PAM-ToD: Plug-and-Play Appearance Modeling for Cross-Time-of-Day 3D Gaussian Splatting

Anonymous authors
Paper under double-blind review

Abstract

Adapting a pre-trained 3D Gaussian Splatting (3DGS) road scene to a new time of day requires learning appearance changes from a few anchor images while preserving consistent, real-time rendering. We propose PAM-ToD, a lightweight plug-in that learns color corrections while keeping the pre-trained 3DGS parameters fixed. PAM-ToD scales each Gaussian's existing color to model illumination changes and uses an additive term for additional brightness, such as when street lamps turn on at night. Under a simplified image formation model, unchanged surface albedo can be eliminated from the relation between source and target appearances, allowing us to learn these corrections without separately estimating albedo and illumination. The model corrects colors across the scene while allowing the corrections to vary by location and by Gaussian. To guide learning from a few anchor images, it discourages abrupt spatial changes in these corrections. We also introduce CARLA-ToD, a benchmark with matching geometry, camera poses, and moving-object trajectories across three times of day. A few target-time anchor images are used to train each plug-in, while separate views are used for evaluation. Across the static and dynamic settings, PAM-ToD achieves higher PSNR and lower LPIPS than the baselines, even when the anchor images come from a single synchronized capture across multiple cameras. Code and Dataset:`https://anonymous.4open.science/r/PAM-ToD-ICLR2027-08C3/README.md`

1 Introduction

Autonomous driving systems must operate reliably under changing lighting conditions, from dawn to night. Evaluating them requires a simulator that can render the same road at different times of day. 3D Gaussian Splatting (3DGS) (Kerbl et al., 2023) turns recorded drives into photorealistic scenes that can be rendered in real time (Yan et al., 2024; Zhou et al., 2024; Chen et al., 2025b), supporting driving simulators and digital twins. However, the reconstructed appearance remains tied to the lighting at capture time. Reconstructing the road separately for each time of day repeats the data collection and training effort, increasing the cost of building a simulator.

A practical time-of-day simulator requires a relighting method that (i) reuses an already reconstructed 3DGS, (ii) relies only on sparse images of the new time of day, and (iii) preserves 3D consistency and real-time rendering. Existing relighting approaches address these requirements with different trade-offs. Inverse rendering approaches make 3DGS relightable by decomposing scene appearance into geometry, material, and illumination (Gao et al., 2024; Liang et al., 2024; Herau et al., 2026). Recent methods further improve light transport modeling by incorporating global illumination (Chen et al., 2025a) and spatially varying local lighting (Du et al., 2025). However, separating material and illumination remains ambiguous when the observations provide limited variation in lighting or viewpoint, especially in outdoor scenes. Moreover, these methods typically construct a dedicated relightable representation rather than directly adapting an existing 3DGS to a new time of day.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Existing outdoor appearance models capture lighting variations through learned appearance embeddings or explicit illumination components (Kulhanek et al., 2024; Zhang et al., 2024; Kaleta et al., 2025; Bai et al., 2025; Feng et al., 2025; Xiao et al., 2026), but they are generally designed for image collections spanning multiple conditions. Diffusion-based relighting offers flexible image and video editing (Xing et al., 2025; Liang et al., 2025; He et al., 2025); however, operating in image space without a shared 3D representation can compromise multi-view consistency, and iterative sampling is substantially slower than 3DGS rendering. These limitations motivate adapting a pre-trained 3DGS using only a few target-time images while preserving consistent, real-time rendering.

We address this need through *Cross-Time Scene Transfer*, where a pre-trained 3DGS remains fixed and its appearance is adapted using a few posed target-time anchor images. Our key idea is to learn appearance corrections directly, without separately recovering materials and illumination. Under a simplified diffuse image-formation model, eliminating the time-invariant albedo yields a multiplicative illumination ratio and an additive correction that accounts for emission changes. This relation provides the physical basis for adapting the source appearance directly, without explicitly recovering the underlying material and illumination.

Building on this formulation, we propose PAM-ToD, a lightweight appearance-adaptation plug-in for pre-trained 3DGS. PAM-ToD models illumination variations across three levels of granularity, comprising a global scene-wide gain, a smooth spatial field, and a per-Gaussian residual. All three levels are optimized jointly; capacity-proportional regularization encourages broad appearance changes to be explained by the global gain and the spatial field, while allowing local variation through the per-Gaussian residual. Furthermore, an additive emission term captures newly active light sources, such as street lamps, which multiplicative scaling cannot generate in dark regions. As a result, PAM-ToD achieves robust optimization even under sparse anchor image supervision.

Our main contributions are as follows:

- **Cross-Time Scene Transfer.** We formulate appearance adaptation of a frozen, pre-trained road-scene 3DGS from a few posed target-time images.
- **PAM-ToD.** We introduce a plug-in module that combines hierarchical multiplicative corrections with additive brightness and regularization for sparse-anchor adaptation, enabling multiple time-of-day appearances without modifying the base 3DGS architecture.
- **CARLA-ToD and Evaluation.** We construct a controlled benchmark with matched geometry, camera poses, and dynamic object trajectories across three times of day, demonstrating improved appearance transfer while preserving real-time rendering speed.

2 Related Work

Inverse Rendering. Inverse rendering approaches relighting by separating the contribution of a surface from that of the light falling on it. Inverse rendering methods make this separation explicit by equipping Gaussian primitives with material and illumination attributes, where material parameters describe surface properties such as intrinsic color and reflectance. Early methods such as Relightable 3D Gaussians (Gao et al., 2024), GaussianShader (Jiang et al., 2024), and GS-IR (Liang et al., 2024) recover surface normals, material, and incident illumination under a physically based rendering model. GI-GS (Chen et al., 2025a) models indirect illumination, while GS-ID (Du et al., 2025) combines spatially varying local lighting with diffusion-based material priors. As noted in the introduction, however, the separation remains ambiguous under the single illumination of a typical capture, and each method must reconstruct the scene in its own relightable representation.

Relighting for Outdoor Scenes. Outdoor captures of the same place differ widely with the time of day, the weather, and the shadows they produce. GS-W (Zhang et al., 2024) models appearance variation through intrinsic and dynamic features, while WildGaussians (Kul-

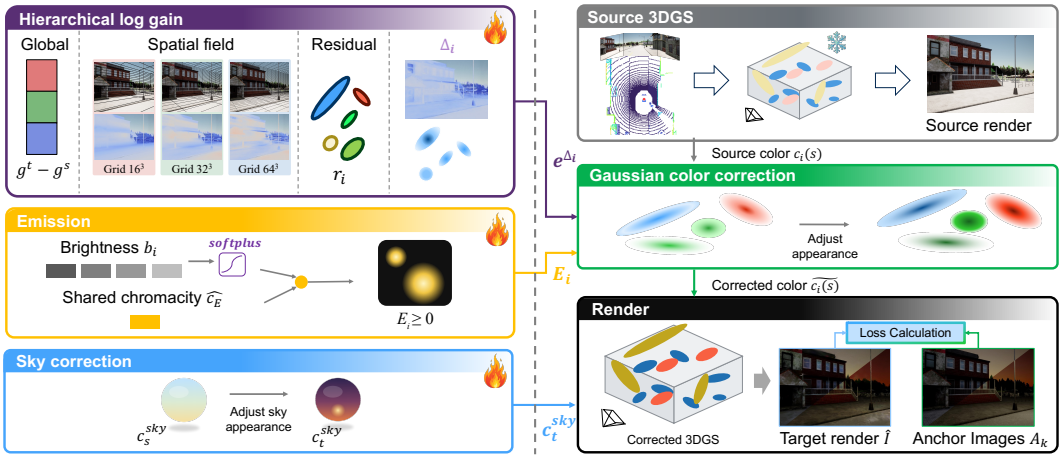


Figure 1: **Overview of PAM-ToD**.Given a frozen 3DGS scene from source time s and sparse anchor images at target time t, PAM-ToD adapts scene appearance in real time using a multiplicative illumination term and an additive emission term.

hanek et al., 2024) combines per-image and per-Gaussian embeddings to predict color transformations. DAVIGS (Lin et al., 2025b) further combines global appearance embeddings with 3D-consistent local features to model spatially varying color transformations. These methods learn appearance variation from image collections rather than explicitly recovering illumination, making changes in lighting conditions only indirectly controllable through the learned appearance representation. A different line of work extends inverse rendering to outdoor scenes. GaRe (Bai et al., 2025), OSDR-GS (Feng et al., 2025), SU-RGS (Zhang et al., 2025b), and LumiGauss (Kaleta et al., 2025) explicitly model sunlight, skylight, and the shadows they cast. For driving scenes, FEGR (Wang et al., 2023), UrbanIR (Lin et al., 2025a), and LightSim (Pun et al., 2023) recover geometry, material, and sun and sky illumination from recorded drives, enabling relighting such as moving the sun or simulating night. A multi-traversal method (Xiao et al., 2026) instead separates material from illumination by recording the same route several times. These approaches address appearance variation through learned appearance representations or explicit material and illumination models. Our method instead derives corrections between source and target appearances and learns them as a plug-in on a frozen 3DGS, without separately estimating albedo and illumination.

Diffusion-based Relighting. A separate line of work changes illumination in images or videos with diffusion models. For single images, IntrinsicDiffusion (Luo et al., 2024) predicts intrinsic layers such as surface color and shading, IC-Light (Zhang et al., 2025a) relights an image according to a text prompt or a background image, and DiffusionLight (Phongthawee et al., 2024) estimates the lighting of a scene from one photograph. Later work extends relighting to video, as in UniRelight (He et al., 2025), UniLumos (Liu et al., 2025), and DiffusionRenderer (Liang et al., 2025), and to multiple viewpoints, as in LightSwitch (Litman et al., 2025), while LumiNet (Xing et al., 2025) transfers the lighting of a reference image to an indoor scene. Without an underlying 3D scene, nothing ensures that the same road looks the same from two viewpoints, and generating each frame takes many sampling steps, far more time than rendering a 3DGS.

3 Method

As illustrated in Figure 1, We propose *PAM-ToD*, a render-time plug-in that adapts a pretrained 3DGS scene to a different time of day while keeping all of its weights fixed. Given a scene reconstructed at source time s and posed anchor images captured at target time t, the plug-in adapts each Gaussian's color through a multiplicative illumination term and an additive emission term. We optimize the plug-in parameters using the anchor images, with implementation details provided in section A.

3.1 Problem Formulation

Preliminaries. The source scene contains M Gaussians $G = \{(\mu_i, q_i, \sigma_i, \alpha_i, c_i(s))\}_{i=1}^{M}$, where the attributes denote centers, rotations, scales, opacities, and source colors, respectively. It also includes a sky model $c_s^{\text{sky}}(\omega)$, where ω is the viewing direction. We freeze all Gaussian parameters and the sky model, and apply learned color corrections to their outputs at render time. Each Gaussian carries a fixed class-probability vector f_i, averaged from SegFormer (Xie et al., 2021) predictions lifted from source images, and class label $\kappa(i) = \arg\max_c f_{i,c}$ (section A.1). Our goal is to reconstruct the target-time appearance from arbitrary viewpoints using the source scene G and a few anchor captures $\{A_k\}_{k=1}^{K}$, each containing synchronized images from the camera rig.

Appearance transport formulations. Motivated by diffuse reflection with additive emission, we adopt a simplified image formation model (Kajiya, 1986; Tappen et al., 2005), $c_i(t) = \rho_i \odot L_i(t) + E_i(t)$, where ρ_i is a time-invariant albedo, $L_i(t)$ is an effective illumination factor, and $E_i(t)$ is an emitted radiance, all of which are three-channel and non-negative. $\odot$ denotes channel-wise multiplication, and the ratios below are likewise taken channel-wise. Assuming that the albedo is unchanged and the source illumination is non-zero in every channel, eliminating ρ_i between the source and target equations yields

$$c_i(t) = \underbrace{c_i(s) \odot \tfrac{L_i(t)}{L_i(s)}}_{\text{illumination term}} + \underbrace{E_i(t) - E_i(s) \odot \tfrac{L_i(t)}{L_i(s)}}_{\text{emission term}}. \tag{1}$$

Eq. 1 expresses the target color as a rescaled source color plus an additive correction, without recovering albedo and illumination separately. Note that the additive correction can be positive or negative even though both emissions are non-negative.

3.2 Parameterization of the Transport Equation

Guided by Eq. 1, PAM-ToD combines a multiplicative illumination term with log gain $\Delta_i \in \mathbb{R}^3$ and a nonnegative additive emission term E_i, together with a separate sky correction. Here, $c_i(s)$ is the view-dependent RGB of the source Gaussian obtained after SH evaluation, i.e., the color that would be rasterized without the plug-in (section A.2). At Gaussian center $x_i = \mu_i$ in world coordinates, the corrected color passed to the rasterizer is

$$\tilde{c}_i = \text{clamp}_{[0,1]}\Big(\underbrace{c_i(s) \odot e^{\Delta_i}}_{\text{illumination term}} + \underbrace{E_i}_{\text{emission term}} \Big). \tag{2}$$

Although the additive correction in Eq. 1 can be negative, we constrain E_i to be non-negative by design: the gain e^{Δ_i} accounts for all darkening, including vanishing source emission, and E_i accounts for additional brightness, including light sources that become active at the target time. This rules out cancellation between an increased multiplicative contribution and a negative additive correction (section A.3).The log gain Δ_i is parameterized hierarchically, with degrees of freedom that expand from a global gain to a spatial field to a per-Gaussian residual. All components are optimized jointly; regularization strengthened in step with this capacity encourages broad appearance changes to be explained by the low-capacity global gain and spatial field, while allowing local variation through the residual.

Illumination term. We model the log illumination field at each time as $\hat{L}(x) = g + \sum_{l=1}^{3} \text{trilerp}(G_l, x)$, where $g \in \mathbb{R}^3$ is the global gain, $G_l \in \mathbb{R}^{3 \times n_l \times n_l \times n_l}$ are the multi-scale log-illumination grids that form the spatial field, $\text{trilerp}(G_l, x)$ denotes trilinear interpolation of G_l at position x, and l indexes three resolution levels with $n_l \in \{16, 32, 64\}$. The log gain is the difference between the fields at times t and s plus a per-Gaussian residual r_i:

$$\Delta_i = \underbrace{(g^t - g^s)}_{\text{global gain}} + \underbrace{\textstyle\sum_{l=1}^{3} \big[\text{trilerp}(G_l^t, x_i) - \text{trilerp}(G_l^s, x_i)\big]}_{\text{spatial field}} + \underbrace{r_i}_{\text{residual}}. \tag{3}$$

The global gain captures scene-wide brightness changes, while the spatial field models smooth, spatially varying appearance changes such as sky gradients and large-scale shading

variations across surfaces, with each grid regularized by a total-variation penalty that grows with its resolution. The residual $r_i \in \mathbb{R}^3$ is a directly optimized per-Gaussian parameter, initialized to zero, that absorbs local changes the spatial field cannot capture; it is smoothed on a semantic k-NN graph whose edge weights $w_{ij} = \max(0, \cos(f_i, f_j))$ suppress smoothing across semantic boundaries (section A.6).

Emission term. The emission E_i is a non-negative per-Gaussian vector that represents artificial light sources and local pools of light in dark regions. To constrain emission color variation, we factor it into a learnable per-Gaussian brightness $b_i \in \mathbb{R}$ and a learnable chromaticity $\hat{c}_E \in \mathbb{R}^3_+$ of unit mean shared across the scene, $E_i = \mathrm{softplus}(b_i)\,\hat{c}_E$, initialized so that the emission is negligible and the plug-in reproduces the source render (section A.4).

Sky model. The frozen sky model is transported to the target time by a global affine transform, $c_t^{\mathrm{sky}}(\omega) = \mathrm{clamp}\big(e^{g_{\mathrm{sky}}} \odot c_s^{\mathrm{sky}}(\omega) + b_{\mathrm{sky}}\big)$ with the learnable $g_{\mathrm{sky}}, b_{\mathrm{sky}} \in \mathbb{R}^3$ initialized to zero; the bias compensates for an absolute deficit of light in the source, as arises when transferring from night to day.

3.3 Optimization

All plug-in parameters are optimized end-to-end by minimizing $\mathcal{L} = \lambda_{\mathrm{anc}}\mathcal{L}_{\mathrm{anchor}} + \lambda_{\mathrm{atp}}\mathcal{L}_{\mathrm{atp}} + \lambda_{\mathrm{unseen}}\mathcal{L}_{\mathrm{unseen}} + \mathcal{L}_{\mathrm{chroma}} + \mathcal{L}_{\mathrm{reg}}$, which combines a robust anchor reconstruction loss, an appearance transport prior that conditions the correction on the source color, a seen-to-unseen prior that propagates the adaptation from observed to unobserved regions, a chromaticity regularizer that suppresses hue noise, and structural regularizers matched to the representation hierarchy. Every set-averaged term is defined as zero when its set is empty; full definitions are given in section A.5.

Robust anchor reconstruction loss. The anchors $\{A_k\}$ are the only direct supervision, yet in practice they may be pseudo-images from a generative model, which may contain fine details that are inconsistent across views. We therefore make the standard L1+SSIM loss robust to such anchors: the rendering $\hat{I}$ and the anchor A are smoothed with a Gaussian filter g_σ to reduce sensitivity to fine-scale differences while retaining broader appearance structure, and the set $\mathbb{T}_\rho$ of the fraction ρ of pixels with the largest channel-averaged absolute errors after smoothing, is excluded from the L1 term:

$$\mathcal{L}_{\mathrm{anchor}} = \operatorname*{mean}_{p \notin \mathbb{T}_\rho} \big|(g_\sigma * \hat{I})_p - (g_\sigma * A)_p\big| + \lambda_{\mathrm{ssim}}\big(1 - \mathrm{SSIM}(g_\sigma * \hat{I}, g_\sigma * A)\big). \tag{4}$$

The sky correction receives gradients from this loss through sky pixels.

Appearance transport prior. Anchor pixels constrain only the Gaussians that project into them, whereas the time-of-day change of a surface is largely predictable from its semantic class and its source brightness: within a class, surfaces that are bright or dark at the source time map to consistently different target colors. For this prior, we use view-independent source RGB values c_i^{DC} computed from the SH DC component. For each class, the source colors and the anchor pixels are separately split into Q luminance quantile bins, and each Gaussian receives as pseudo target τ_i the mean anchor RGB of the bin whose rank matches that of its source color. The loss is an L1 pull toward these targets of the gain-corrected source color $\bar{c}_i$, which omits the emission so that the prior constrains the gain alone (section A.5): $\mathcal{L}_{\mathrm{atp}} = \frac{1}{3|\mathbb{A}|}\sum_{i\in\mathbb{A}} \|\bar{c}_i - \tau_i\|_1$, over the set $\mathbb{A}$ of Gaussians with a valid τ_i, whether visible from the anchors or not.

Seen-to-unseen transport prior. Gaussians outside every anchor view receive no gradient from $\mathcal{L}_{\mathrm{anchor}}$, so their log gain is left to whatever the smooth field extrapolates, which surfaces as inconsistent appearance in novel views; this prior transfers the fitted illumination change to them. Let $\mathbb{V}$ be the set of Gaussians visible from at least one anchor view and $\mathbb{U}$ its complement. Assuming that Gaussians of the same semantic class κ and the same distance bin d to the nearest light source (Gaussians labeled pole or traffic light serve as proxies) undergo a similar illumination change, we compute the per-channel median $m_{\kappa,d}$ of

Δ_i over the visible members of each (κ, d) cell, and during the second half of training pull the unseen set toward it, $\mathcal{L}_{\text{unseen}} = \frac{1}{|\mathbb{U}_{\text{valid}}|} \sum_{i \in \mathbb{U}_{\text{valid}}} \|\Delta_i - \text{sg}(m_i)\|^2$ with $m_i \equiv m_{\kappa(i),d(i)}$, where $\mathbb{U}_{\text{valid}} \subseteq \mathbb{U}$ is the subset of unseen Gaussians for which a reference median m_i is available (section A.5), and sg denotes the stop-gradient operator; detaching the target makes the propagation one-directional and preserves the fit in the visible region.

Chromaticity regularization. A change of time of day alters brightness far more than hue, whereas the per-Gaussian residual and the propagation to unseen Gaussians are free enough to introduce hue noise, visible as colored blotches on uniform surfaces such as roads. We therefore penalize hue changes in the log gain while leaving luminance free. With the chromaticity operator $\text{ch}(v) = v - \frac{1}{3}\sum_c v_c$, the regularizer consists of three parts:

$$\mathcal{L}_{\text{chroma}} = \frac{\lambda_{\text{rch}}}{3M} \sum_{i=1}^{M} \|\text{ch}(r_i)\|^2 + \frac{\lambda_{\text{ch}}}{|\mathbb{U}_{\text{valid}}|} \sum_{i \in \mathbb{U}_{\text{valid}}} \|\text{ch}(\Delta_i - \text{sg}(m_i))\|^2 + \frac{\lambda_{\text{chu}}}{|\mathbb{C}|} \sum_{i \in \mathbb{C}} \|\text{ch}(\Delta_i) - \text{sg}(\tilde{h}_{\kappa(i)})\|^2. \tag{5}$$

The first part restricts the residual r_i to luminance changes; the second suppresses spurious hue shifts when propagating to unseen Gaussians; and the third drives the chromaticity of uniformly textured classes $\mathbb{C}$ (road and sidewalk) toward $\tilde{h}_\kappa$, the detached per-class median of $\text{ch}(\Delta_i)$ over $\mathbb{C}$ (section A.5).

Structural regularization. Finally, because a few anchors constrain far fewer pixels than the plug-in has parameters, each level of the representation hierarchy receives one smoothness or sparsity regularizer matched to its capacity, $\mathcal{L}_{\text{reg}} = \lambda_{\text{tv}}\mathcal{R}_{\text{tv}} + \lambda_{\text{knn}}\mathcal{R}_{\text{knn}} + \lambda_E\mathcal{R}_E$, where $\mathcal{R}_{\text{tv}}$ is the resolution-weighted total-variation penalty on the spatial field grids, $\mathcal{R}_{\text{knn}}$ the semantic k-NN smoothness of the residual, and $\mathcal{R}_E$ the ℓ_1 sparsity penalty on the emission (section A.6). Optimizer settings and all hyperparameters are listed in section A.7.

4 CARLA-ToD Dataset

Evaluating time-of-day adaptation pixel by pixel requires image pairs in which geometry, camera poses, and dynamic objects are held fixed and only the illumination changes. Because real driving data cannot satisfy this condition, we construct the CARLA-ToD dataset with the CARLA simulator. The dataset comprises 24 sequences (200 frames $\times$ 3 cameras each), covering four driving scenes at three times of day (Morning, Noon, Night) in two configurations (Static, Dynamic).

Generation protocol. An ego vehicle is driven along a predefined route in synchronous mode with the same route and timing at every time of day, so the scene geometry and camera poses coincide exactly between corresponding frames (the difference in extrinsics is identically zero) and only illumination and emission change. The weather is fixed to clear, and only the sun altitude is varied. Morning ($2.0°$) reproduces the warm appearance of a low sun, Noon ($90.0°$) a neutral illumination under a high sun, and Night ($-1.0°$) an environment in which emission from artificial light sources, such as street lights and building windows, coexists with dark regions. The Static configuration places no dynamic actors and is used to evaluate baselines that assume static scenes. The Dynamic configuration populates the scene with 40–60 vehicles and 40–55 pedestrians that follow identical trajectories at every time of day.

Sensors and annotations. The ego vehicle carries a rigid rig of three RGB cameras (Front, Front-left, Front-right), each recording 1920×1300 sRGB images with a horizontal field of view of $60°$ and no lens distortion, and a roof-mounted 64-channel LiDAR that records about 1.1×10^5 points per frame. The mounting geometry is given in Appendix B. We provide a binary sky mask for every RGB frame and, for the Dynamic configuration, binary masks of dynamic objects (all, vehicles, pedestrians) and 3D bounding boxes.

Table 1: Quantitative relighting comparison on the CARLA-ToD Static Dataset. The first, second, and third best performances are highlighted in First, Second, and Third, respectively.

Methods	Reconstruction			Novel View Synthesis			FPS ↑
	SSIM ↑	PSNR ↑	LPIPS ↓	SSIM ↑	PSNR ↑	LPIPS ↓	
GS-W	0.748	16.76	0.312	0.729	16.58	0.324	22.9
GS-IR	0.663	16.42	0.372	0.671	16.57	0.361	61.8
GI-GS	0.663	16.37	0.353	0.667	16.60	0.342	16.3
LumiGauss	0.491	11.29	0.597	0.467	10.89	0.622	152.7
StreetGS	0.471	10.17	0.487	0.463	10.13	0.486	66.7
+Ours ($K = 1$)	0.740	22.05	0.298	0.729	21.89	0.297	53.4
+Ours ($K = 8$)	0.826	25.66	0.249	0.814	25.45	0.247	53.4

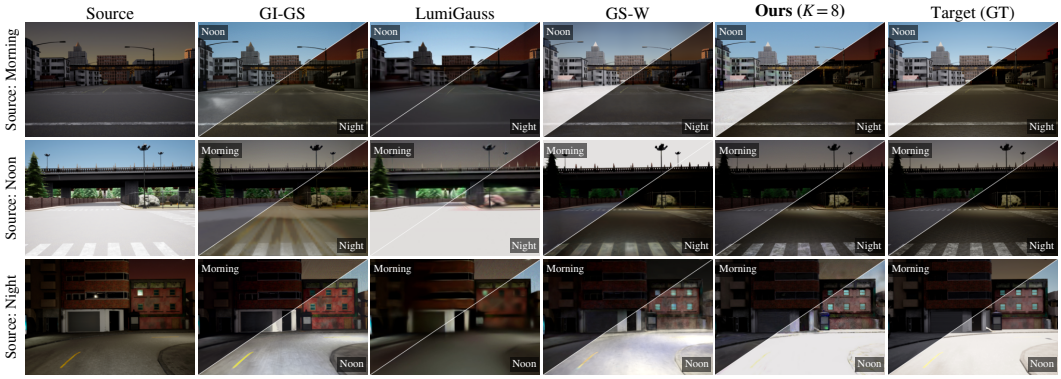


Figure 2: Qualitative comparison on the CARLA-ToD Static Dataset.

5 Experiments

To validate the proposed method, we evaluate on the newly constructed CARLA-ToD dataset and on the Waymo Open Dataset (WOD) (Sun et al., 2020). Comparisons with inverse rendering and appearance-code baselines use the CARLA-ToD Static. The proposed method uses StreetGS as its base model and is evaluated with $K = 1$ and $K = 8$ synchronized anchor captures. Comparisons with diffusion-model baselines use the CARLA-ToD Dynamic setting, where each diffusion model is applied to the renderings of StreetGS.

Implementation details. For each dataset, we trained all methods on a single NVIDIA A6000 GPU and compared their performance on 3D reconstruction and novel view synthesis (protocol details in section B.3). We further evaluated the contribution of each major component of our method by individually removing it and quantitatively measuring its impact. Since GS-W requires images captured at different times of day in its training set, we trained it using a combination of eight frames from different times of day. For GS-IR, GI-GS, and LumiGauss, we trained a separate model for each time of day and evaluated relighting by swapping the resulting environment maps across times of day.

Comparison on CARLA-ToD. Tables 1 and 2 report quantitative results on CARLA-ToD Static and Dynamic datasets, averaged over all six transfer directions (per-direction details in Appendix). In Table 1, our method achieves the best ToD adaptation quality while retaining real-time rendering. On NVS ($K = 8$), it achieves the highest accuracy, improving over baselines by up to 0.085 SSIM, 8.87 dB PSNR, and 0.077 LPIPS with negligible computational overhead, maintaining real-time rendering at 53.4 FPS. Even with $K = 1$, our physical inductive bias and hierarchical regularization suppress overfitting and significantly surpass baselines. As shown in Figure 2, inverse rendering methods struggle with ill-posed material decomposition, whereas appearance-embedding methods lack expressiveness for large illumination changes. In Table 2, the proposed method ($K = 1, 8$) likewise outperforms state-of-the-art diffusion-model methods. Diffusion-model methods

Table 2: Quantitative relighting comparison on the CARLA-ToD Dynamic Dataset.

Methods	Reconstruction SSIM ↑	PSNR ↑	LPIPS ↓	Novel View Synthesis SSIM ↑	PSNR ↑	LPIPS ↓	FPS ↑
LumiNet	0.569	14.44	0.422	0.560	14.45	0.421	0.08
DiffusionRenderer	0.541	14.49	0.496	0.529	14.49	0.566	0.08
UniRelight	0.597	15.89	0.484	0.560	15.49	0.577	0.07
StreetGS	0.467	10.54	0.486	0.462	10.51	0.481	59.3
+Ours ($K = 1$)	0.715	21.29	0.318	0.706	21.15	0.311	49.4
+Ours ($K = 8$)	0.803	24.79	0.269	0.791	24.45	0.263	49.4

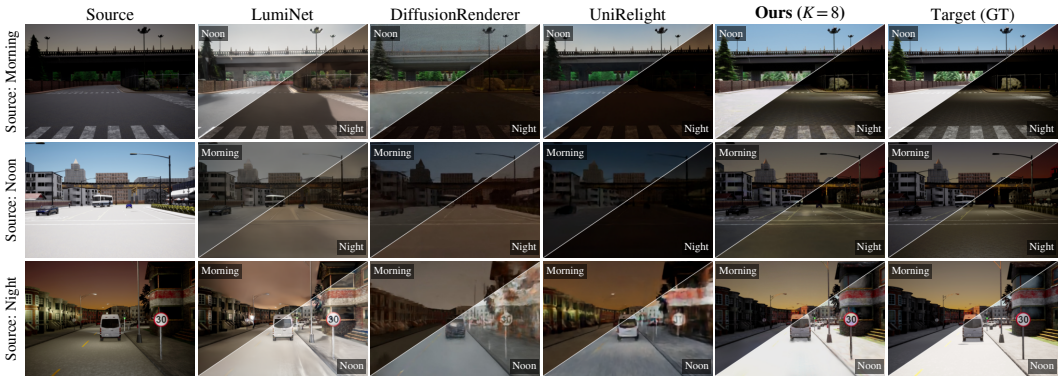


Figure 3: Qualitative comparison on the CARLA-ToD Dynamic Dataset.

such as LumiNet and UniRelight operate in 2D image or video space and therefore cannot preserve 3D geometric consistency or photometric coherence across viewpoints. Their multi-step denoising also makes inference extremely slow, at 0.07–0.08 FPS. Figure 3 presents the qualitative comparisons. LumiNet loses spatial consistency on continuous structures such as roads. Meanwhile, Diffusion Renderer and UniRelight struggle to produce reference-faithful relighting and accurately capture fine details like road markings and moving vehicles.

Ablation on the number of anchors. Table 3 varies the number of synchronized anchor captures K. All metrics improve monotonically with K, with diminishing gains that saturate around $K = 8$. Even a single synchronized capture is within about 4 dB of $K = 16$, indicating that the low-degree-of-freedom global gain and spatial field explain most of the time-of-day change. The remaining gap stems from regions unobserved by the anchors or saturated near zero in the dark source, where the emission term is constrained only by anchor coverage.

Ablation on the loss terms. Table 4 evaluates the contribution of individual loss terms under fixed structural regularization. Removing the anchor reconstruction loss causes a noticeable drop in accuracy; nevertheless, performance remains above the base model as the appearance transport prior captures global shifts from per-class luminance statistics. Omitting the appearance transport prior slightly degrades all metrics, validating its complementary effect. The seen-to-unseen prior shows a minor effect here, given that eight synchronized anchor captures yield dense coverage and leave only a small unobserved set $\mathbb{U}$. Finally, removing chromaticity regularization maintains PSNR but degrades SSIM and LPIPS, as it explicitly suppresses subtle hue blotches on flat surfaces that impact perceptual quality more than raw pixel errors.

Ablation on the correction channels and hierarchy level. Table 5 shows that both correction channels are essential. Removing the multiplicative channel causes the largest drop at both anchor counts, whereas omitting the additive term yields smaller degradation. The hierarchy levels also exhibit anchor-dependent behaviors. At K=8, global gain captures most illumination changes, with spatial grids providing further gains and residuals offering marginal value. At K=1, however, global gain alone is insufficient, making each hierarchy level necessary. Although the residual-only variant approaches the full model in PSNR at

Table 3: Ablation on the number of K.

K	SSIM ↑	PSNR ↑	LPIPS ↓
1	0.720	21.11	0.327
2	0.741	21.94	0.323
4	0.772	23.48	0.314
8	0.808	24.72	0.285
16	**0.825**	**25.23**	**0.273**

Table 4: Ablation on the loss terms.

Loss	SSIM ↑	PSNR ↑	LPIPS ↓
w/o $\mathcal{L}_{\text{anchor}}$	0.622	15.05	0.429
w/o $\mathcal{L}_{\text{atp}}$	0.803	24.54	0.289
w/o $\mathcal{L}_{\text{unseen}}$	0.808	24.71	0.286
w/o $\mathcal{L}_{\text{chroma}}$	0.804	24.75	0.291
Full	**0.808**	**24.72**	**0.285**

Table 5: Ablation on the correction channels and hierarchy levels with K=1, 8.

Method	$K=1$ SSIM ↑	$K=1$ PSNR ↑	$K=1$ LPIPS ↓	$K=8$ SSIM ↑	$K=8$ PSNR ↑	$K=8$ LPIPS ↓
Base	0.471	10.17	0.487	0.471	10.17	0.487
w/o Illumination term	0.565	12.83	0.431	0.659	16.73	0.375
w/o Emission term	0.717	20.86	0.329	0.783	22.69	0.311
Global Only	0.597	16.91	0.439	0.786	23.86	0.303
Residual Only	0.714	20.60	0.399	0.790	23.58	0.289
Global + Sparse Grid	0.689	20.01	0.358	0.804	24.70	0.289
Full	**0.720**	**21.11**	**0.327**	**0.808**	**24.72**	**0.285**

K=1, its worse LPIPS and lower score at K=8 confirm that multi-scale grids are critical for cross-view consistency. Overall, the full model consistently achieves the best performance, validating our multi-scale hierarchy under sparse anchor supervision, where global correction alone is insufficient.



Figure 4: Qualitative comparison on the Waymo Open Dataset.

Real-Scene Adaptation on Waymo with Synthesized Anchors. We additionally evaluate whether PAM-ToD can adapt a 3DGS trained from WOD. For this experiment, we use 4scenes as the source domain and synthesize sparse target-time anchor images using the OpenAI Image Edit API (ope, Accessed:2026-09-19). Across all reference-conditioned criteria, our method outperforms UniRelight, achieving lower errors in ΔE_{00} (2.69 vs. 8.10), ΔL^* (1.78 vs. 5.26) (Sharma et al., 2005), and EMD_L (1.10 vs. 4.01) (Rubner et al., 2000). UniRelight suffers from degraded reference fidelity and chunk-wise brightness inconsistencies, whereas our approach faithfully retains target illumination. However, our method remains susceptible to floaters inherited from the source 3DGS, which diffusion-based models naturally eliminate as image enhancers.

6 Conclusion

We formalize Cross-Time Scene Transfer, adapting a frozen pre-trained road-scene 3DGS from sparse target-time images, and propose PAM-ToD as a lightweight plug-in. Motivated by the cancellation of time-invariant albedo, PAM-ToD models appearance changes using a multiplicative correction and a non-negative additive emission term, avoiding explicit material-illumination decomposition. Additionally, its hierarchical parameterization and near-identity initialization ensure stable adaptation even from a single synchronized anchor. On CARLA-ToD, it substantially outperforms inverse rendering, appearance embedding, and diffusion baselines while maintaining real-time rendering. **Limitations and Future Work.** Our method requires posed anchor images at the target time of day. Additionally, it struggles to capture specularities, sharp shadow boundaries, and physically consistent shadow rendering. Future work includes extending our method to integrating physical shadow priors to handle severe illumination extremes.

AI USE STATEMENT

In this study, generative AI tools were used to assist in designing figures, as well as implementing and debugging experimental code. For figure creation, AI assisted in exploring layout choices. Generative AI was also used to check the consistency of mathematical formulations and method descriptions, survey related literature, and polish the manuscript text. Analyses generated by AI were not directly used as scientific evidence but served as supplementary material subject to strict author verification. The authors take full responsibility for the final contents, including any parts prepared with AI assistance.

REPRODUCIBILITY STATEMENT

To enable full reproducibility of the proposed method and correct interpretation of experimental results, detailed descriptions of our methodology and evaluation settings are provided in the main text and appendix. Furthermore, the code and datasets used in our experiments are shared in the submission and will be made publicly available upon acceptance.

REFERENCES

gpt-image-2.5-sunburst, Accessed:2026-09-19. URL `https://developers.openai.com/api/docs/guides/images-vision?api-mode=responses`.

Haiyang Bai, Jiaqi Zhu, Songru Jiang, Wei Huang, Tao Lu, Yuanqi Li, Jie Guo, Runze Fu, Yanwen Guo, and Lijun Chen. Gare: Relightable 3d gaussian splatting for outdoor scenes from unconstrained photo collections. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Hongze Chen, Zehong Lin, and Jun Zhang. GI-GS: Global illumination decomposition on gaussian splatting for inverse rendering. In *International Conference on Learning Representations (ICLR)*, 2025a.

Ziyu Chen, Jiawei Yang, Jiahui Huang, Riccardo de Lutio, Janick Martinez Esturo, Boris Ivanovic, Or Litany, Zan Gojcic, Sanja Fidler, Marco Pavone, Li Song, and Yue Wang. OmniRe: Omni urban scene reconstruction. In *International Conference on Learning Representations (ICLR)*, 2025b.

Kang Du, Zhihao Liang, Yulin Shen, and Zeyu Wang. GS-ID: Illumination decomposition on gaussian splatting via adaptive light aggregation and diffusion-guided material priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Wei Feng, Kangrui Ye, Qi Zhang, Qian Zhang, and Nan Li. 2d gaussian splatting for outdoor scene decomposition and relighting. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 990–998, 2025.

Jian Gao, Chun Gu, Youtian Lin, Zhihao Li, Hao Zhu, Xun Cao, Li Zhang, and Yao Yao. Relightable 3d gaussians: Realistic point cloud relighting with BRDF decomposition and ray tracing. In *European Conference on Computer Vision (ECCV)*, 2024.

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. *arXiv preprint arXiv:2506.15673*, 2025.

Quentin Herau, Tianshuo Xu, Depu Meng, Jiezhi Yang, Chensheng Peng, Spencer Sherk, Yihan Hu, and Wei Zhan. Spectralsplat: Appearance-disentangled feed-forward gaussian splatting for driving scenes, 2026.

Yingwenqi Jiang, Jiadong Tu, Yuan Liu, Xifeng Gao, Xiaoxiao Long, Wenping Wang, and Yuexin Ma. GaussianShader: 3d gaussian splatting with shading functions for reflective surfaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5322–5332, 2024.

James T Kajiya. The rendering equation. In *Proceedings of the 13th annual conference on Computer graphics and interactive techniques*, pp. 143–150, 1986.

Joanna Kaleta, Kacper Kania, Tomasz Trzcinski, and Marek Kowalski. Lumigauss: Relightable gaussian splatting in the wild. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2025.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (TOG)*, 42(4), 2023.

Jonas Kulhanek, Songyou Peng, Zuzana Kukelova, Marc Pollefeys, and Torsten Sattler. Wildgaussians: 3d gaussian splatting in the wild. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Ruofan Liang, Zan Gojcic, Huan Ling, Jacob Munkberg, Jon Hasselgren, Zhi-Hao Lin, Jun Gao, Alexander Keller, Nandita Vijaykumar, Sanja Fidler, and Zian Wang. Diffusionrenderer: Neural inverse and forward rendering with video diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Zhihao Liang, Qi Zhang, Ying Feng, Ying Shan, and Kui Jia. GS-IR: 3d gaussian splatting for inverse rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Chih-Hao Lin, Bohan Liu, Yi-Ting Chen, Kuan-Sheng Chen, David A. Forsyth, Jia-Bin Huang, Anand Bhattad, and Shenlong Wang. UrbanIR: Large-scale urban scene inverse rendering from a single video. In *International Conference on 3D Vision (3DV)*, pp. 512–523, 2025a.

Jiaqi Lin, Zhihao Li, Binxiao Huang, Xiao Tang, Jianzhuang Liu, Shiyong Liu, Xiaofei Wu, Fenglong Song, and Wenming Yang. Decoupling appearance variations with 3d consistent features in gaussian splatting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025b.

Yehonathan Litman, Fernando De la Torre, and Shubham Tulsiani. Lightswitch: Multi-view relighting with material-guided diffusion. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Ropeway Liu, Hangjie Yuan, Bo Dong, Jiazheng Xing, Jinwang Wang, Rui Zhao, Yan Xing, Weihua Chen, and Fan Wang. Unilumos: Fast and unified image and video relighting with physics-plausible feedback. *arXiv preprint arXiv:2511.01678*, 2025.

Jundan Luo, Duygu Ceylan, Jae Shin Yoon, Nanxuan Zhao, Julien Philip, Anna Frühstück, Wenbin Li, Christian Richardt, and Tuanfeng Y. Wang. Intrinsicdiffusion: Joint intrinsic layers from latent diffusion models. In *ACM SIGGRAPH 2024 Conference Papers*, 2024.

Pakkapon Phongthawee, Worameth Chinchuthakun, Nontaphat Sinsunthithet, Amit Raj, Varun Jampani, Pramook Khungurn, and Supasorn Suwajanakorn. Diffusionlight: Light probes for free by painting a chrome ball. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Ava Pun, Gary Sun, Jingkang Wang, Yun Chen, Ze Yang, Sivabalan Manivasagam, Wei-Chiu Ma, and Raquel Urtasun. Neural lighting simulation for urban scenes. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000.

Gaurav Sharma, Wencheng Wu, and Edul N Dalal. The ciede2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations. *Color Research & Application*, 30(1):21–30, 2005.

Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

M.F. Tappen, W.T. Freeman, and E.H. Adelson. Recovering intrinsic images from a single image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(9): 1459–1472, 2005.

Zian Wang, Tianchang Shen, Jun Gao, Shengyu Huang, Jacob Munkberg, Jon Hasselgren, Zan Gojcic, Wenzheng Chen, and Sanja Fidler. Neural fields meet explicit geometric representations for inverse rendering of urban scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8370–8380, 2023.

Yangyi Xiao, Siting Zhu, Baoquan Yang, Tianchen Deng, Yongbo Chen, and Hesheng Wang. Appearance decomposition gaussian splatting for multi-traversal reconstruction, 2026.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Xiaoyan Xing, Konrad Groh, Sezer Karaoglu, Theo Gevers, and Anand Bhattad. Luminet: Latent intrinsics meets diffusion models for indoor scene relighting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Yunzhi Yan, Haotong Lin, Chenxu Zhou, Weijie Wang, Haiyang Sun, Kun Zhan, Xianpeng Lang, Xiaowei Zhou, and Sida Peng. Street gaussians: Modeling dynamic urban scenes with Gaussian splatting. In *European Conference on Computer Vision (ECCV)*, 2024.

Dongbin Zhang, Chuming Wang, Weitao Wang, Peihao Li, Minghan Qin, and Haoqian Wang. Gaussian in the wild: 3d gaussian splatting for unconstrained image collections. In *European Conference on Computer Vision (ECCV)*, 2024.

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 *International Conference on Learning Representations (ICLR)*, 2025a.

Qi Zhang, Chi Huang, Qian Zhang, Nan Li, and Wei Feng. Su-rgs: Relightable 3d gaussian splatting from sparse views under unconstrained illuminations. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 26859–26868, 2025b.

Xiaoyu Zhou, Zhiwei Lin, Xiaojun Shan, Yongtao Wang, Deqing Sun, and Ming-Hsuan Yang. DrivingGaussian: Composite gaussian splatting for surrounding dynamic autonomous driving scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

A Method Details

This appendix collects the implementation details omitted from section 3.

A.1 Semantic Labels and Anchors

We run SegFormer (Xie et al., 2021) with Cityscapes classes on the source images and lift the per-pixel class probabilities to the Gaussians by visibility-weighted averaging over all source views, where the weight of a view combines the Gaussian's opacity with an occlusion test against the rendered depth. The averaged distribution is the semantic feature f_i and its arg-max is the class label $\kappa(i)$; averaging also resolves conflicting observations across views. Both quantities are computed once and kept fixed during plug-in optimization. An anchor A_k is one synchronized capture of the three cameras (front, front-left, front-right) of the same rig as the source scene, so K anchors provide $3K$ posed images.

A.2 Color Space and Point of Application

The correction of Eq. 2 is applied to the view-dependent RGB of each Gaussian obtained after spherical-harmonics evaluation and before rasterization, in the display-encoded color space of the training images; no linearization is performed. Eq. 1 therefore serves as a structural approximation in that space rather than as a radiometric identity, which is also why we do not interpret its two factors as physically recovered illumination and emission. Two color representations of the source Gaussians are thus used: $c_i(s)$, the view-dependent RGB after spherical-harmonics evaluation, which is what rendering corrects, and c_i^{DC}, the view-independent RGB given by the degree-0 SH coefficient alone, which the appearance transport prior of section A.5 uses because its pseudo targets are per-Gaussian statistics computed once without a viewing direction.

A.3 Sign of the Additive Correction

The additive correction in Eq. 1 can be negative, e.g., when a source-time emitter is switched off at the target time, whereas E_i in Eq. 2 is non-negative. Since ρ_i, L_i, and E_i are never observed separately, the split in Eq. 1 is not unique: for any source color $c_i(s) > \mathbf{0}$, a negative additive part can be absorbed into the multiplicative term by

$$e^{\Delta_i} \leftarrow \min\big(e^{\Delta_i},\, c_i(t) \oslash c_i(s)\big), \tag{6}$$

leaving the non-negative remainder $\max\big(\mathbf{0},\, c_i(t) - c_i(s) \odot e^{\Delta_i}\big)$, where $\oslash$ denotes channel-wise division. For an unconstrained per-Gaussian source–target RGB pair, the non-negativity of E_i therefore does not reduce the set of representable target colors; the additive term is strictly necessary only where $c_i(s) \approx \mathbf{0}$, in which case the remainder is non-negative. This argument concerns a single color pair and does not extend to the parameterization of Eq. 2, where the emission chromaticity is shared across Gaussians and the gain is shared across viewing directions. Fixing the sign removes the ambiguity between a large gain and a cancelling negative emission, which we found to destabilize optimization with few anchors. The final clamp in Eq. 2 plays no role in this argument.

A.4 Parameterization and Initialization

Residual. $r_i \in \mathbb{R}^3$ is a directly optimized per-Gaussian parameter, initialized to zero; it is regularized, not aggregated, on the k-NN graph of section A.6.

Emission. $b_i \in \mathbb{R}$ is a learnable per-Gaussian scalar that controls the emission magnitude through softplus; it is initialized to $b_i = -6$, so that $\mathrm{softplus}(b_i) \approx 2.5 \times 10^{-3}$ and the plug-in initially reproduces the source render. The shared chromaticity $\hat{c}_E = 3\,\mathrm{softplus}(\theta_E)/\|\mathrm{softplus}(\theta_E)\|_1$ is a positive vector of unit mean with learnable $\theta_E \in \mathbb{R}^3$, initialized to $\hat{c}_E = \mathbf{1}$.

Sky. $g_{\text{sky}}, b_{\text{sky}} \in \mathbb{R}^3$ are the only trainable sky parameters; both are unconstrained and initialized to zero, so the transform starts as the identity, while the source sky model itself remains frozen.

A.5 Loss Definitions

Robust anchor loss. Let $e_p = \frac{1}{3}\sum_c |(g_\sigma * \hat{I})_{p,c} - (g_\sigma * A)_{p,c}|$ be the channel-mean residual after blurring. $\mathbb{T}_\rho$ contains the fraction ρ of pixels with the largest e_p within the anchor image rendered at the current iteration (section A.7; $\rho = 10\%$, $\sigma = 2$), without any additional masks. Trimming applies only to the L1 term; the SSIM term is evaluated on all pixels of the blurred images.

Appearance transport prior. For each class κ, the view-independent source colors c_i^{DC} of its Gaussians and the anchor pixels labeled κ by 2D segmentation of the anchor images are separately sorted by luminance ($0.299R + 0.587G + 0.114B$) and split into Q=32 equal-count quantile bins. Each Gaussian receives as pseudo target τ_i the mean anchor RGB of the bin whose rank matches that of c_i^{DC}; an empty anchor bin inherits the mean of the nearest lower non-empty bin (the class-wide anchor mean if no such bin exists), and classes with fewer than Q Gaussians or fewer than 2,000 anchor pixels are skipped. τ_i is computed once before optimization and is a rank-matching pseudo target rather than a pixel correspondence. The gain-corrected source color of section 3.3 is $\bar{c}_i = \text{clamp}_{[0,1]}(c_i^{\text{DC}} \odot e^{\Delta_i})$, which omits the emission term, so E_i receives no gradient from this prior; for efficiency, the loss is evaluated on a random subset of $\mathbb{A}$ at each step.

Seen-to-unseen prior. The visibility weight of Gaussian i is the sum over the anchor views of (opacity) $\times$ (in-frame indicator) $\times$ (non-occlusion indicator against the rendered depth); $\mathbb{V}$ is the set of Gaussians with a positive weight and $\mathbb{U}$ its complement. Since true light-source positions are unknown, we use as proxies the Gaussians whose lifted segmentation has its arg-max at *pole* or *traffic light*; $d(i)$ is the Euclidean distance from x_i to the nearest proxy center, quantized with the fixed, logarithmically spaced edges $\{2, 4, 8, 16, 32\}$ m into six bins. The bins depend only on the frozen source model and are computed once; in a scene without any proxy, the class-only median m_κ is used instead. The medians $m_{\kappa,d}$ are first computed over the visible set at the midpoint of optimization and recomputed periodically thereafter; $\mathcal{L}_{\text{unseen}}$ and the second part of $\mathcal{L}_{\text{chroma}}$ are active only from the midpoint on. A cell (κ, d) with too few visible members for a reliable median falls back to m_κ, and a class with too few visible members receives no prior (table 6 lists the thresholds). We denote by $\mathbb{U}_{\text{valid}} \subseteq \mathbb{U}$ the unseen Gaussians whose class or cell has a valid reference median; the sum and the normalization of $\mathcal{L}_{\text{unseen}}$ and of the second part of $\mathcal{L}_{\text{chroma}}$ run over $\mathbb{U}_{\text{valid}}$, so that unseen Gaussians without a reference contribute neither to the loss nor to its normalization.

Chromaticity regularization. In Eq. 5, $\mathbb{C}$ is the set of Gaussians labeled road or sidewalk (a class with fewer than 100 Gaussians is excluded), and $\tilde{h}_\kappa$ is the per-channel median of $\text{ch}(\Delta_i)$ over the members of $\mathbb{C}$ with class κ, recomputed from the current values at every step and detached. All three parts act on log-gain quantities; the first part is averaged over all $3M$ entries of the residual, as $\mathcal{R}_E$ is over the emission. In dynamic scenes, the dynamic-object Gaussians share the global gain and spatial field with the static scene but carry their own residual and emission, to which the first part of $\mathcal{L}_{\text{chroma}}$, $\mathcal{R}_{\text{knn}}$, and $\mathcal{R}_E$ are applied in the same way; the appearance transport, seen-to-unseen, and uniform-class priors act on the static Gaussians only.

A.6 Regularizers

The three terms of $\mathcal{L}_{\text{reg}}$ are

$$\mathcal{R}_{\text{tv}} = \sum_{u\in\{s,t\}} \sum_{l=1}^{3} \frac{n_l}{n_1} \underset{\text{axes, cells, channels}}{\text{mean}} \big(\nabla G_l^u\big)^2, \tag{7}$$

$$\mathcal{R}_{\text{knn}} = \frac{\sum_i \sum_{j\in\mathcal{N}_k(i)} w_{ij} \, \|r_i - r_j\|_2^2}{\sum_i \sum_{j\in\mathcal{N}_k(i)} w_{ij}}, \qquad w_{ij} = \max\big(0, \cos(f_i, f_j)\big), \tag{8}$$

$$\mathcal{R}_E = \frac{1}{3M} \sum_{i=1}^{M} \|E_i\|_1. \tag{9}$$

In Eq. 7, ∇G_l is the per-channel finite difference between neighboring cells along each axis, squared and averaged over the three axes, all cells, and the three channels; the penalty is applied to the source and target grids separately rather than to their difference, and the weight n_l/n_1 penalizes finer levels more strongly. In Eq. 8, $\mathcal{N}_k(i)$ are the k=8 Euclidean nearest neighbors of the Gaussian centers and f_i is the fixed semantic feature of section A.1; edges with $w_{ij} = 0$ contribute nothing and edges between Gaussians of different classes typically carry small but non-zero weights, so smoothing across semantic boundaries is suppressed rather than removed.

A.7 Optimization and Hyperparameters

All plug-in parameters, i.e., the global gain, the spatial-field grids, the per-Gaussian residual, the emission, and the sky correction, are optimized jointly for 8,000 iterations with a single Adam optimizer at a learning rate of 10^{-2}; no level of the hierarchy is frozen, introduced later, or trained with a separate schedule. Each iteration renders one anchor image, cycling through the $3K$ anchor images in a fixed order. The seen-to-unseen prior and the second part of $\mathcal{L}_{\text{chroma}}$ are the only terms with a schedule: they are enabled at iteration 4,000, when their reference medians are first computed from the visible Gaussians. Table 6 lists the loss weights and the remaining constants; the same values are used for all scenes and directions.

Table 6: Hyperparameters of PAM-ToD, shared by all experiments.

Loss weights	$\lambda_{\text{anc}} = 20$, $\lambda_{\text{ssim}} = 0.2$, $\lambda_{\text{atp}} = 1$, $\lambda_{\text{unseen}} = 0.1$, $\lambda_{\text{rch}} = 1$, $\lambda_{\text{ch}} = 1$, $\lambda_{\text{chu}} = 2$
Regularizers	$\lambda_{\text{tv}} = 10^{-2}$, $\lambda_{\text{knn}} = 10^{-2}$, $\lambda_E = 10^{-4}$
Spatial field	3 levels ($n_l \in \{16, 32, 64\}$), TV weight $n_l/16$
Residual graph	$k = 8$ neighbors
Robust anchor	$\sigma = 2$ pixels, $\rho = 10\%$
ATP prior	$Q = 32$ bins, classes with $< 2{,}000$ anchor px or $< Q$ Gaussians skipped, 65,536 samples/iter
Seen-unseen	edges $\{2, 4, 8, 16, 32\}$ m, min. 500 Gaussians/cell, update every 200 iters
Emission init	$b_i = -6$, $\hat{c}_E = \mathbf{1}$

Table 7: Sensor mounting relative to the front camera, in a camera basis with x forward, y left, and z up. Location in m, rotation in deg.

Sensor	Loc (x, y, z)	Rot (roll, pitch, yaw)
Front camera (ID 0)	$(0, 0, 0)$	$(0, 0, 0)$
Front-left camera (ID 1)	$(-0.045, 0.116, -0.001)$	$(-0.05, -0.58, 55.00)$
Front-right camera (ID 2)	$(-0.049, -0.070, 0.001)$	$(-0.55, -0.44, -55.00)$
LiDAR	$(-1.541, 0.021, -0.300)$	$(-0.70, 0.42, -0.01)$

B CARLA-ToD Datasets Details

This appendix complements the dataset description of the main paper with the generation protocol, the sensor configuration, the time-of-day settings, and the evaluation splits of CARLA-ToD.

B.1 Generation Protocol

Route and timing. All sequences are generated deterministically in synchronous mode with a fixed simulation tick, and the sensors are time-synchronized at 10 Hz; each sequence lasts 20 s (200 frames).

Static and Dynamic configurations. The Static configuration spawns no actor other than the ego vehicle, so each sequence contains only the static geometry of the town, and its dynamic-object masks are empty. The Dynamic configuration additionally spawns 40–60 vehicles and 40–55 pedestrians that move through the scene; their per-frame poses and box sizes are exported and coincide exactly across the three times of day, so corresponding Dynamic frames again differ only in illumination. The ego route is the same in both configurations, so the Static and the Dynamic sequences of a town share their camera poses as well.

Time-of-day settings. The weather is fixed to clear and only the sun altitude angle is changed: 2.0° for Morning, 90.0° for Noon, and -1.0° for Night; all other weather parameters and the camera settings are shared by the three conditions. With the sun below the horizon, the simulator switches on its artificial light sources (street lights and illuminated building windows), so Night combines dark unlit regions with local emitters, whereas Morning and Noon are lit by the sun alone and differ in exposure and color temperature.

B.2 Sensor Mounting and Imaging Parameters

Sensor rig. A rigid rig of three RGB cameras (Front, Front-left, Front-right) and a roof-mounted LiDAR is attached to the ego vehicle. The two side cameras are yawed by $\pm 55^\circ$, so the three 60° fields of view cover about 170° horizontally with a small overlap between neighboring cameras. Table 7 gives the mounting geometry relative to the front camera; the per-sequence extrinsics are constant and are exported for every frame as camera-to-world matrices in the same basis.

Image formation. Each camera records 1920×1300 8-bit sRGB images at 10 Hz with a 60° horizontal field of view; the cameras are ideal pinholes without lens distortion. The intrinsics follow from the field of view and the image size,

$$f_x = f_y = \frac{W}{2\tan(\mathrm{FOV}/2)}, \qquad c_x = \frac{W}{2}, \qquad c_y = \frac{H}{2}, \tag{10}$$

which gives $f_x = f_y \approx 1662.77$, $c_x = 960$, and $c_y = 650$ for W=1920, H=1300, and FOV=60°, shared by the three cameras.

LiDAR. The LiDAR has 64 channels and runs at 10 Hz with a range of 100 m and a vertical field of view of $[-25^\circ, 15^\circ]$; each scan contains about 1.1–1.2×10^5 returns with a per-point intensity and is exported together with its pose. The scans are used only to initialize the Gaussians and the scene bounds of the base model; neither the plug-in optimization nor the evaluation uses them.

Masks and annotations. A binary sky mask, aligned with the RGB frame, is provided for every image. The Dynamic configuration additionally provides per-frame binary masks of the dynamic objects (all, vehicles, and pedestrians) for every camera and, for every actor, per-frame 3D bounding boxes given as an object-to-world pose and a box size; the base model uses the boxes to separate the dynamic-object Gaussians from the static scene.

Table 8: Model size on CARLA-ToD Static.

Method	Gaussians	Params	Size
GS-IR	1.44 M	122.9 M	492 MB
GI-GS	1.58 M	125.9 M	504 MB
GS-W	1.91 M	173.7 M	695 MB
StreetGS	1.48 M	106.2 M	425 MB
+ Ours	1.48 M	113.9 M (+7.7 M)	456 MB (+31 MB)

B.3 Evaluation Protocol

Source training views. Each sequence contains 200 synchronized captures (600 images). Under the *Reconstruction* protocol the source 3DGS is trained on all 200 captures of the source time. Under the *NVS* protocol every eighth capture is held out, and the source 3DGS is trained on the remaining captures.

Target anchors. The K anchor captures are taken at the target time at frames spaced uniformly along the route, $\lfloor (k+\frac{1}{2})\,200/K \rfloor$ for $k = 0, \ldots, K-1$ (frame 100 for K=1; frames $12, 37, \ldots, 187$ for K=8). Under the NVS protocol an anchor that falls on a held-out frame is moved to the nearest training frame, so no held-out view is ever observed by the plug-in.

Evaluation views. NVS is scored on the held-out captures only, which are unseen by both the source 3DGS and the plug-in.

C Additional Results

C.1 Additional Results on Model size.

As shown in Table 8, PAM-ToD adds only four floats per Gaussian for the residual and emission terms alongside scene-shared spatial grids totaling 7 MB, resulting in 7.7 M parameters and representing a mere 7% increase over the frozen StreetGS base model. In contrast, inverse rendering methods assign normals, albedo, roughness, and metallic parameters to every Gaussian while additionally maintaining lighting and shading volumes. Furthermore, GS-W attaches per-Gaussian appearance features $F \in \mathbb{R}^{16(K+1)\times H\times W}$. While StreetGS requires a separate 3DGS model for each time of day, our method attaches a lightweight 31 MB plug-in per target time to a single frozen base, reducing the total footprint for three time slots from 1.3 GB to 0.49 GB. Although conventional baselines enable relighting within a unified framework, they suffer from trade-offs between reconstruction quality and rendering speed. On the other hand, our approach preserves high reconstruction fidelity and real-time rendering speed while minimizing model capacity and reusing pre-trained 3DGS assets.

C.2 Results Details on CARLA-ToD Static.

Tables 9 and 10 present detailed results for time-of-day adaptation across individual time slots in scene reconstruction and novel view synthesis, respectively. Existing inverse rendering methods achieve time-of-day adaptation without significant accuracy drops when transferring between morning and night. However, they suffer severe quality degradation under transitions such as noon to morning or night, as well as night or morning to noon. This indicates that conventional inverse rendering fails to predict consistent material properties required for physically based rendering across different times of day for the same scene, demonstrating that simply swapping environment maps is insufficient for faithful time-of-day adaptation. Additionally, LumiGauss frequently suffers from optimization instability, failing to converge during training. While GS-W using appearance embeddings outperforms inverse rendering baselines, it still lacks sufficient expressive capacity. In contrast, our method consistently yields superior average performance even with a single synchronized anchor capture.

Table 9: Quantitative comparison of reconstruction on the CARLA-ToD Static Dataset. Mo, No, and Ni denote Morning, Noon, and Night, respectively.

ToD	Metric	GI-GS	GS-IR	LumiGauss	GS-W	Ours (K=1)	Ours (K=8)
Mo → No	PSNR ↑	11.91	12.37	7.50	19.47	20.93	24.03
	SSIM ↑	0.697	0.689	0.428	0.815	0.777	0.827
	LPIPS ↓	0.362	0.382	0.609	0.265	0.251	0.237
No → Mo	PSNR ↑	19.81	18.65	7.66	15.63	25.30	27.00
	SSIM ↑	0.736	0.707	0.441	0.823	0.798	0.858
	LPIPS ↓	0.306	0.318	0.587	0.287	0.278	0.234
Mo → Ni	PSNR ↑	19.97	20.24	19.23	19.67	22.39	26.93
	SSIM ↑	0.680	0.686	0.624	0.739	0.737	0.835
	LPIPS ↓	0.312	0.336	0.529	0.309	0.289	0.233
Ni → Mo	PSNR ↑	20.86	20.46	19.13	19.13	23.77	27.34
	SSIM ↑	0.681	0.699	0.627	0.731	0.741	0.832
	LPIPS ↓	0.317	0.337	0.550	0.292	0.315	0.263
No → Ni	PSNR ↑	15.93	16.14	7.20	14.15	22.01	26.24
	SSIM ↑	0.581	0.572	0.408	0.706	0.704	0.821
	LPIPS ↓	0.387	0.408	0.637	0.345	0.324	0.250
Ni → No	PSNR ↑	9.74	10.63	7.04	12.50	17.93	22.44
	SSIM ↑	0.602	0.628	0.415	0.675	0.689	0.786
	LPIPS ↓	0.433	0.452	0.667	0.374	0.336	0.281

Table 10: Quantitative comparison of novel view synthesis (NVS) on the CARLA-ToD Static Dataset. Mo, No, and Ni denote Morning, Noon, and Night, respectively.

ToD	Metric	GI-GS	GS-IR	LumiGauss	GS-W	Ours (K=1)	Ours (K=8)
Mo → No	PSNR ↑	12.74	12.98	7.51	16.41	20.70	23.56
	SSIM ↑	0.719	0.715	0.400	0.780	0.757	0.806
	LPIPS ↓	0.339	0.352	0.634	0.310	0.252	0.235
No → Mo	PSNR ↑	20.42	19.11	7.99	15.65	25.04	27.00
	SSIM ↑	0.744	0.731	0.437	0.780	0.788	0.848
	LPIPS ↓	0.289	0.301	0.601	0.310	0.273	0.223
Mo → Ni	PSNR ↑	19.72	20.19	19.03	20.97	22.30	26.60
	SSIM ↑	0.672	0.684	0.600	0.724	0.724	0.820
	LPIPS ↓	0.315	0.330	0.551	0.304	0.290	0.237
Ni → Mo	PSNR ↑	20.93	20.56	19.28	19.02	23.53	27.09
	SSIM ↑	0.680	0.698	0.597	0.723	0.729	0.820
	LPIPS ↓	0.305	0.332	0.574	0.292	0.312	0.260
No → Ni	PSNR ↑	15.95	16.17	7.21	15.46	21.94	26.10
	SSIM ↑	0.581	0.578	0.410	0.711	0.697	0.812
	LPIPS ↓	0.382	0.397	0.647	0.334	0.322	0.249
Ni → No	PSNR ↑	9.84	10.40	6.88	11.99	17.85	22.33
	SSIM ↑	0.604	0.618	0.399	0.654	0.677	0.774
	LPIPS ↓	0.424	0.452	0.690	0.394	0.333	0.278

C.3 Qualitative Evaluation of Ablation Studies.

Ablation on the number of anchors. Figure 5 shows the results of the comparison of the number of anchors. All metrics improve monotonically with the number of anchor captures, with gains that diminish beyond K=8: at that density the anchors already see most of the scene, and further captures mostly re-observe Gaussians that are already constrained. A single synchronized capture recovers most of the change, because the dominant part of a time-of-day transfer is a global change of exposure and color temperature, which the low-capacity global gain and spatial field extrapolate to the whole scene from a single capture. The benefit of additional anchors is strongly direction-dependent. The dark regions improve more than the lit ones as anchors are added: the former are explained by the gain, whose field becomes better constrained, while the latter depend on the emission of the observed Gaussians.

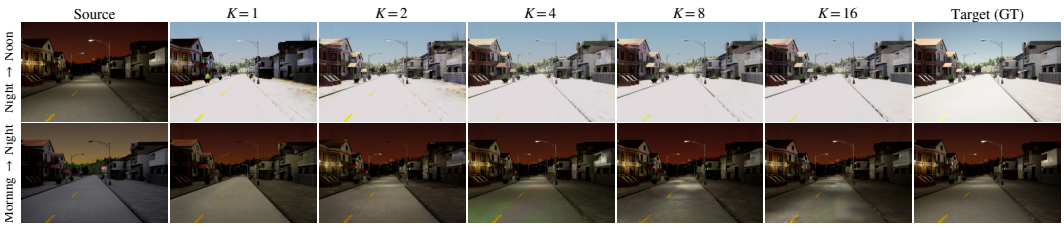


Figure 5: Qualitative comparison on of the number of anchors.

Ablation on our key illumination and emission term. Figure 6 presents qualitative results of the ablation study on our key illumination and emission terms. Removing these two components leads to fundamentally different failure modes. Without the illumination term, appearance transport relies entirely on a non-negative offset, restricting the model to brightening the scene. Consequently, transitions to darker times remain largely identical to the unadapted source, making it difficult to adjust local colors for brighter time slots because an offset merely shifts the color spectrum rather than scaling the underlying source texture. Conversely, removing the emission term still allows the gain to capture exposure, color temperature shifts, and large-scale shading variations, which functions reasonably well when the target is darker than the source. However, near-black source colors require large gains, which amplify their residual color noise into severe chromatic artifacts on road surfaces and walls; the additive term provides another way to increase brightness without such amplification. In the displayed examples (Figure 6), the decomposition into bright and dark regions is consistent with this functional separation: the emission term concentrates on active light sources and changes little in the remaining dark regions, so on its own it does not model fine local appearance shifts. Nevertheless, when only a single synchronized capture is available, the contribution of the emission term remains limited because per-Gaussian degrees of freedom cannot be fully constrained from a single capture, allowing the low-capacity illumination term to absorb most of the appearance shift. Overall, these two terms are non-redundant and mutually complementary, as an offset cannot darken a scene and a multiplicative gain cannot lift exactly zero channels, demonstrating that both components are indispensable for robust adaptation.

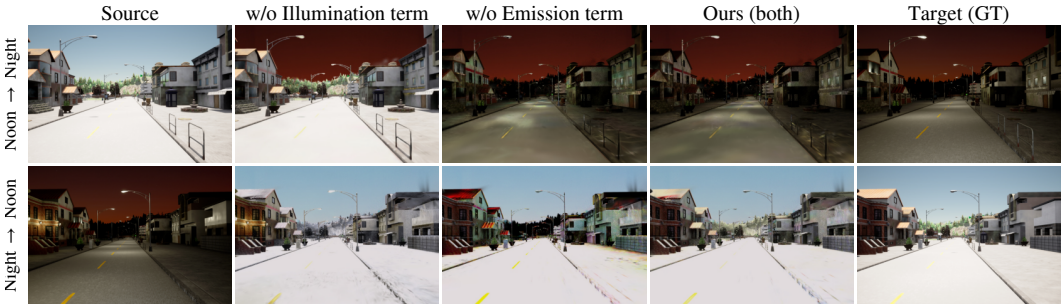


Figure 6: Qualitative comparison on our key illumination and emission term.

C.4 Additional ToD-Dataset and Real-World Datasets Results.

Figure 7 and Figure 8 present additional cross-sectional qualitative results regarding various segments of the CARLA-ToD Dataset. Figure 9 shows the qualitative results from the Waymo Open Datasets.

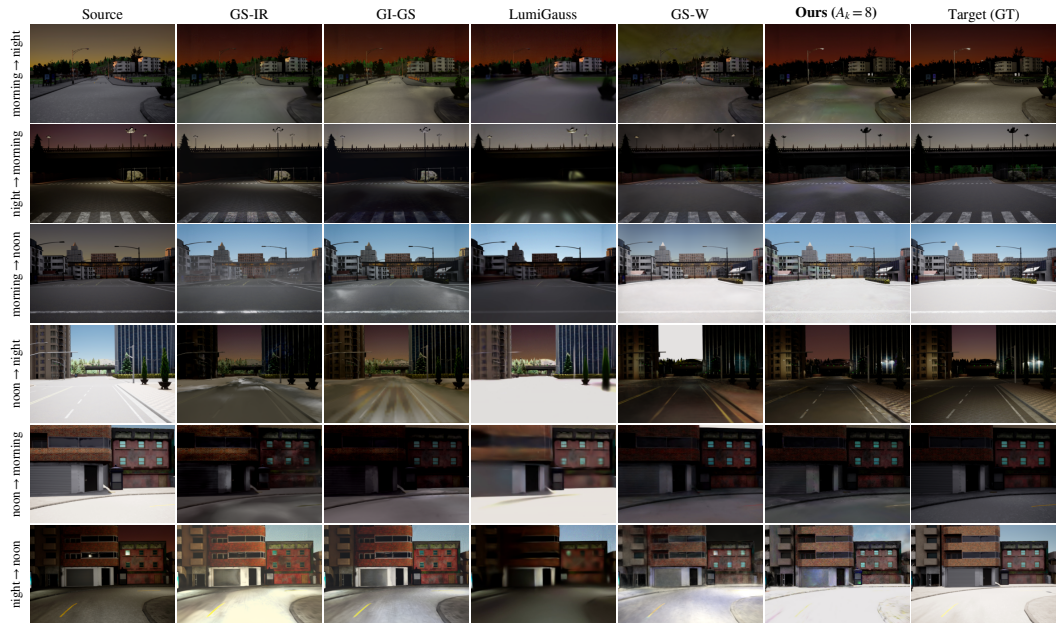


Figure 7: Additional qualitative results for NVS on CARLA-ToD static scenes.

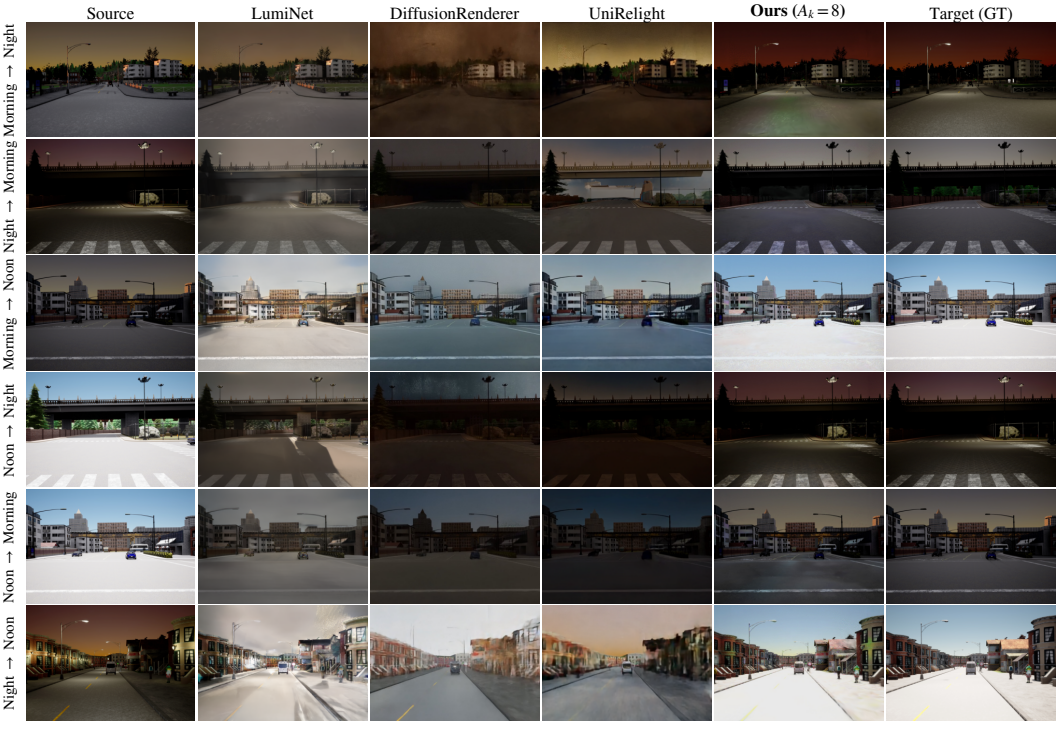


Figure 8: Additional qualitative results for NVS on CARLA-ToD dynamic scenes.

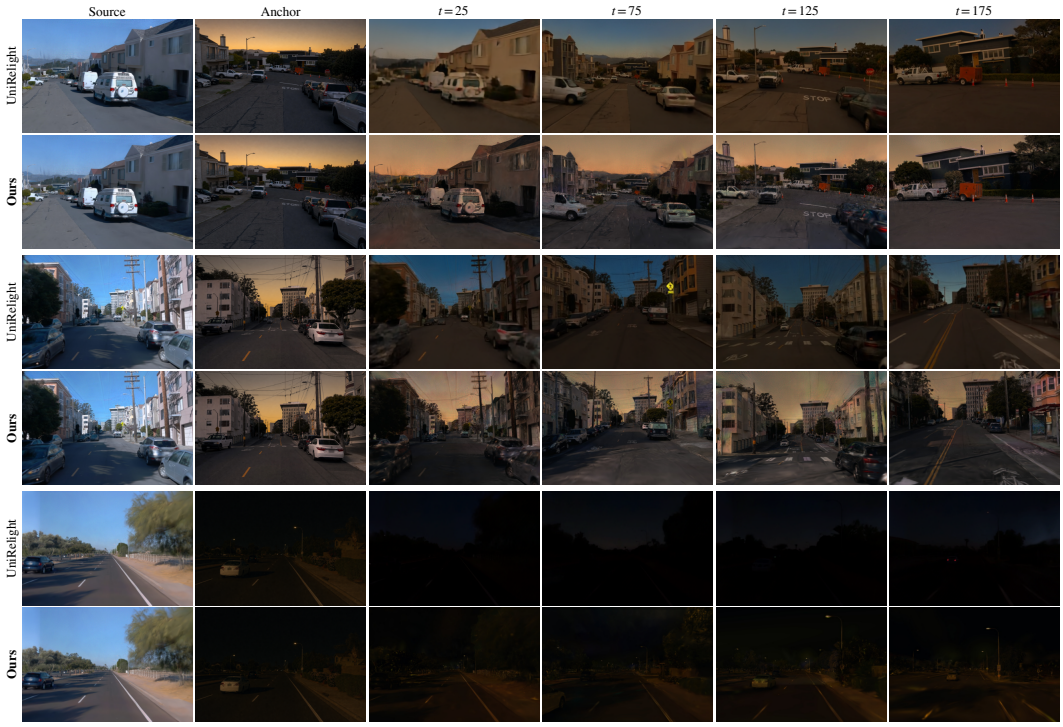


Figure 9: Additional qualitative results for NVS on WOD.