# A. Appendix

## A.1. Evaluation Using VBench and Other Metrics

To further validate our model's performance and the soundness of our evaluation paradigm, we employed VBench [7], a widely adopted benchmark for video generation. Note that Hi-Light relit videos were downsampled to 1080p to satisfy VRAM constraints for the motion smoothness and dynamic degree calculations. In video relighting, the primary objective is to modify illumination while preserving the original scene's details and motion dynamics. Consequently, metrics such as Motion Smoothness, Dynamic Degree, Subject Consistency and Background Consistency should be nearly identical to those of the original video. As shown in Table 3, Hi-Light demonstrates exceptional performance, matching the original video's Motion Smoothness and Dynamic Degree almost perfectly. In contrast, competitors like TC-Light and LAV exhibit significantly increased Dynamic Degrees, indicating a possible introduction of motion artifacts. Furthermore, Hi-Light preserves Subject Consistency effectively, whereas other models degrade details in identity. Notably, LAV (AnimateDiff) achieves a counter-intuitively high Background Consistency score; we attribute this to the model producing blurred yet consistent backgrounds, which the metric misinterprets as stability.

However, VBench has distinct limitations when applied to the video relighting task rather than generation. First, the Temporal Flickering metric, which relies on pixel-wise Mean Absolute Error (MAE) for static scenes, inherently misinterprets legitimate motion in high-resolution videos as artifacts. Consequently, lower-fidelity models with blurry edges receive artificially high scores because their lack of detail minimizes pixel differences, while the sharp, high-resolution motion in original and Hi-Light videos is penalized. Second, the Aesthetic Quality metric is biased toward conventionally "pretty" (bright and saturated) images, and may unfairly penalize physically accurate but intentionally dramatic or dark lighting. Therefore, while VBench serves as a useful supplementary benchmark, our specific evaluation paradigm is better tailored to assess detail degradation and lighting stability in relighting tasks.

In addition to the Motion Smoothness and CLIP scores computed in the SOTA work [21], we further investigated the Fréchet inception distance (FID) [6]. FID is a metric used to evaluate the quality of generated models by comparing the distribution of generated images to the distribution of real images. With reference to the evaluation of FID in Light-A-Video [21], we have done a similar evaluation as shown in Table 3.

## A.2. Comprehensive Ablation Study

### A.2.1. Ablation Study on the Architecture Design

It appears that the combination of LAB-DF and Lightness Prior has a negative impact on $S_{\dot{I}}$. A possible explanation is that LAB-DF restores edges in the L channel, so the local gradients get steeper inside bright areas. The HMA-LSF reduces boundary flips, but any residual flow error causes small intensity shifts within the bright set. Together, they may amplify tiny alignment errors inside bright regions, affecting the $S_{\dot{I}}$.

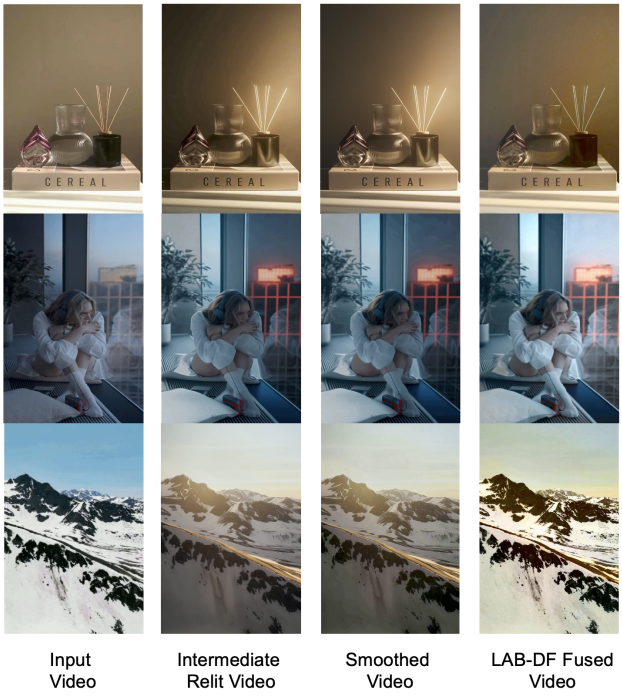


Figure 8. A visualization of the ablation study of the architecture.

As shown in Table 5, the Lightness Prior and HMA-LSF modules require around 117 seconds each, while the LAB-DF module executes much faster at only 7 seconds on average. The execution time of HMA-LSF scales strongly with resolution because both optical flow estimation and bilateral filtering are dense, pixel-level operations. Optical flow must compute and apply motion vectors for every pixel between frames, while bilateral filtering evaluates spatial–intensity neighbourhoods for each pixel to preserve edges. As resolution increases, the number of operations grows quadratically, leading to significantly higher runtimes for high-resolution videos compared to lower-resolution inputs. Also, increasing the denoising time step will increase the time taken for the lightness prior anchored and guided diffusion process.

Figure 9 shows an example of the effects of the in-

| | **Original Video** | **Hi-Light (ours)** | **LAV (Animatediff)** | **LAV (CogVideoX)** | **LAV (Wan)** | **TC-Light** |
|---|---|---|---|---|---|---|
| Dynamic Degree (↑) | 0.626 | 0.630 | 0.625 | 0.780 | 0.614 | 0.629 |
| Motion Smoothness (↑) | 0.985 | 0.982 | 0.982 | 0.983 | 0.987 | 0.982 |
| Temporal Flickering (↑) | 0.972 | 0.971 | 0.975 | 0.979 | 0.978 | 0.970 |
| Temporal Style (↑) | 0.192 | 0.194 | 0.197 | 0.192 | 0.201 | 0.205 |
| Subject Consistency (↑) | 0.939 | 0.933 | 0.917 | 0.933 | 0.935 | 0.911 |
| Background Consistency (↑) | 0.953 | 0.943 | 0.948 | 0.929 | 0.939 | 0.931 |
| Aesthetic Quality (↑) | 0.538 | 0.553 | 0.537 | 0.528 | 0.557 | 0.560 |
| Imaging Quality (↑) | 0.689 | 0.684 | 0.527 | 0.563 | 0.611 | 0.557 |
| FID (↓) | 0 | 76 | 241 | 133 | 135 | 120 |

Table 3. VBench and metrics from prior works.

| Method | | | SSIM (↑) | $S_I$(↑) | $S_c$(↑) | $S_i$(↑) | $S_{LS}$(↑) |
|---|---|---|---|---|---|---|---|
| LAB-DF | HMA-LSF | Lightness Prior | | | | | |
| ✗ | ✗ | ✗ | 0.607 | 0.165 | 0.259 | 0.430 | 0.285 |
| ✗ | ✗ | ✓ | 0.615 | 0.205 | 0.321 | 0.533 | 0.353 |
| ✗ | ✓ | ✗ | 0.612 | 0.425 | 0.450 | 0.511 | 0.462 |
| ✗ | ✓ | ✓ | 0.623 | 0.385 | 0.495 | 0.534 | 0.476 |
| ✓ | ✗ | ✗ | 0.939 | 0.295 | 0.400 | 0.475 | 0.390 |
| ✓ | ✗ | ✓ | 0.941 | 0.370 | 0.522 | 0.545 | 0.479 |
| ✓ | ✓ | ✗ | 0.942 | 0.549 | 0.523 | 0.373 | 0.482 |
| ✓ | ✓ | ✓ | **0.943** | **0.572** | **0.537** | **0.417** | **0.509** |

Table 4. A Complete Ablation Study Result on the Architecture of Hi-Light.

| Module | Time taken (s) |
|---|---|
| Lightness Prior | 98 |
| HMA-LSF | 117 |
| LAB-DF | 7 |

Table 5. The table contains the average time taken for executing each module in the architecture.

troduced light smoothing filter. The filter mitigates the light flickering problem by stabilizing the video's lighting over time. The plots for "Average Intensity" and "Number of Bright Pixels" show that the unsmoothed video (red line) exhibits large and erratic fluctuations. In contrast, the smoothed video (green line) displays much more consistent and stable values for these metrics. The most direct evidence is in the "Frame-to-Frame Change" graph. The unsmoothed video shows frequent, high-amplitude spikes, which quantitatively represent severe flicker. The smoothed video, however, maintains a line consistently closer to zero, indicating that the change in brightness between consecutive frames is smooth.

Besides the framework architecture ablation study, we also investigated how the fusion strength affects the LAB-DF process. As shown in Figure 10, as the fusion strength increases, more weight is assigned to the $L$ channel of the smoothed intermediate relit video, resulting in a decrease in both SSIM and the light stability score. Notably, the decrease in SSIM appears to be almost linear, but there is a relatively sharp drop in the Light Stability Score when the fusion strength increases from 0.3 to 0.5.

#### A.2.2. Ablation Study on The Video Diffusion Model Backbone

To determine the optimal VDM backbone for our Hi-Light framework, we conducted a rigorous ablation study evaluating three open-source models: AnimateDiff [5], CogVideoX [17], and Wan [13]. There are a total of 30 videos involved with a fixed time step of 10, and noise strength of 0.3. The quantitative results, presented in Table 6, reveal a distinct trade-off between computational efficiency and the quality of the intermediate video generated. The Wan backbone achieved superior performance, attaining the highest scores in both detail preservation and light stability. CogVideoX followed closely in quality but at a significant computational cost, requiring over 35% more processing time. Conversely, while AnimateDiff offered the fastest inference, it did so at the expense of a substantial degradation in both structural fidelity and light stability. Based on this analysis, Wan emerges as the most balanced choice. Therefore, we adopt Wan as the default VDM backbone for comparative experiments.

| VDM | SSIM (↑) | $S_I$ (↑) | $S_c$ (↑) | $S_{\dot{I}}$ (↑) | $S_{LS}$ (↑) | Ave. Time Taken (s) |
|---|---|---|---|---|---|---|
| AnimateDiff | 0.749 | 0.360 | 0.228 | 0.389 | 0.326 | 371 |
| CogVideoX | 0.932 | 0.581 | 0.459 | 0.509 | 0.517 | 966 |
| Wan | 0.940 | 0.606 | 0.461 | 0.517 | 0.528 | 712 |

Table 6. Quantitative results for the ablation study on the VDM backbones.

| # of Time Step | SSIM (↑) | $S_I$ (↑) | $S_c$ (↑) | $S_{\dot{I}}$ (↑) | $S_{LS}$ (↑) | Ave. Time Taken (s) |
|---|---|---|---|---|---|---|
| 5 | 0.952 | 0.590 | 0.407 | 0.506 | 0.501 | 617 |
| 10 | 0.956 | 0.600 | 0.467 | 0.609 | 0.559 | 771 |
| 15 | 0.951 | 0.599 | 0.472 | 0.591 | 0.554 | 1044 |
| 20 | 0.956 | 0.607 | 0.510 | 0.594 | 0.570 | 1258 |
| 25 | 0.952 | 0.607 | 0.492 | 0.598 | 0.566 | 1477 |

Table 7. Quantified results for the ablation study on the number of time steps.

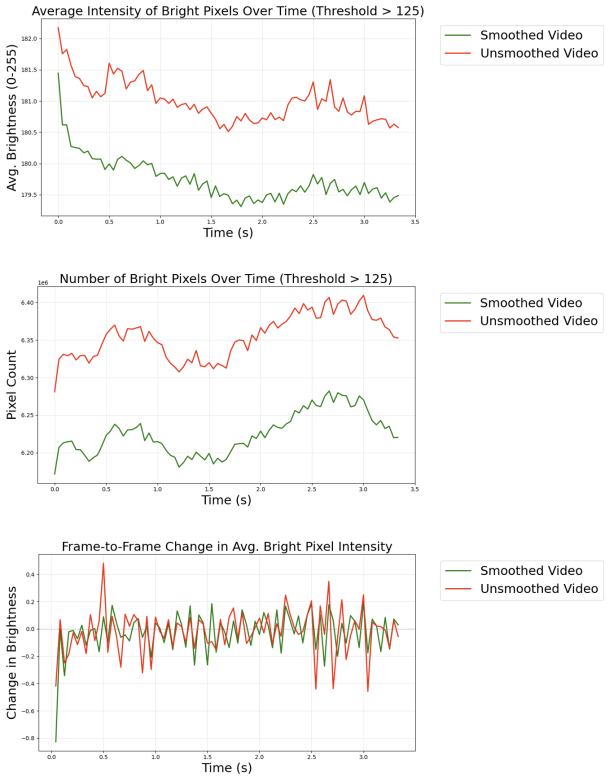


Figure 9. The plot shows the effect of the hybrid motion-adaptive light smoothing filter.

### A.2.3. Ablation Study on The Time Step

To study the effect of the number of denoising time steps in our framework, which governs the fundamental trade-off between light stability and computational cost. We performed a rigorous ablation study on a sample of 30 videos, with results presented in Table 8. Our analysis reveals a non-monotonic relationship between the number of steps

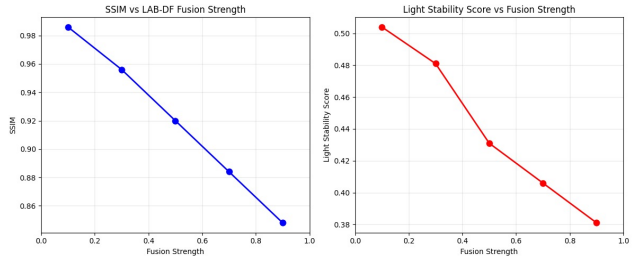


Figure 10. The plot shows how the strength of LAB-DF fusion affects SSIM and Light Stability Score.

and the final output quality. While computational cost scales linearly, the Light Stability Score $S_{LS}$ improves from 0.501 to a peak of 0.570 at 20 steps, a 13.9% gain in the light stability. Beyond this point, however, we observe diminishing returns; increasing to 25 steps not only incurs a substantial additional runtime of over 200 seconds but also results in a slight degradation of both stability ($S_{LS}$ drops to 0.566) and detail preservation. These empirical results lead us to a principled choice, identifying 20 time steps as the optimal setting for our framework, as it maximises temporal coherence without introducing unnecessary computational overhead or performance degradation. Notably, 10 steps can also yield a quality result in a much shorter time.

### A.2.4. Ablation Study on The Noise Addition

The ablation study on noise strength, conducted over 30 videos, shows a clear trade-off between structural fidelity and lighting stability. As noise strength increases from 0.1 to 0.6, SSIM decreases gradually, indicating progressive loss of fine-grained details. Meanwhile, the Light Stability Score also drops, reflecting reduced temporal smoothness. Interestingly, moderate noise levels (0.3–0.4) yield a balance, where the light stability scores remain relatively high while detail preservation is only slightly degraded. How-

| Noise Strength | SSIM (↑) | $S_I$ (↑) | $S_c$ (↑) | $S_{\dot{I}}$ (↑) | $S_{LS}$ (↑) |
|---|---|---|---|---|---|
| 0.1 | 0.937 | 0.674 | 0.497 | 0.486 | 0.552 |
| 0.2 | 0.936 | 0.635 | 0.492 | 0.492 | 0.540 |
| 0.3 | 0.934 | 0.609 | 0.489 | 0.492 | 0.530 |
| 0.4 | 0.932 | 0.642 | 0.471 | 0.490 | 0.534 |
| 0.5 | 0.930 | 0.609 | 0.475 | 0.527 | 0.537 |
| 0.5 | 0.925 | 0.572 | 0.500 | 0.462 | 0.511 |
| 0.6 | 0.923 | 0.500 | 0.462 | 0.465 | 0.476 |

Table 8. Quantified results for the ablation study on the diffusion noise strength.

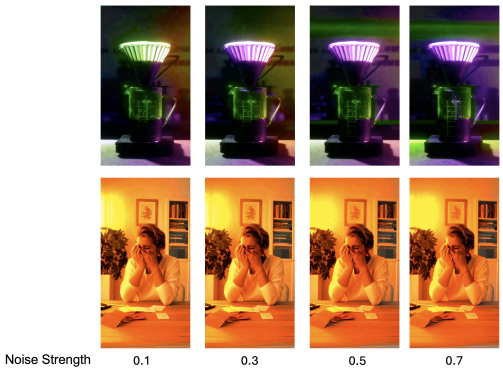


Figure 11. A visualisation of the effect of noise strength.

ever, beyond 0.5, both stability and fidelity decline sharply, suggesting over-noising destabilises the diffusion process.

Although lower noise strengths tend to produce more stable and detail-preserving results, they can also limit the diversity of relighting outcomes by keeping the latent space of the relit video too close to the original input. Increasing the noise strength injects additional stochasticity into the diffusion process, enabling the model to explore a broader range of illumination effects and generate more varied relighting results as shown in Figure 11. However, this comes at the cost of reduced structural fidelity and light stability, highlighting an inherent trade-off. In practice, noise strength thus becomes a tunable option that users can adjust depending on whether they prioritise stability and fidelity or prefer more diverse lighting variations.

### A.3. Ablation Study on the $\alpha$ Term in the Optical Flow Light Smoothing Filter

The ablation results show that increasing the fixed $\alpha$ from 0.1 to 0.9 consistently improves all light-stability metrics, while SSIM remains nearly unchanged. This indicates that a large $\alpha$ provides the best trade-off for suppressing highlight flicker without visibly degrading per-frame details. However, at very high values, we occasionally observe subtle motion trails in fast-moving regions when inspected closely. In contrast, the adaptive $\alpha$ variant achieves slightly better performance across all light stability scores. By reducing the smoothing weight only in high-flow areas, the adaptive formulation follows the temporal behaviour of the fixed $\alpha$ but avoids oversmoothing motion boundaries. The adaptive scores track just below the corresponding fixed-$\alpha$, confirming that they preserve most of the temporal stability while preventing motion trails on fast-moving subjects.

| Fixed $\alpha$ | SSIM | $\mathbf{S_I}$ (↑) | $\mathbf{S_c}$ (↑) | $\mathbf{S_{\dot{I}}}$ (↑) | $\mathbf{S_{LS}}$ (↑) |
|---|---|---|---|---|---|
| 0.1 | 0.657 | 0.349 | 0.378 | 0.500 | 0.409 |
| 0.2 | 0.657 | 0.362 | 0.386 | 0.509 | 0.419 |
| 0.3 | 0.656 | 0.374 | 0.394 | 0.513 | 0.427 |
| 0.4 | 0.656 | 0.389 | 0.405 | 0.528 | 0.441 |
| 0.5 | 0.655 | 0.401 | 0.416 | 0.541 | 0.453 |
| 0.6 | 0.655 | 0.418 | 0.428 | 0.553 | 0.466 |
| 0.7 | 0.654 | 0.436 | 0.439 | 0.561 | 0.478 |
| 0.8 | 0.653 | 0.450 | 0.453 | 0.568 | 0.490 |
| 0.9 | 0.654 | 0.460 | 0.463 | 0.574 | 0.499 |
| Adaptive $\alpha$ | | | | | |
| 0.1 | 0.657 | 0.354 | 0.381 | 0.502 | 0.412 |
| 0.3 | 0.657 | 0.374 | 0.395 | 0.510 | 0.426 |
| 0.5 | 0.656 | 0.403 | 0.419 | 0.542 | 0.455 |
| 0.7 | 0.655 | 0.438 | 0.447 | 0.556 | 0.480 |
| 0.9 | 0.653 | 0.470 | 0.467 | 0.584 | 0.507 |

Table 9. Ablation results for the fixed and adaptive $\alpha$ terms in the optical-flow light smoothing filter.

### A.4. Ablation Study on the $\beta$ Term in LAB-DF

We conducted an ablation study on 60 videos to examine the effect of the merging-strength parameter $\beta$. As shown in Table 10, increasing $\beta$ leads to a consistent drop in both SSIM and $S_{LS}$, indicating reduced detail preservation and lower light stability. A greater drop occurs between $\beta = 0.4$ and $\beta = 0.7$, suggesting that excessively large merging strengths overly suppress the fine-grained information during LAB-DF. Notably, even with $\beta$ as high as 0.9, our method continues to surpass all baseline methods by a significant margin, demonstrating robust performance under extreme merging strengths.

### A.5. Ablation Study on Videos with Different Motion Speeds

We further evaluate the robustness of our framework under varying scene dynamics. A total of 60 videos with manually selected motion levels (equally split among fast, medium, and slow) were used in this ablation. The results in Table 11 show that motion speed has a negligible impact on detail preservation. SSIM varies by at most 0.005 across the three groups, and the change in $S_{LS}$ remains within 1%. These findings indicate that our proposed method maintains consistent performance regardless of motion intensity, demonstrating strong robustness to variations in scene dynamics.

| $\beta$ | SSIM | $S_I$ (↑) | $S_c$ (↑) | $S_{\dot{I}}$ (↑) | $S_{LS}$ (↑) |
|---|---|---|---|---|---|
| 0.1 | 0.974 | 0.485 | 0.474 | 0.536 | 0.498 |
| 0.2 | 0.962 | 0.475 | 0.457 | 0.534 | 0.489 |
| 0.3 | 0.945 | 0.461 | 0.448 | 0.527 | 0.479 |
| 0.4 | 0.926 | 0.436 | 0.433 | 0.518 | 0.462 |
| 0.5 | 0.906 | 0.408 | 0.419 | 0.506 | 0.444 |
| 0.6 | 0.887 | 0.373 | 0.399 | 0.494 | 0.422 |
| 0.7 | 0.868 | 0.369 | 0.387 | 0.491 | 0.416 |
| 0.8 | 0.850 | 0.359 | 0.379 | 0.477 | 0.405 |
| 0.9 | 0.833 | 0.347 | 0.363 | 0.487 | 0.399 |

Table 10. Merge Strength Ablation Study results.

| Motion | SSIM | $S_I$ (↑) | $S_c$ (↑) | $S_{\dot{I}}$ (↑) | $S_{LS}$ (↑) |
|---|---|---|---|---|---|
| Slow | 0.947 | 0.468 | 0.459 | 0.511 | 0.480 |
| Medium | 0.952 | 0.501 | 0.461 | 0.537 | 0.499 |
| Fast | 0.949 | 0.487 | 0.454 | 0.530 | 0.490 |

Table 11. Ablation on motion speed. Performance metrics across videos with slow, medium, and fast motion show minimal variation, indicating strong robustness to scene dynamics.

## A.6. Comparative Experimental Results on Models' Runtime

We measured the average time taken for the models to relight 60 videos, each of which has 81 frames. This set of experiments was conducted using an H100 GPU.

| Model | Scaled Average Relighting Time |
|---|---|
| LAV (AnimateDiff) | 276s |
| TC-Light | 398S |
| LAV (Wan) | 476s |
| LAV (CogVideoX) | 503s |
| Hi-Light (ours) | 530s |

Table 12. Runtime comparison results in ascending order. The average time taken for the models to relight videos of 81 frames.

## A.7. Empirical Choice of the Brightness Threshold $\tau$ in $S_{LS}$ Calculation

The brightness threshold $\tau = 125$ in the calculation of $S_{LS}$ is an empirical selection. As shown in the Figure 12, the distribution of the relit video clearly shifts toward higher brightness compared to the original, with a pronounced increase in pixel density beginning around L = 120 to 130. Below this range, both original and relit distributions overlap considerably, indicating that fluctuations are more likely due to scene content or motion rather than relighting. By contrast, above 125, the relit histogram diverges strongly

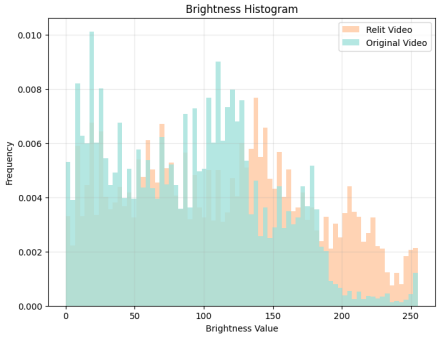


Figure 12. Histogram of pixel brightness of source videos and relit videos.

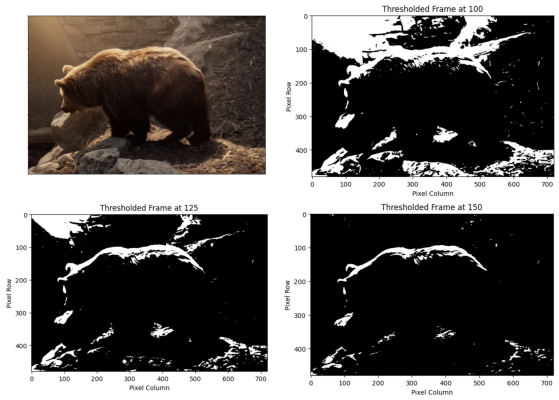


Figure 13. Visualisation of a relit video frame at different brightness.

from the original, capturing the excess brightness introduced by relighting. Choosing 125, therefore, strikes a balance: it is high enough to exclude darker regions where variations are unrelated to illumination changes, while low enough to consistently include the portion of the distribution where relighting effects dominate. This cutoff ensures that the $S_{LS}$ focuses on the visually meaningful relighted regions without being contaminated by background fluctuations and motion.

The Figure 13 also illustrates the effect of different luminance thresholds (100, 125, 150) on isolating bright regions within a frame. At a low threshold of 100, the mask includes a large portion of the background and textured regions, potentially introducing noise unrelated to relighting. At the high threshold of 150, only a few sparse highlights remain, missing much of the relevant relit areas on the bear's back and surrounding surfaces. In contrast, the threshold of 125 yields a relatively balanced mask: it successfully captures the prominent illuminated regions (e.g., along the bear's contour and the lit rock surfaces) while excluding most of the darker, non-relit areas. This visually demonstrates that 125 is high enough to avoid contamination from shadow and

| Model | 105 | 115 | 125 | 135 | 145 |
|---|---|---|---|---|---|
| TC-Light | 0.300 | 0.296 | 0.285 | 0.277 | 0.238 |
| LAV (AnimateDiff) | 0.119 | 0.123 | 0.109 | 0.115 | 0.120 |
| LAV (CogVideoX) | 0.202 | 0.208 | 0.209 | 0.231 | 0.239 |
| LAV (Wan) | 0.263 | 0.278 | 0.271 | 0.272 | 0.270 |
| Hi-Light | **0.515** | **0.509** | **0.520** | **0.521** | **0.536** |

Table 13. Sensitivity test results for the brightness threshold $\tau$. The table shows the $S_{LS}$ under different $\tau$ values.

texture variation, yet good enough to preserve the meaningful bright zones where relighting actually occurs.

### A.7.1. Sensitivity Study on the Brightness Threshold $\tau$

To examine the robustness of the proposed $S_{LS}$ against variations in the brightness threshold $\tau$, we conducted a sensitivity analysis across values ranging from 105 to 145. As shown in Table 13, the relative ranking of models remains consistent regardless of the chosen threshold, with Hi-Light achieving the highest scores across all cases. While small numerical fluctuations are observed as $\tau$ varies, the overall performance gap between Hi-Light and competing methods remains substantial. This consistency indicates that $S_{LS}$ is not overly sensitive to the precise threshold choice and that our evaluation reliably captures temporal lighting stability. Importantly, the stability of Hi-Light improves slightly at higher thresholds, suggesting that the framework not only generalises across different luminance regimes but also benefits when the evaluation focuses on regions with more prominent illumination. Together, these findings confirm that our metric and results are robust and not an artefact of the specific choice of $\tau$.

## A.8. Human Survey Result Analysis

To quantitatively substantiate the perceptual relevance of our proposed Light Stability Score, we performed a Spearman's rank correlation analysis between our metric's rankings and the results of a 30-participant human evaluation. As shown in Figure 14 and Table 14, the analysis yielded a perfect positive correlation (Spearman's $\rho = 1.0$), indicating that the model rankings produced by our metric are identical to those derived from human judgment. This statistically significant alignment provides strong empirical evidence that the Light Stability Score serves as an effective and reliable proxy for human perception of light stability in the video relighting task.

| Model | $S_{LS}$ | $S_{LS}$ Rank | Ave. Human Rank | Human Rank |
|---|---|---|---|---|
| Hi-Light (ours) | 0.509 | 1 | 1.14 | 1 |
| TC-Light [10] | 0.281 | 2 | 2.59 | 2 |
| LAV (Wan) [21] | 0.279 | 3 | 2.92 | 3 |
| LAV (CogVideoX) [21] | 0.267 | 4 | 3.86 | 4 |
| LAV (AnimateDiff) [21] | 0.098 | 5 | 4.50 | 5 |

Table 14. Comparison between the results of Light Stability Score and human survey.

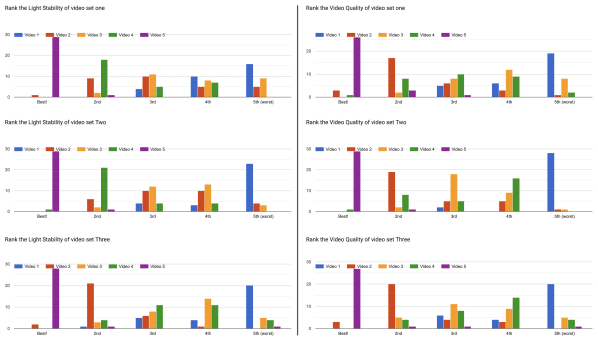


Figure 14. The left side shows the results of the human survey for light stability. The right side shows the results of the human survey for video quality. Video 1: LAV-AnimateDiff, Video 2: LAV-Wan, Video 3: LAV-CogVideoX, Video 4:TC-Light, Video 5:Hi-Light.

## A.9. Additional Plot Results and Showcases

This section provides more visual results for the paper. The plots below provide a clear visual comparison of the light stability of the relit results. Hi-Light (ours) demonstrates the best smoothness in the plots among all the baselines.

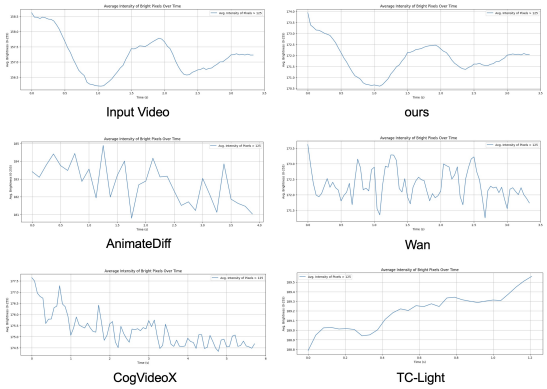


Figure 15. Plot of average intensity of bright pixels (brightness$\geq$125) over time.

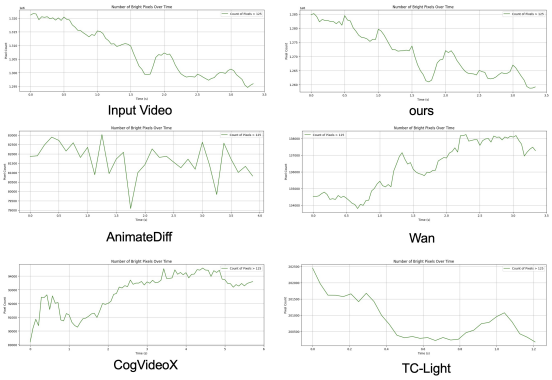


Figure 16. Plot of number of bright pixels (brightness$\geq$125) over time.

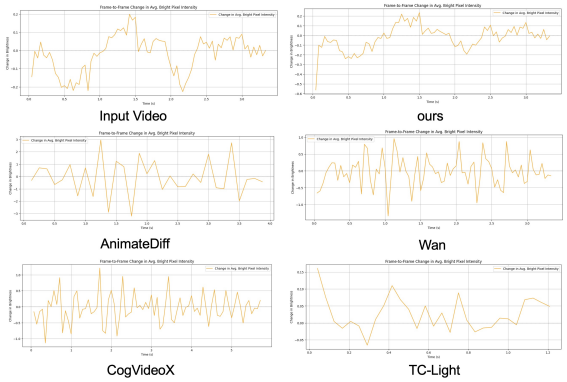


Figure 17. Plot of frame-to-frame change in average bright pixel (brightness$\geq$125) intensity.

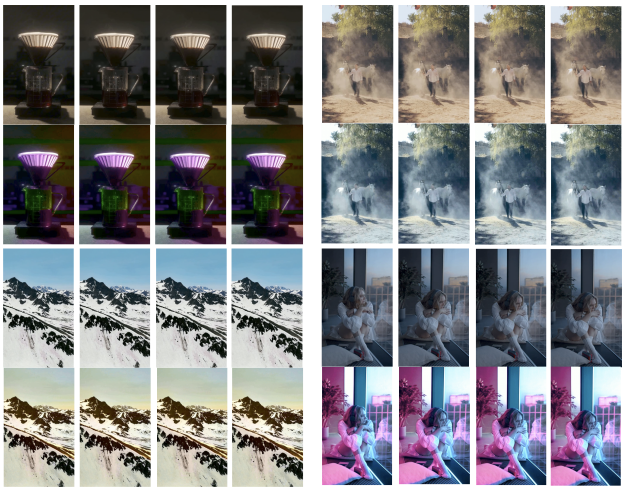

Figure 18. More Hi-Light relit video showcase.

Figure 19. More Hi-Light relit video showcase.