Light-A-Video: Training-free Video Relighting via Progressive Light Fusion

Supplementary Material

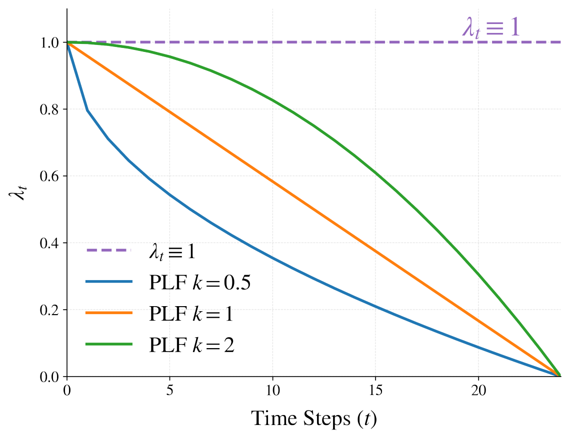


Figure 1. **Evolution of λ_t over time steps t for different PLF strategies.** λ_t determines the proportion of the relight target mixed into the fusion target.

A. Comprehensive Ablation Studies

In this section, we conduct comprehensive ablation studies to explore the effects of the hyper-parameter γ of the Consistent Light Attention **(CLA)** and various Progressive Light Fusioin **(PLF)** strategies on the quality of relighted video generation. Specifically, the values of γ are uniformly sampled within the range of $[0, 1]$, where a larger γ indicates a higher proportion of the cross-frame averaged feature in the CLA. Notably, when $\gamma = 0$, it corresponds to the vanilla IC-Light with standard self-attention. For the PLF strategy, the parameter λ_t determines the proportion of the relight target mixed into the fusion target at each step. Several different PLF strategies are also proposed, with λ_t defined as:

$$\lambda_t = 1 - \left(\frac{t}{T_m}\right)^k \tag{1}$$

Here, $T_m = 25$ denotes the total number of noise-adding steps for the source video, and different values of k indicate different rates of decay for λ_t over time. $\lambda_t \equiv 1$ means directly replacing the fusion target with the relight target for all steps. Fig. 1 illustrates the curves of λ_t as it varies with time step t.

A quantitative comparison of various settings is provided in Fig. 2, where the three evaluation metrics (FID, Temporal Clip score, and Motion Preservation score) introduced in the main text are employed to evaluate the per-frame image quality and temporal consistency of the relighted video generated by our Light-A-Video method. Specifically, Fig. 2 (a) depicts the variation of the FID score with different values of the trade-off parameter γ. An excessively large γ results in a significant degradation of the overall relighting image quality. This is attributed to the overemphasis on the cross-frame averaged feature in the CLA module, which leads to temporal over-smoothing and diminishes the lighting specificity, thereby negatively impacting the relighting effect. However, when γ is chosen appropriately (between 0.2 and 0.5), the FID score remains stable and can even be enhanced, especially when employing PLF strategies with $k = 1$ or $k = 0.5$.

The temporal consistency evaluation, as depicted in Fig. 2 (b), demonstrates a steady increase in the Temporal Clip score with the rise of the parameter γ. This trend underscores the remarkable efficacy of the CLA module in augmenting the temporal consistency of the relighted video. These results reflect that the CLA module is highly effective in enhancing the temporal consistency of the relighted video. In a parallel vein, the Motion Preservation score serves as an indicator of motion consistency with the source video. Specifically, when the value of γ is selected within the range of 0.2 to 0.5, the relighted video can achieve a high degree of motion consistency with the original video.

It is worth noting that, as evidenced by the three figures, employing a constant $\lambda_t \equiv 1$ significantly underperforms the method of progressively decreasing λ_t in PLF, both in terms of relight image quality and temporal consistency. Although a constant λ_t yields a higher Temporal Clip score when $\gamma > 0.5$, the overall motion deviates substantially from the source video, resulting in an unacceptable motion preservation effect. These results effectively demonstrate the efficacy of our PLF strategy. The explanation for this observation is twofold:

- Compared to a dynamically mixed target, a constant target with rich additional illumination information in the denoising process is more likely to deviate from the sampling trajectory of the Video Diffusion Model (VDM). When this deviation exceeds the refinement capability of the VDM, it perturbs the motion priors, consequently leading to visible temporal jitter.
- Repeatedly injecting constant relight appearance across multiple iterations is analogous to cyclically relighting the same image using the image relight model. This process causes the input distribution to progressively diverge from the training distribution of the image relight model, ultimately degrading the quality of the relighted images.

B. More Quantitative Evaluation

In this section, we provide more quantitative evaluation on Light-A-Video. Tab. 1 shows the results of Setting 2 (fore-

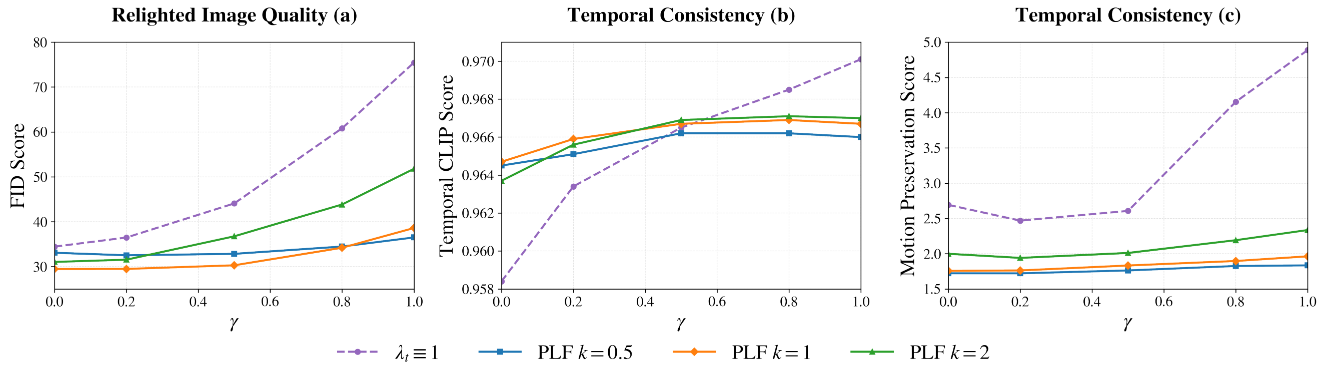


Figure 2. **The relative effectiveness of different PLF strategy on Light-A-Video performance.** (a) FID scores, (b) Temporal CLIP scores, and (c) Motion Preservation scores are shown for four strategies: PLF with constant λ ($\lambda_t \equiv 1$), and PLF with $k = 0.5, 1, 2$. Lower FID/Motion Preservation scores and higher Temporal Clip scores indicate better performance.

ground sequences relighting) on 180 foreground sequences extracted by SAM2. For a fair comparison, Motion Preservation score is only calculated on the relighted foreground objects. Then, as shown in Tab. 2, a quantitative ablation study is conducted on each Light-A-Video components. With the help of CLA module, motion consistency can be improved. Moreover, PLF module further enhance motion quality. Note that FID is calculated with vanilla IC-Light results, thus only CLA (without collaboration with VDMs) achieves a lower FID value.

Methods	FID Score (↓)	CLIP Score (↑)	Motion Preservation (↓)
IC-Light	/	0.8908	4.029
IC-Light + SDEdit-0.2	**16.44**	0.9066	3.877
IC-Light + SDEdit-0.6	46.60	0.9452	3.717
Light-A-Video (Ours)	<u>35.11</u>	**0.9777**	**1.385**

Table 1. Quantitative evaluation on Light-A-Video Setting 2.

Methods	FID Score (↓)	CLIP Score (↑)	Motion Preservation (↓)
IC-Light	/	0.9040	5.969
IC-Light + CLA	13.10	0.9308	5.279
IC-Light + CLA + PLF	29.63	0.9667	1.833

Table 2. Ablation studies on Light-A-Video components.

C. Additional Qualitative Results

In this section, we present additional qualitative results. In Fig. [3-4], we show examples of foreground sequences relighting with background generation on AnimateDiff. In Fig. [5-6], we showcase the application of Light-A-Video directly to the video relighting task. And finally, as illustrated in Fig. 7, we present the video relighting results on DiT-based video models, such as CogVideoX.

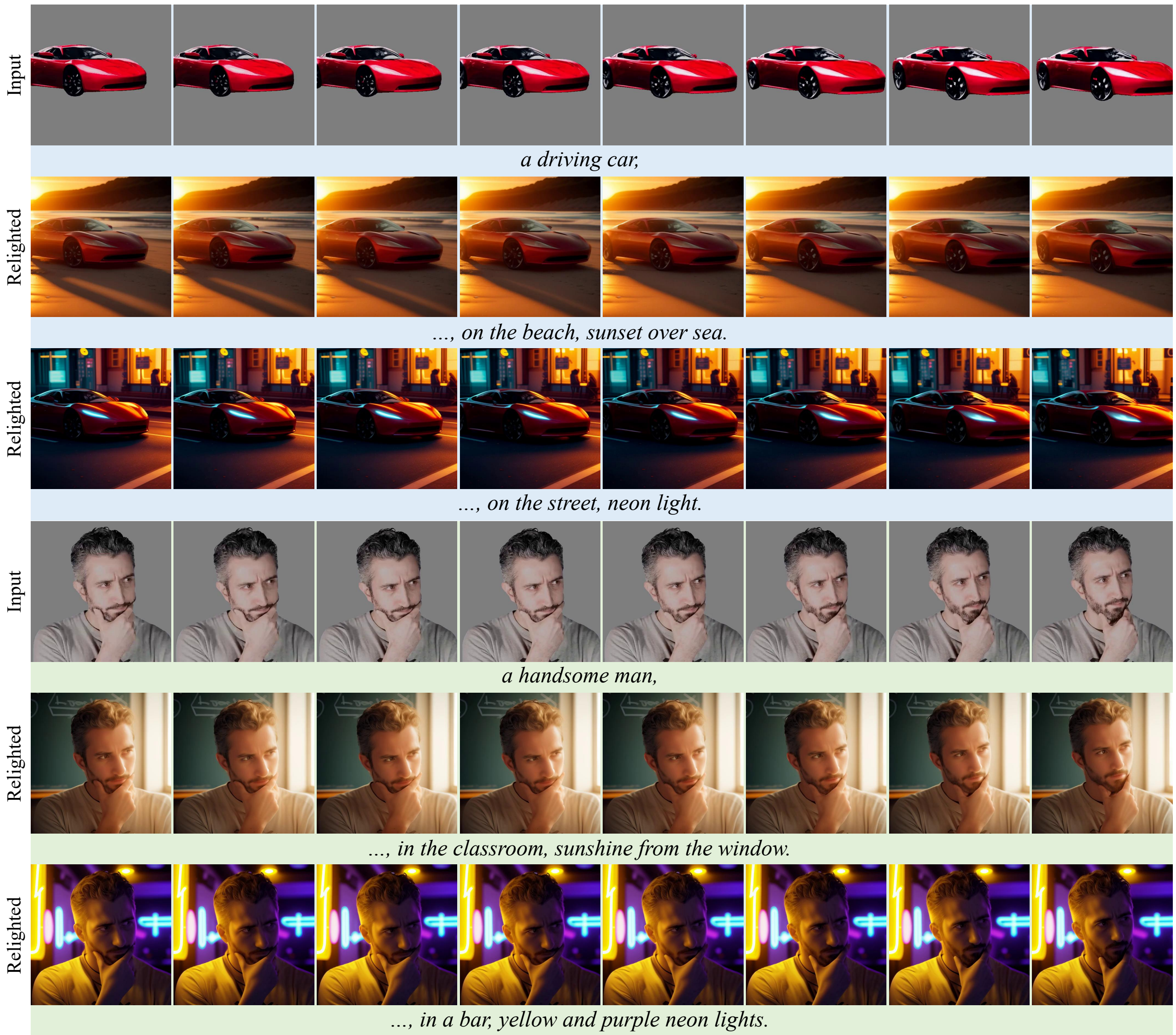
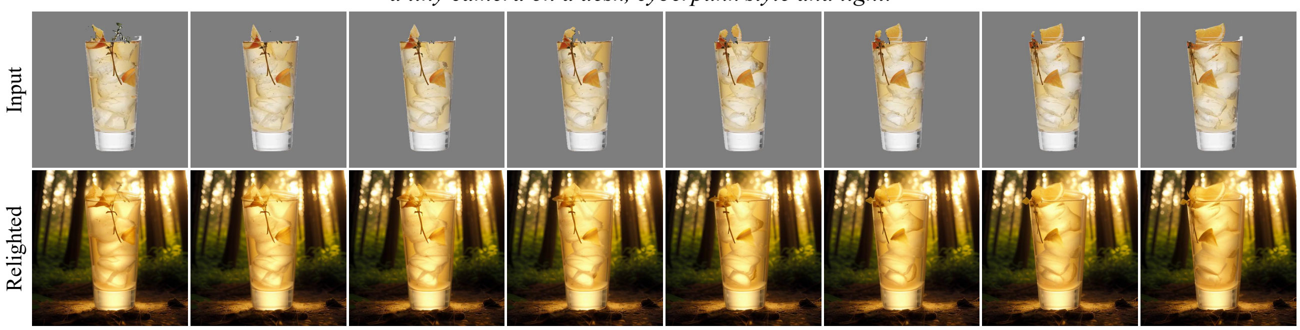


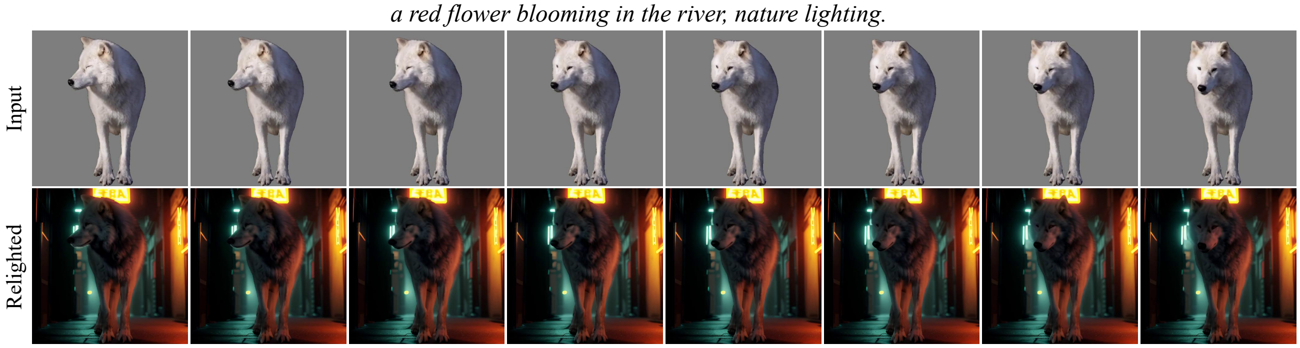
Figure 3. **More results of Light-A-Video in foreground sequences relighting with background generation.**

a tiny camera on a desk, cyberpunk style and light.



a glass of water, in the forest, magic golden lit.

a red flower blooming in the river, nature lighting.



a wolf stands in an alley, detailed face, neon, Wong Kar-wai, warm.

Figure 4. **More results of Light-A-Video in foreground sequences relighting with background generation.**

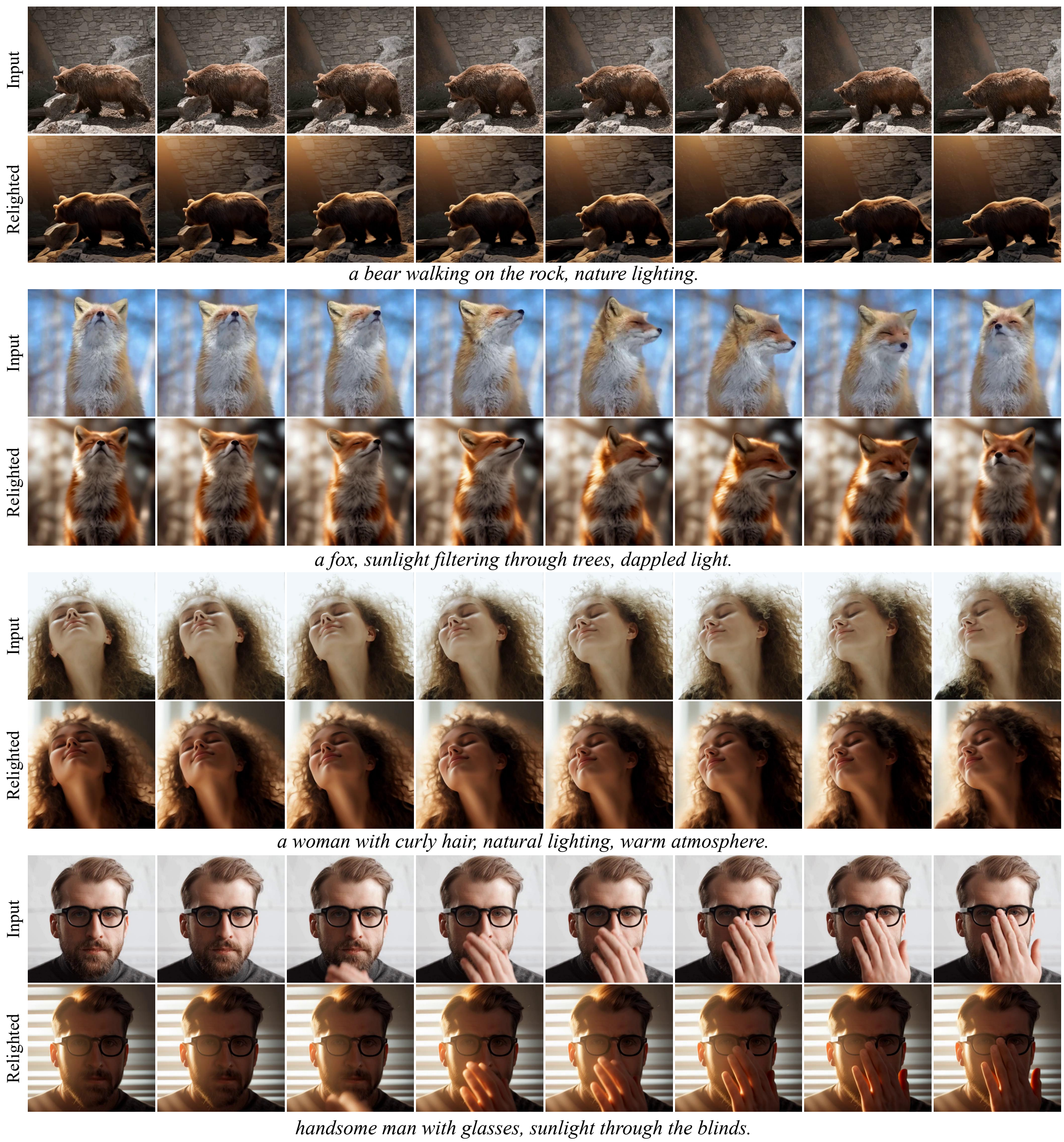


Figure 5. **More results of Light-A-Video in video sequences relighting.**

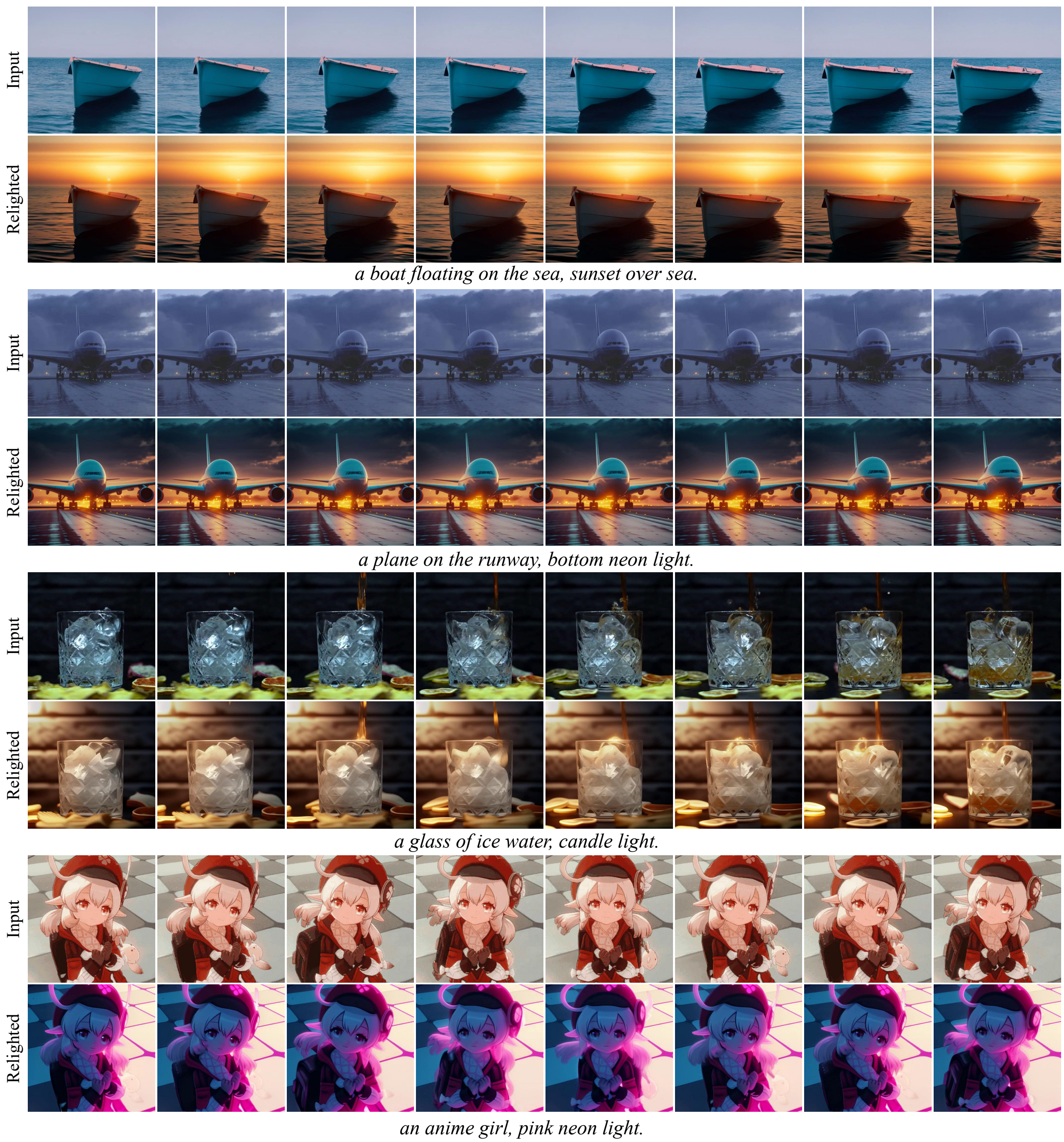


Figure 6. **More results of Light-A-Video in video sequences relighting.**

Figure 7. **More results of Light-A-Video in video sequences relighting on CogVideoX.**