FaithFusion: Harmonizing Reconstruction and Generation via Pixel-wise Information Gain

Supplementary Material

A. Additional Implementation Details

A.1. Driving Scene Representation

FaithFusion's scene representation is built upon the decoupled 3D Gaussian Splatting (3DGS) structure widely adopted for driving scenes [8, 53, 69]. This representation is explicitly decomposed into three core components: a *Static Background* ($\mathcal{G}^{\text{bg}}$), *Dynamic Rigid Objects* ($\bar{\mathcal{G}}_v^{\text{rigid}}$ in local canonical space), and a *Sky Model* (C_{sky}), to precisely model geometry and motion through distinct semantic representations.

Static Background. The static background component is represented by a set of static Gaussian primitives $\mathcal{G}^{\text{bg}}$. These Gaussians are defined and optimized directly in the global world coordinate system, and all their attributes remain invariant over time.

Rigid Body Representation and Transformation. Gaussians belonging to a rigid object v are defined in its local canonical space. The set of Gaussians in this space, $\bar{\mathcal{G}}_v^{\text{rigid}}$, do not change their internal attributes (mean $\bar{\boldsymbol{\mu}}$, rotation $\bar{\mathbf{q}}$, scale $\bar{\mathbf{s}}$, opacity $\bar{o}$, and SH coefficients $\bar{\mathbf{c}}$) over time t. The object's motion is captured entirely by an external rigid transformation $\mathbf{T}_v(t) \in \text{SE}(3)$, which transforms the Gaussians into world space $\mathcal{G}_v^{\text{rigid}}(t)$:

$$\mathcal{G}_v^{\text{rigid}}(t) = \mathbf{T}_v(t) \otimes \bar{\mathcal{G}}_v^{\text{rigid}}. \tag{S1}$$

The transformation operator $\otimes$ specifically updates the mean position $\boldsymbol{\mu}(t)$ and rotation $\mathbf{q}(t)$ of the Gaussians when moved to the world coordinate system, while other attributes (scale, opacity, and SH coefficients) remain unchanged. We decompose the rigid pose as $\mathbf{T}_v(t) = (\mathbf{R}_v(t), \mathbf{t}_v(t))$, and the world-space mean position $\boldsymbol{\mu}(t)$ is obtained by:

$$\boldsymbol{\mu}(t) = \mathbf{R}_v(t)\bar{\boldsymbol{\mu}} + \mathbf{t}_v(t), \tag{S2}$$

the rotation $\mathbf{q}(t)$ is updated by composing the object's rotational component $\mathbf{R}_v(t)$ with the canonical rotation $\bar{\mathbf{q}}$:

$$\mathbf{q}(t) = \text{Rot}(\mathbf{R}_v(t), \bar{\mathbf{q}}), \tag{S3}$$

where $\text{Rot}(\cdot)$ denotes rotating the quaternion $\bar{\mathbf{q}}$ by the rotation matrix $\mathbf{R}_v(t)$. In this manner, the motion of dynamic objects is accurately modeled, ensuring geometric consistency across the time sequence.

Sky Model Compositing. The sky model is treated as a separate optimizable environmental texture map C_{sky} to fit large-scale appearance. The final pixel color C is obtained by α-blending the rendered Gaussian image $C_{\mathcal{G}}$ with the sky image C_{sky}, where $C_{\mathcal{G}}$ is rendered from all Gaussians ($\mathcal{G}^{\text{bg}}$ and $\{\mathcal{G}_v^{\text{rigid}}\}$):

$$C = C_{\mathcal{G}} + (1 - O_{\mathcal{G}})C_{\text{sky}}, \tag{S4}$$

where $O_{\mathcal{G}}$ is the rendered opacity mask accumulated from all Gaussian primitives. This strategy addresses the challenge of reconstructing unbounded distant scenes.

A.2. Evaluation Metrics

We conduct comprehensive quantitative evaluations following the protocol established by DriveDreamer4D [68], utilizing metrics that jointly assess the 3DGS novel view synthesis capability and the resulting spatiotemporal consistency under complex conditional shifts (e.g., lane shifts).

Spatiotemporal Coherence Metrics. (↑) To rigorously evaluate the coherence of dynamic elements and static scene structure, we employ two core metrics. Both quantify accuracy by comparing features detected in the rendered image with ground truth features that are geometrically projected from the original 3D scene onto the new trajectory view.

- **Novel Trajectory Agent IoU (NTA-IoU):** Measures the spatiotemporal accuracy of foreground dynamic agents (vehicles). Its computation involves detecting 2D bounding boxes on the rendered frames and comparing them to the projected ground-truth 3D bounding boxes. High NTA-IoU ensures accurate agent placement and adherence to the underlying 3D structure, leading to precise corrections in under-constrained regions.
- **Novel Trajectory Lane IoU (NTL-IoU):** Measures the geometric fidelity and spatiotemporal coherence of background lane lines. By comparing lane lines detected in the synthesized image against projected ground truth (often derived from the HDMap), it specifically verifies the integrity of the environment's static geometry. This metric reflects minimal disturbance to the original scene structure and guarantees environmental consistency.

Perceptual Quality Metric. (↓) We use Fréchet Inception Distance (FID) [15] to evaluate the overall visual realism and distributional quality of the 3DGS rendered novel view frames. FID calculates the distance between two multivariate Gaussian distributions fitted to the deep feature representations (from an Inception network) of generated frames and real frames. This score reflects the distribution-level similarity in a high-level perceptual space. A lower FID score indicates superior visual quality and consistent behavior across diverse viewpoints.

A.3. Expected Information Gain Derivations

We provide the detailed derivation for the Expected Information Gain (EIG) approximation utilized in the main paper (Equation 4). This derivation strictly follows the unified framework for Bayesian optimal experimental design and information-theoretic approximations [19, 25].

Motivation. The analytical computation of the EIG definition in Equation 4 is intractable, requiring evaluation of complex posterior parameter distributions and expectations over the observation space. To obtain a highly efficient and differentiable acquisition function for 3DGS, our goal is to derive a *computable upper bound* of the EIG. This is achieved by combining the *Laplace approximation* (to simplify the entropy terms, Prop. 3.2/3.5 in [25]) and the *log-determinant inequality* (to obtain the final trace form upper bound, Lemma 5.1 in [25]).

EIG Definition. The EIG quantifies the *predicted reduction in uncertainty of 3DGS model parameters* $\mathbf{\Omega}$ if a new observation (Y^{gt}_{NVS}) at the novel view X_{NVS} is acquired. Following the mutual information definition of EIG in [25] (Section 5.1), this uncertainty reduction is formally the mutual information between $\mathbf{\Omega}$ and Y^{gt}_{NVS} (conditioned on X_{NVS}):

$$\text{EIG} = I\left[\mathbf{\Omega}; Y^{\text{gt}}_{NVS} \mid X_{NVS}\right] = \mathbb{H}[\mathbf{\Omega}] - \mathbb{E}_{p(Y^{\text{gt}}_{NVS}|X_{NVS})}\left[\mathbb{H}\left[\mathbf{\Omega} \mid Y^{\text{gt}}_{NVS}, X_{NVS}\right]\right], \tag{S5}$$

where $\mathbb{H}[\mathbf{\Omega}]$ denotes the *prior entropy*, and $\mathbb{H}[\mathbf{\Omega} \mid Y^{\text{gt}}_{NVS}, X_{NVS}]$ denotes the *posterior entropy*.

! Note on Observation Index i and EIG Decomposition. The index i in the main paper's equations is used to denote different scopes of observation:

- In the optimization objective (Eq. 2), i denotes an individual training view.
- In the EIG definition (Eq. 4), i denotes an individual view Y_i^{NVS} within the novel view sequence Y_{NVS}. This formula calculates the information gain from a single such view.

The full EIG for the entire novel view sequence is obtained via view additivity. The final computable bound (Eq. 5) is then derived by applying Fisher information additivity principle (Prop. 4.2 in [25]) across all pixels j in every view i. Therefore, the summation index i in the final trace form (Eq. 5) is implicitly a flattened sum over all pixels in novel view sequence.

Laplace Proxy Justification. In Eq. (S5), $p(Y^{\text{gt}}_{NVS} \mid X_{NVS})$ is the true predictive distribution of real observations. We use the deterministic 3DGS rendered result Y_{NVS} as a computationally tractable proxy for Y^{gt}_{NVS}. This is justified by the Laplace approximation (Prop. 3.2 in [25]), which models 3DGS parameters $\mathbf{\Omega}$ as a Gaussian posterior around the MAP parameters ω^* ($\mathbf{\Omega} \sim \mathcal{N}(\omega^*, (H''[\omega^*])^{-1})$). Here, $H''[\omega^*]$ is the Hessian of the negative log-posterior, serving as the inverse prior covariance. Since Y_{NVS} is a deterministic function of $\mathbf{\Omega}$, Y_{NVS} (conditioned on ω^*) approximates $p(Y^{\text{gt}}_{NVS} \mid X_{NVS})$ for this Gaussian parameter posterior.

Laplace Approximation of Entropy. To compute the entropy terms in Eq. (S5), we apply the Gaussian differential entropy formula:$\mathbb{H}[\mathcal{N}(\mu, \mathbf{\Sigma})] = \frac{1}{2}\log\det(2\pi e\mathbf{\Sigma})$, where the covariance $\mathbf{\Sigma}$ is the inverse of the *observed information matrix* (Hessian of the negative log-posterior, Prop. 3.2 in [25]). For EIG, we distinguish two key observed information matrices:

- **Prior observed information:** $H''[\omega^*]$ (Hessian of the negative log-posterior of $\mathbf{\Omega}$ evaluated at ω^*, corresponding to $\mathbb{H}[\mathbf{\Omega}]$);
- **Posterior observed information:** $H''[\mathbf{\Omega} \mid Y_{NVS}] = H''[\omega^*] + H''[Y_{NVS} \mid \omega^*]$ (sum of prior information and novel-view information, via the information additivity principle in Prop. 4.2 of [25]), where $H''[Y_{NVS} \mid \omega^*]$ is the Hessian of the negative log-likelihood of Y_{NVS} (conditioned on ω^*).

Substituting these Gaussian entropy approximations into Eq. (S5) yields:

$$\begin{aligned}
\text{EIG} &\approx \frac{1}{2}\log\det\left(2\pi e(H''[\omega^*])^{-1}\right) - \mathbb{E}_{p(Y_{NVS}|X_{NVS},\omega^*)}\left[\frac{1}{2}\log\det\left(2\pi e\left(H''[\omega^*] + H''[Y_{NVS} \mid \omega^*]\right)^{-1}\right)\right] && \text{(S6)}\\
&= \frac{1}{2}\left[\log\det\left((H''[\omega^*])^{-1}\right) - \mathbb{E}_{p(Y_{NVS}|X_{NVS},\omega^*)}\left[\log\det\left(\left(H''[\omega^*] + H''[Y_{NVS} \mid \omega^*]\right)^{-1}\right)\right]\right] && \text{(S7)}\\
&= \frac{1}{2}\mathbb{E}_{p(Y_{NVS}|X_{NVS},\omega^*)}\left[\log\det\left(H''[\omega^*] + H''[Y_{NVS} \mid \omega^*]\right) - \log\det\left(H''[\omega^*]\right)\right] && \text{(S8)}\\
&= \frac{1}{2}\mathbb{E}_{p(Y_{NVS}|X_{NVS},\omega^*)}\left[\log\det\left(\mathbf{I} + H''[Y_{NVS} \mid \omega^*](H''[\omega^*])^{-1}\right)\right]. && \text{(S9)}
\end{aligned}$$

Trace Form Upper Bound and Pixel-Level Decomposition. We apply the log determinant inequality ($\log\det(\mathbf{I} + \mathbf{A}) \leq \text{tr}(\mathbf{A})$) (Lemma 5.1 in [25]) to Eq. (S9). By the linearity of the trace operator, and substituting the expectation of the novel-view Hessian $\mathbb{E}_{p(Y_{NVS}|X_{NVS},\omega^*)}\left[H''[Y_{NVS} \mid \omega^*]\right]$ with its Fisher information (Prop. 4.1 in [25]), we get:

$$\text{EIG} \leq \frac{1}{2}\mathbb{E}_{p(Y_{NVS}|X_{NVS},\omega^*)}\left[\text{tr}\left(H''[Y_{NVS} \mid \omega^*](H''[\omega^*])^{-1}\right)\right] \tag{S10}$$

$$= \frac{1}{2}\text{tr}\left(\mathbb{E}_{p(Y_{NVS}|X_{NVS},\omega^*)}\left[H''[Y_{NVS} \mid \omega^*]\right](H''[\omega^*])^{-1}\right) \tag{S11}$$

$$= \frac{1}{2}\text{tr}\left(H''[Y_{NVS} \mid X_{NVS},\omega^*](H''[\omega^*])^{-1}\right). \tag{S12}$$

Finally, leveraging the Fisher information additivity principle (Prop. 4.2 in [25]), the total Fisher information of Y_{NVS} is the sum of pixel-level Fisher information $H''[Y_{i,NVS} \mid X_{i,NVS},\omega^*]$ (one per pixel i). Substituting these pixel-wise Fisher information into Eq. (S12) yields the final trace-form approximation (main paper Equation 5):

$$\text{EIG} \leq \frac{1}{2}\sum_i \text{tr}\left(H''[Y_{i,NVS} \mid X_{i,NVS},\omega^*](H''[\omega^*])^{-1}\right). \tag{S13}$$

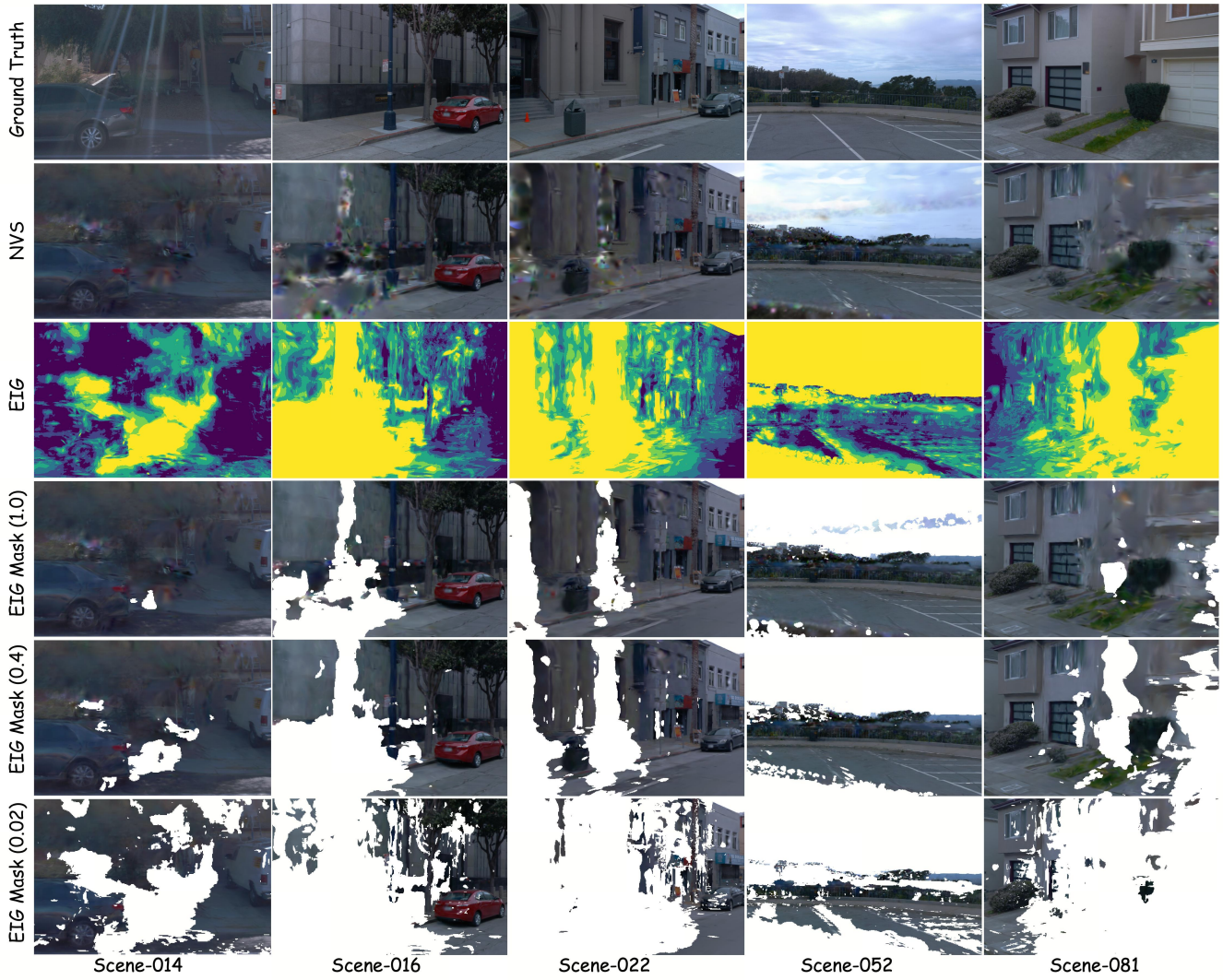

Figure S1. Visualization of EIG as a novel view synthesis quality proxy on representative Waymo [44] scenes. **Rows (Top to Bottom):** (1) Ground Truth, (2) Novel View Synthesis, (3) Pixel-wise EIG map (yellow = high EIG), (4-6) NVS masked by EIG thresholds ($\tau = 1.0, 0.4, 0.02$). White areas are excluded high-EIG pixels, confirming high EIG aligns with NVS artifacts. Sky regions are handled by a separate model and are excluded from this EIG analysis.

Figure S2. **Extended Qualitative Comparison on Waymo [44].** This figure provides additional novel view renderings for the same trajectory across representative methods, complementing the results shown in Fig. 5 of the main paper. Our method (last column) consistently maintains superior detail and fidelity across challenging regions, highlighted by the orange boxes, compared to methods [39, 49, 50].

B. Additional Visualization Results

B.1. EIG Correlation Validation and Evaluation Protocol

As quantified in Fig. 3 of the main paper, our cross-camera evaluation validates that pixel-level EIG is highly correlated with NVS quality. We detail the specific evaluation protocol here.

We first compute co-visible frame sequences between target cameras based on their Field of View (FoV) to ensure sufficient multi-view observation redundancy—a factor critical for optimizing 3D geometry and rendering fidelity. Considering that large low-frequency regions (e.g., solid colors, low-light scenes, or smooth surfaces) disproportionately inflate PSNR scores in standard NVS evaluations, which fails to reflect the model's ability to capture complex 3D structure, we filtered out frames predominantly containing such low-frequency information when assessing EIG-NVS correlation. This procedure ensures our validation efforts focus on EIG's efficacy in high-frequency detail and intricate geometry, effectively eliminating the inherent PSNR bias. Following this rigorous filtering process, we identified $4,245$ evaluation pairs for correlation validation.

For qualitative validation, Fig. S1 visualizes the EIG map and subsequent masking results on representative Waymo [44] scenes. The figure confirms the correlation: high EIG consistently aligns with NVS artifacts, and progressively masking these high-EIG pixels yields substantially improved perceptual clarity. This direct visual evidence not only validates the intuitive link between EIG and synthesis quality but also underscores EIG's unique advantage as a reliable, pixel-level proxy: it requires no manual annotation of artifacts, operates in a fully unsupervised manner, and provides fine-grained spatial guidance for targeted synthesis refinement.

B.2. More Qualitative Results

To complement the qualitative analysis in the main paper (Figs. 5 and 6), extended visualization results are provided in Fig. S2 and Fig. S3, where our key conclusions regarding scene synthesis quality and consistency are further validated. All high-resolution visuals and frame-by-frame trajectory comparisons (covering original and novel trajectories with 3 meters/6 meters lane shifts) are available in the accompanying *qualitative_supplement* folder.

Figure S3. **Detailed Ablation Study of EIG-Guided Components.** This figure provides an extended analysis by quantitatively measuring the incremental performance of integrating EIG-guided components into the OmniRe baseline (Sec. 4.3 in the main paper for the overview). The comprehensive results further confirm the significant role of EIG guidance in coherently integrating diffusion edits and distilling them back into the 3DGS structure, thus mitigating over-restoration and geometric drift.

Fig. S2 extends the main paper comparisons, showing *FaithFusion* significantly outperforms baselines ([39, 49, 50]) in preserving fine details (e.g., lane markings, building facades) and 3D coherence under large viewpoint shifts. Notably, the black regions in the upper part of the FreeVS [49] results were manually padded to align with the visualization scale of other methods, as its original output crops the sky region. Our method avoids spurious artifacts (ground bending, semantic mismatches) even at 6 meters lane offsets, aligning with our conclusion that EIG-guided control enables precise "generate-preserve" decisions.

Fig. S3 details our ablation results, validating EIG's role as a unified policy that harmonizes diffusion and 3DGS. Consistent with our core insight of replacing heuristics with information-theoretic guidance, EIG suppresses over-restoration in high-confidence regions (preserving 3DGS fidelity) and refines under-constrained areas (enhancing quality). This mechanism resolves the reconstruction-generation trade-off, successfully delivering the three key goals: consistency, quality, and faithfulness.