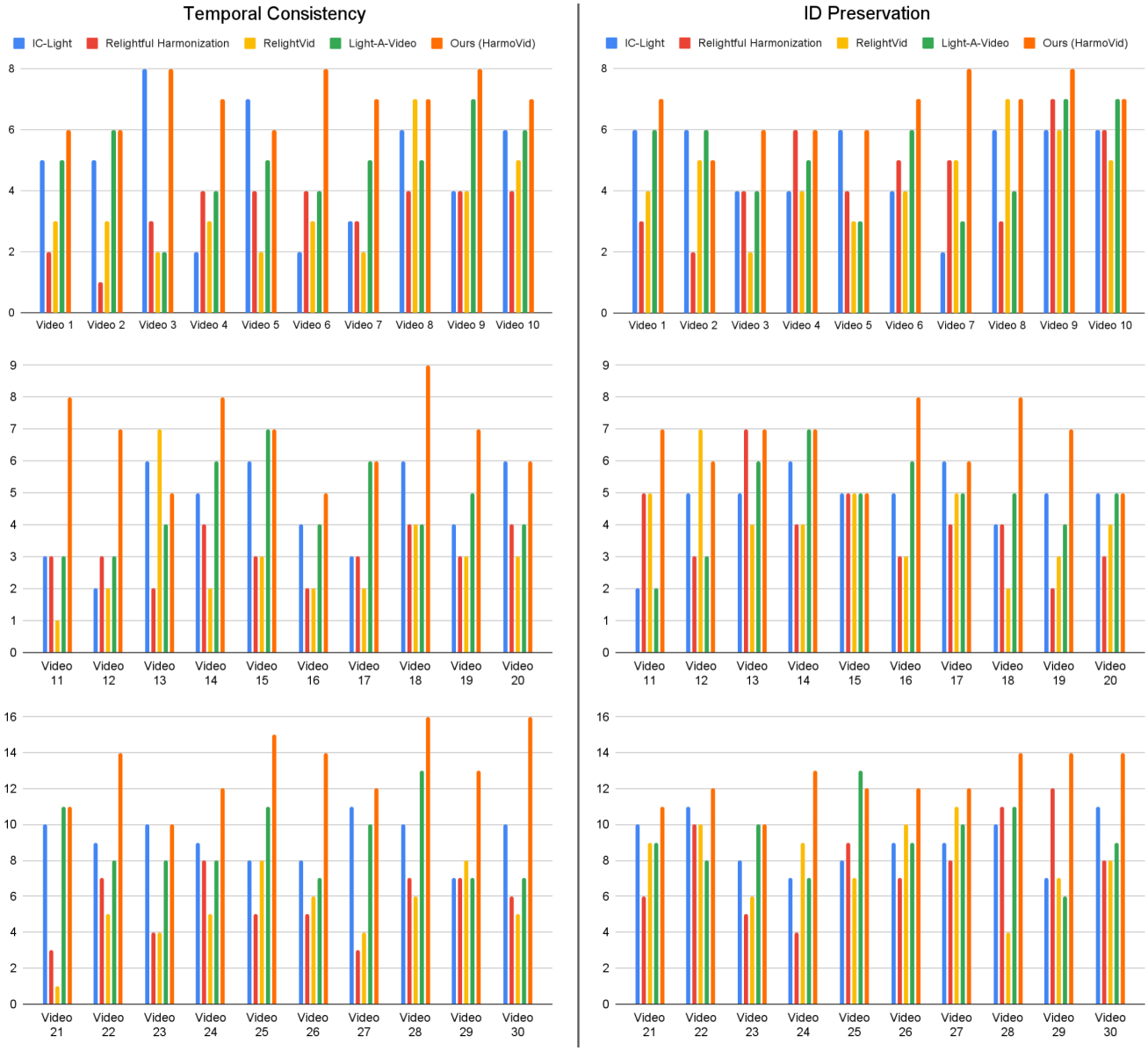

Figure 8. The user study results for temporal consistency and identity preservation.

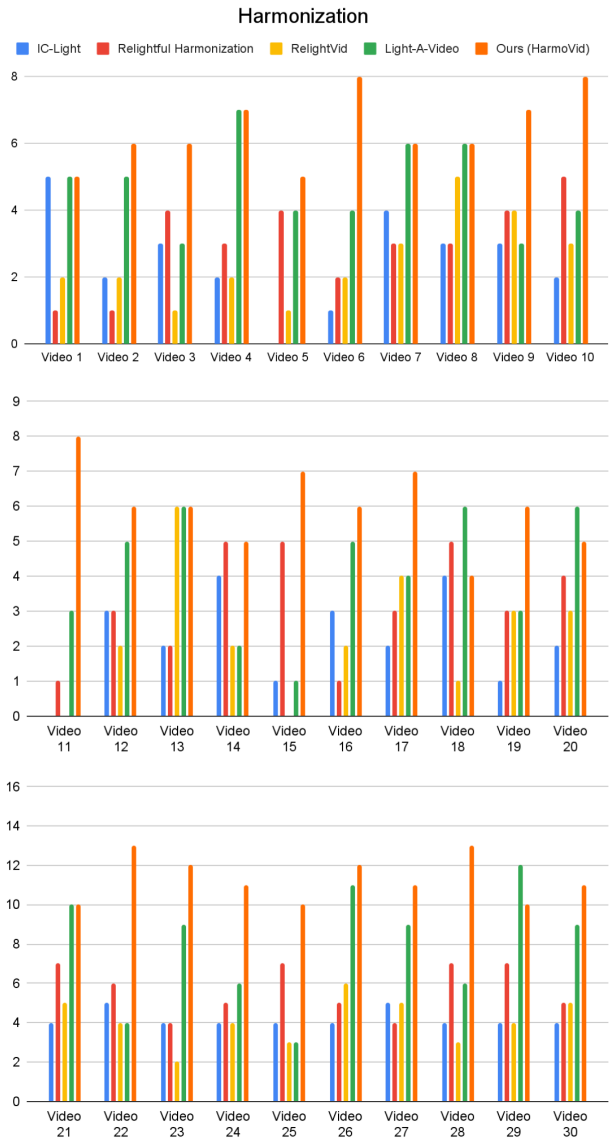

Figure 9. The user study results for overall harmonization quality.