WildRelight: A Real-World Dataset and Benchmark for Single-Image Relighting

Supplementary Material

1. Physics-Guided Test-Time Adaptation

To illustrate how WildRelight's real-world supervision can be practically exploited, we design a reference framework that integrates physics-guided posterior sampling for inverse decomposition regularization, together with sampling-aware TTA to better align forward relighting dynamics.

Rather than aiming to introduce a fully optimized solution, this framework serves as a structured case study demonstrating how dataset-driven supervision can be incorporated at both inference and adaptation stages. The resulting bidirectional consistency between scene representation and image formation highlights the practical value of WildRelight, while leaving substantial room for future methodological improvements.

1.1. Physics-Guided Inverse Rendering via Diffusion Posterior Sampling

To enforce physical validity in G-buffer prediction without retraining, we introduce a physics-guided inference strategy based on Diffusion Posterior Sampling [4]. At diffusion timestep t, the estimated clean latent $\hat{x}_0$ is decoded by the VAE $D(\cdot)$ into intrinsic components and rendered under illumination L via a differentiable Cook–Torrance renderer $\mathcal{R}(\cdot)$ [5, 33]. We enforce consistency with the observed image I_{gt} via a measurement loss:

$$\mathcal{L}_{\text{render}} = \left\| \mathcal{R}\big(D(\hat{x}_0),\, L\big) - I_{\text{gt}} \right\|_2^2 . \tag{1}$$

Guided by the gradient $g_t = \nabla_{x_t} \mathcal{L}_{\text{render}}$, the DDIM sampling trajectory is refined as:

$$x_{t-1} \leftarrow x_{t-1} - \zeta_t \, g_t, \tag{2}$$

where ζ_t is the guidance strength. To ensure stability and efficiency during this optimization, we employ a split-sum approximation for image-based lighting while keeping the diffusion network parameters frozen.

1.2. Evaluation of Proposed Methodology

Evaluation Protocol. Unlike global finetuning which relies on fixed splits, our inference-time methodology (DPS + TTA) enables instance-specific adaptation. We evaluate this universality across all 30 scenes using a "leave-one-lighting-out" protocol: for N lightings, one is the test target while the other $N-1$ serve as self-supervised signals. This rigorously simulates real-world deployment without ground-truth supervision.

1.2.1. Ablation Study and Analysis

We dissect the contribution of each component by incrementally integrating them into the pre-trained DiffusionRenderer [16] baseline. The quantitative and qualitative results, averaged over the full dataset, are presented in Table 4 and Figure 7.

Baseline (Zero-Shot). The synthetic-trained baseline suffers a substantial sim-to-real gap (21.63 dB PSNR), failing to disentangle complex outdoor illumination and reflectance. This deficit quantifies the limitations of synthetic data rather than the architecture. We use this baseline to represent synthetic priors, demonstrating that real-world adaptation—enabled by our data—is critical for bridging this gap.

Impact of Physics-Guided DPS. Introducing Diffusion Posterior Sampling (+DPS) injects physical constraints, yielding consistent gains (+0.95 dB PSNR). Crucially, DPS acts as a geometric anchor: by enforcing rendering equation consistency, it prevents hallucinated shadows, ensuring generation adheres to the underlying scene structure.

Impact of Temporal TTA. TTA alone drastically minimizes photometric error (+2.5 dB PSNR) but slightly degrades LPIPS ($0.390 \rightarrow 0.392$). This reveals a photometric-perceptual trade-off in unconstrained self-supervision: optimization overfits pixel intensities at the cost of the natural image manifold, leading to artifacts that penalize perceptual metrics.

Synergy: Constrained Adaptation. The full pipeline (+DPS & TTA) resolves this trade-off, achieving the best overall performance (25.04 dB PSNR, 0.345 LPIPS). Here, DPS regularizes the TTA trajectory, ensuring the model adapts to scene-specific lighting dynamics without sacrificing physical plausibility or high-frequency details.

Comparison with Global Finetuning. Remarkably, our inference-time approach rivals fully supervised global finetuning reported in Sec. 4.2 (25.04 vs. 25.95 dB). This underscores our framework's efficiency: effectively bridging the sim-to-real gap with near-supervised performance without expensive retraining.

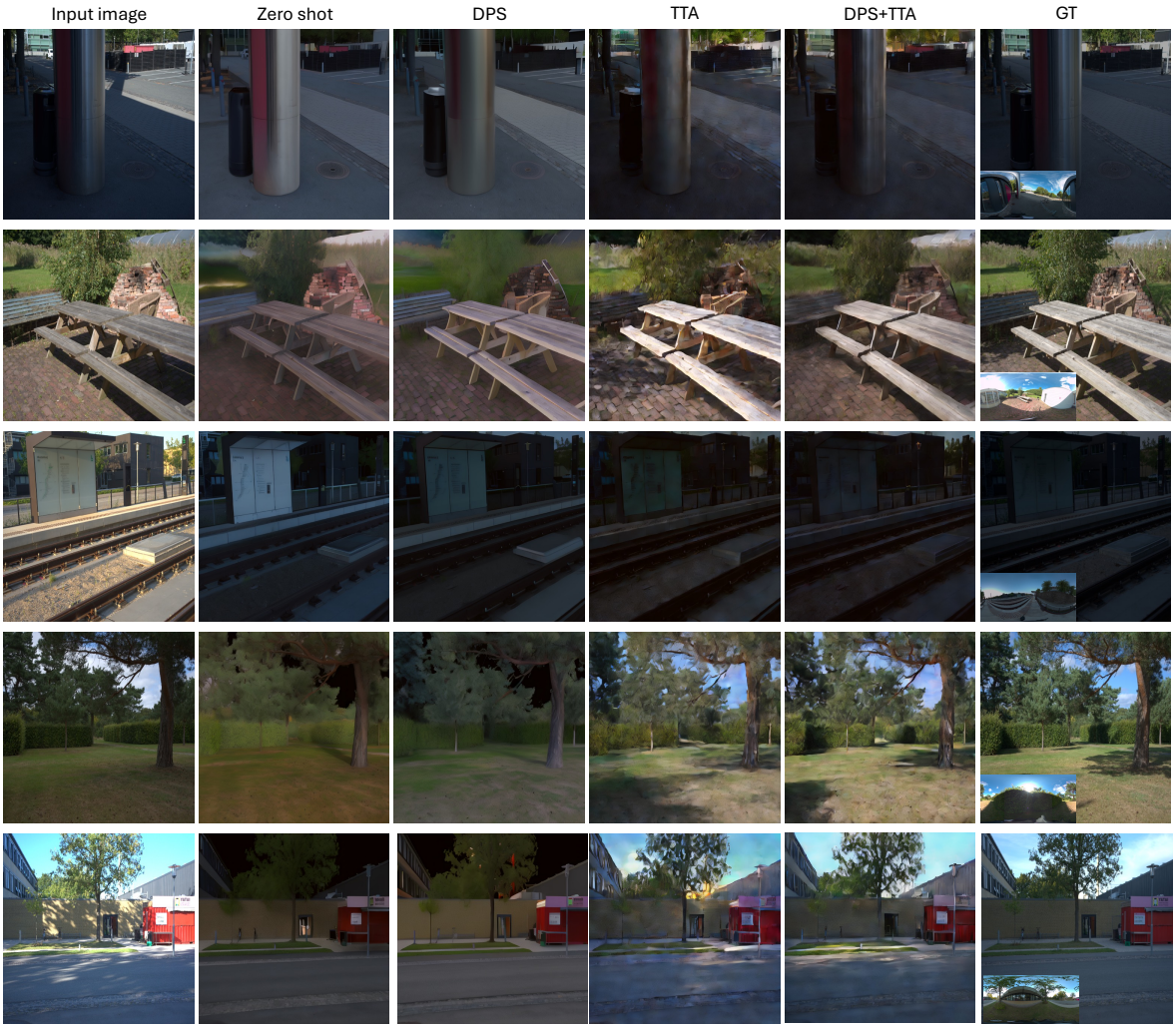

Figure 7. Qualitative Ablation Study Results. We visualize the results of our proposed framework on the *WildRelight* dataset. The envmap on bottom left of GT image correspond to target illuminations used for relighting.

Table 4. **Ablation Study of the Proposed Methodology.** Evaluated on the full 30-scene *WildRelight* dataset. While TTA alone significantly improves PSNR, it inadvertently degrades perceptual quality (higher LPIPS) due to unconstrained optimization overfitting to pixel-level intensities. Incorporating DPS restores physical consistency, and their combination yields the best overall performance, effectively bridging the sim-to-real gap without supervised training.

Configuration	Mechanism	PSNR ↑	SSIM ↑	LPIPS ↓
Baseline (Pre-trained)	Zero-shot	21.63	0.6311	0.3901
+ DPS	Inference Prior	22.58	0.6578	0.3825
+ TTA	Optimization	24.10	0.6451	0.3923
+ DPS & TTA	**Constrained Adapt.**	**25.04**	**0.6829**	**0.3453**

2. Quantitative Validation of Illumination Alignment

We provide a rigorous quantitative validation based on metadata timestamp statistics and solar angular displacement analysis. This proves that the temporal gap in our acquisition pipeline results in physically negligible illumination misalignment for the task of relighting.

2.1. Temporal Synchronization Statistics

We extracted and analyzed the acquisition timestamps from the metadata of all image pairs (Sony A7 scene images and Insta360 envmaps). The distribution of the time delay Δt is summarized in Table 5.

Table 5. Statistics of Temporal Synchronization (Δt) between scene and environment map capture.

Metric	Value
Median Time Delta	38.00 seconds
Mean Time Delta	40.14 seconds
Max Time Delta	114.00 seconds

While the maximum delay is approximately 1.7 minutes, we demonstrate below that this is well within the tolerance for accurate outdoor lighting estimation.

2.2. Physical Error Analysis: Solar Angular Displacement

The primary source of directional illumination outdoors is the sun. The Earth rotates at approximately 15° per hour (0.00417° per second). We calculate the angular displacement of the sun, θ_{shift}, caused by the capture delay:

$$\theta_{\text{shift}} = \Delta t \times 0.00417^\circ/s \tag{3}$$

For our mean delay (40.14s):

$$\theta_{\text{mean}} \approx 0.17^\circ \tag{4}$$

For our worst-case delay (114s):

$$\theta_{\text{max}} \approx 0.48^\circ \tag{5}$$

To contextualize these values, the angular diameter of the sun is approximately 0.5°. In the average case, the sun moves only $1/5$ of its own diameter, which is perceptually imperceptible in lighting effects. Even in the worst-case scenario, the displacement (0.48°) is still less than the angular size of the light source itself (0.5°).

2.3. Impact on Relighting Tasks

Modern single-image relighting algorithms typically operate on environment maps that are downsampled (e.g., to 256×128) to estimate lighting coefficients (like Spherical Harmonics) or to condition diffusion models.

At a width of 256 pixels, the horizontal angular resolution is $360^\circ/256 \approx 1.4^\circ$ per pixel.

$$\text{Pixel Shift}_{\text{max}} = \frac{0.48^\circ}{1.4^\circ/\text{pixel}} \approx 0.3 \text{ pixels} \tag{6}$$

Thus, even our maximum temporal delay results in a sub-pixel shift (0.3 pixels) in the effective resolution used by state-of-the-art relighting models. This quantitatively confirms that our dataset maintains high-quality illumination alignment suitable for training and evaluating inverse rendering methods.

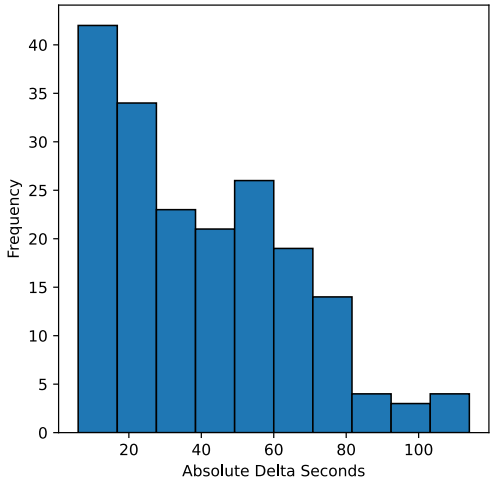

Figure 8. Distribution of capture time differences (Δt). The histogram shows the frequency of absolute time delays between the scene image and the environment map capture across the dataset. The distribution is heavily right-skewed, with the vast majority of samples having a delay of less than 20 seconds, confirming that large delays are rare.

We also plot the distribution of the absolute time differences (Δt) in Figure 8. As illustrated, the distribution is strongly skewed towards zero. The dominant peak in the first 3 bins (< 40s) corroborates that for the majority of our collected scenes, the temporal gap is minimal. The long tail extending to ~ 100s represents a small fraction of outlier cases. Combined with the solar angular displacement analysis, this distribution confirms that the impact of these delays on illumination alignment remains physically negligible for the purpose of relighting.

3. Dynamic Element Masking

A significant challenge in our longitudinal, "in-the-wild" capture is the inevitable presence of dynamic scene elements, such as wind-blown foliage and moving cloud formations, despite a static camera rig. To preserve the photometric integrity of our GT images, we avoid computational alignment (e.g., warping) which would alter the pixel data. Instead, we provide meticulously hand-annotated binary masks for all non-static regions. This approach allows researchers to optionally exclude these dynamic areas during metric computation, thereby isolating the evaluation of relighting performance from artifacts caused by scene motion. After determining that automated methods were unreliable for our complex natural scenes, we developed a rigorous manual

Figure 9. Advantage of RAW format with HDR. When using HDR photo, the details of the photo are preserved. Therefore, by adjusting the exposure settings, the original colors of the photo can be accurately restored.

Figure 10. Example of Dynamic Scene Elements. As showed in the figure, the left and right images were captured at different times with a fixed camera position. Despite the static camera setup, subtle movements of the leaves occur due to external factors such as wind. To address this, we manually created masks for these dynamic regions, allowing researchers to determine whether to include them when computing metrics.

annotation pipeline based on pairwise temporal comparisons. The complete details of this pipeline, including our annotation interface and specific exclusion criteria (e.g., for water and reflections), are available in the supplementary materials.

4. Advantages of RAW-Based HDR Image

The data recorded by a camera sensor in RAW format exhibits a fundamentally linear relationship with the light intensity of the actual scene. This linearity is a core advantage in computational photography, particularly for HDR image synthesis. Leveraging this linear property of RAW data allows us to directly synthesize an HDR image by simply performing a linear combination of multiple exposures. This method circumvents the complex process of calculating and inverting the Camera Response Function (CRF). It significantly simplifies the HDR generation pipeline while improving fidelity.

In contrast to the RAW format, the conventional JPG format has inherent limitations for HDR applications. JPG files are typically stored and compressed using an 8-bit color depth, a process that results in a significant loss of color and luminance information. This loss of information is particularly severe under extreme lighting conditions, as illustrated in Figure 9. The key differences are:

1. **Shadow Detail and Color Fidelity:** In low-light environments, the RAW format, with its high bit depth (typically 12 or 14 bits), captures extensive detail in the dark regions. By increasing the exposure in post-processing, the original information can be recovered with minimal loss. Conversely, since this information is already discarded during the in-camera processing of a JPG image, attempting to boost its exposure does not recover meaningful detail and instead leads to severe color distortion, banding, and noise, as shown in Fig. 9.
2. **Highlight Information Retention:** In highlight regions, while both a RAW-based HDR image and a JPG image may appear as pure white on a Standard Dynamic Range (SDR) display due to exceeding the display's maximum brightness, the amount of information they contain is fundamentally different. The RAW data fully retains the color and tonal information within these bright areas. When the exposure is reduced in post-processing, the details in these "clipped" areas can be clearly recovered. As shown in the cloud section of Fig. 9, at an exposure compensation of -2EV, the RAW-based HDR image reveals subtle gradations in the clouds. The corresponding area in the JPG image, however, remains a flat white expanse because the information was permanently lost.

RAW-based HDR synthesis offers unparalleled advantages over the JPG format, both in the recovery of shadow detail and the preservation of highlight information.

Figure 11. Side-by-side camera setup, a non-aligned envmap camera will record direct sun light, but scene camera records a shadow.

4.1. Necessity of Strict Spatial Alignment

Precise co-location of the environment map camera (Insta360) and the scene camera (Sony A7) is a prerequisite for pixel-aligned evaluation.

Spatial Parallax is Fatal. A non-confocal setup (e.g., a standard side-by-side placement with a 10cm baseline) fundamentally alters occlusion relationships between the scene geometry and the light source. For instance, foliage that occludes the sun in the environment camera's view may not occlude it in the scene camera's view. This discrepancy creates "false" shadows in the Ground Truth, shadows that exist in the illumination map but not in the photograph (or vice versa). Such geometric inconsistencies render strict, pixel-aligned quantitative evaluation impossible.

5. Methodology for Determining the Nodal Point (No-Parallax Point)

In the domain of photographic composition, the elimination of parallax error is of paramount importance. Parallax manifests as an apparent displacement of foreground objects relative to the background as the camera is rotated. To mitigate this artifact, it is crucial to rotate the camera system around a specific pivot point. While colloquially referred to as the "nodal point," the technically precise term for this locus is the center of the entrance pupil. The entrance pupil represents the virtual image of the physical aperture stop as viewed through the front elements of the lens. It is the conjugate point through which all chief rays appear to pass before refraction. Rotating the camera and lens assembly around the center of the entrance pupil ensures that the perspective remains consistent across multiple exposures, thereby facilitating seamless image stitching.

The empirical determination of the entrance pupil's location, or the no-parallax point, is a foundational procedure in panoramic photography. The following methodology outlines a systematic approach to identifying this point.

5.1. Experimental Setup

The camera must be securely mounted on a panoramic head affixed to a stable tripod. This specialized head allows for the adjustment of the camera's position along the longitudinal axis of the lens. For the purpose of this procedure, a focal length of 40mm was selected. Two distinct, vertically-oriented objects, positioned at different distances from the camera, were chosen as reference points. A lamppost and a more distant utility pole served as suitable subjects. The camera's position was initially adjusted so that, when viewed through the camera's live-view display, the nearer reference object precisely occluded the more distant one at the center of the frame.

5.2. Procedure for Parallax Elimination

The core of the methodology lies in an iterative process of rotation and observation. With the reference objects aligned, the camera is panned horizontally by an angle of approximately 30 degrees to the left and then to the right. The relative position of the two reference objects is carefully observed during this rotation. If the nearer object appears to shift its position relative to the farther object, parallax is present. This indicates that the axis of rotation is not coincident with the entrance pupil. Adjustments must then be made to the fore-aft position of the camera on the panoramic head, and the rotational test is repeated.

The objective is to achieve a state where, upon panning the camera to the left and right, the two reference objects remain in perfect alignment, with no discernible relative displacement, as shown in Fig. 12. When this condition is met, the camera's axis of rotation is correctly aligned with the no-parallax point of the lens at the selected focal length. This position ensures that images captured from different angles will be free of parallax-induced stitching errors.

5.3. Details of Dynamic Scene Elements Annotation

A significant challenge in capturing longitudinal, "in-the-wild" datasets is the presence of dynamic scene elements. While our capture rig ensures a static viewpoint, the long temporal intervals between acquisitions mean that elements such as wind blown foliage, grass, and cloud formations inevitably move.

Although computational methods exist for image alignment (e.g., optical flow warping), applying such modifications would alter the pixel values and compromise the photometric integrity of our ground truth (GT) images. To preserve the dataset as a true GT reference, we opted instead to provide binary masks that identify these non-static regions. This approach allows researchers to optionally exclude these dynamic areas during metric computation, thereby isolating relighting performance from inconsistencies caused by scene motion.

We initially explored automated segmentation methods, including optical flow and contour detection, to identify these moving regions. However, these approaches produced unsatisfactory results, struggling with the subtle motions and complex natural textures present in our scenes. Consequently, we adopted a meticulous manual annotation process.

Figure 12. Nodal point alignment. When the camera is not positioned at the nodal point, rotating the camera causes the nearby utility pole to fail to occlude the poles behind it. In contrast, when the camera is at the nodal point, the nearby utility pole can occlude the poles behind it.

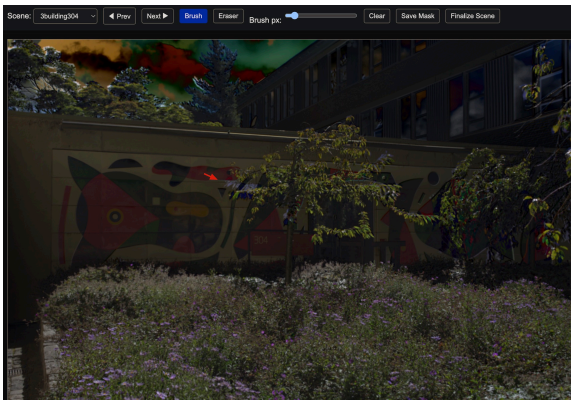

Figure 13. UI Interface for Marking Dynamic Scene Elements. Volunteers utilize a brush tool to create masks or an eraser tool to remove incorrectly marked regions. In the illustrated example, the areas indicated by red arrows correspond to discrepancies between the two photographs, highlighting regions where inconsistencies exist.

Our annotation pipeline is as follows:

1. **Pairwise Comparison:** For each scene, annotators performed a sequential, pairwise comparison of adjacent time steps (e.g., t_0 vs. t_1, t_1 vs. t_2, etc.).
2. **Difference Visualization:** To aid the human annotators, we generated absolute pixel-difference images for each pair. This visualization technique effectively accentuates the contours of misaligned objects, where pixel gradients are highest, making the boundaries of dynamic elements more conspicuous.
3. **Manual Annotation:** Annotators manually painted masks over all identified dynamic regions for each image pair. The primary targets for masking were clouds and moving vegetation (leaves, branches, and grass).
4. **Mask Aggregation:** The final mask for the entire scene is generated by computing the union of all pairwise masks. This ensures that any element that moved at any point during the capture sequence is included in the aggregate mask.

We explicitly excluded two categories of dynamic effects from masking. First, water surfaces (e.g., lakes and seas) were not annotated due to their highly complex and stochastic textures, which are intractable to mask reliably. Second, dynamic reflections and refractions were intentionally left unmasked, as we consider the ability to model these complex, illumination dependent light transport phenomena to be a core aspect of the relighting challenge itself. To facilitate the annotation of standard masks, we employ a dedicated user interface (UI) that enables manual annotation tasks (Fig. 13). The interface provides interactive tools for adjusting the brush size to generate masks, as well as an eraser function to correct erroneous regions.