ReFlex: Faithful Panorama Reconstruction from a Single Image with Reflections

Yikai Pan **Zhiyuan Zhou** **Yuxiang Yan** **Xuetong Yang** **Runhe Yang** **Jian Pu**[†]
ykpan25@m.fudan.edu.cn, jianpu@fudan.edu.cn
Institute of Science and Technology for Brain-inspired Intelligence, Fudan University

Abstract

Reflections reveal scene content beyond the camera's direct field of view, but faithfully recovering this hidden information is challenging as reflective observations are heterogeneous, incomplete, and physically modulated. Existing methods either rely on explicit reflection modeling with privileged geometric priors or defer to generative synthesis that may hallucinate plausible but unfaithful content. To address these challenges, we propose **ReFlex**, an observation-constrained framework that casts faithful panorama reconstruction as a controlled inverse problem. We first develop an *Observation-Consistent Estimator* with hierarchical token prediction and signed routing to integrate heterogeneous reflection cues while stabilizing global structure. Subsequently, we propose a *Reliability-Guided Refiner*, which estimates spatial reliability from token predictions and guides a frozen diffusion model to refine uncertain regions without overwriting observation-supported structures. To validate the effectiveness of ReFlex, we further build ReFlex-Bench, a paired benchmark with 110,760 reflective RGB inputs and aligned target panoramas across diverse scenes, reflective geometries, and observation regimes. Experiments show that ReFlex delivers superior reconstruction fidelity without normals, field-of-view annotations, masks, or other privileged information, and obtains the best performance on eight of nine downstream full-view perception protocols.

1 Introduction

Reflective objects, such as mirrors, metals, glass, and glossy surfaces, naturally embed secondary observations within a single RGB image [1, 2]. By capturing indirect views beyond the camera frustum, they reveal broader environmental information [3, 4]. This rich information is implicitly contained in the reflected regions of the input image itself, requiring no extra camera views, depth measurements, or ground truth annotations. This raises a natural question: can reflections in a single RGB image serve as faithful geometric and semantic observation cues for regions beyond the camera frustum?

Existing approaches have made preliminary explorations by formulating the problem as a panorama reconstruction task, generally diverging into two distinct paradigms. As shown in Fig. 1(b), early methods rely on privileged priors, such as object geometry [4, 5] or reflectance map oracles [6] to explicitly extract the environmental information from reflections. While these methods maintain a certain level of faithfulness to the observations, they suffer from a significant bottleneck stemming from their heavy reliance on privileged physical priors. Propelled by the rapid advancements in generative models [7, 8], recent works have shifted towards a conditional generation paradigm, as shown in Fig. 1(c). By treating the observed image as a conditional constraint, these approaches leverage their exceptional capacity for high-fidelity content synthesis to achieve strong visual realism [9, 10, 11]. However, these generative models often struggle to faithfully attend to the actual observation cues embedded in reflections. Instead of faithful reconstruction, they tend to hallucinate plausible but

[†]Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

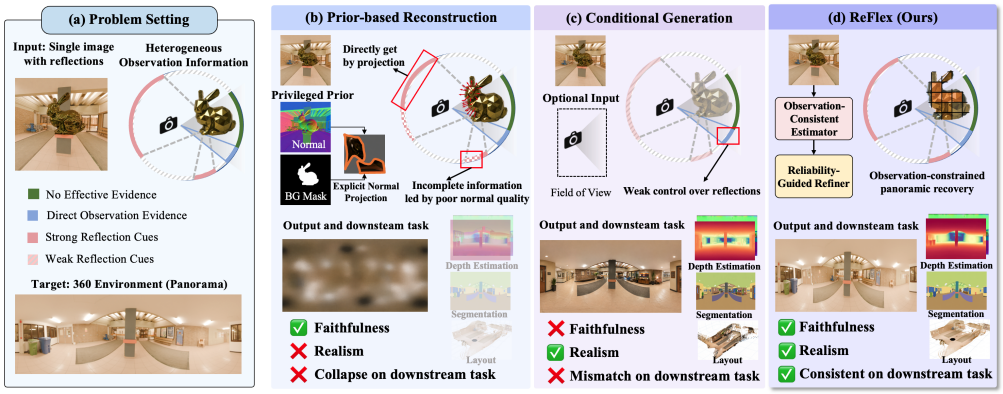

Figure 1: Motivation and problem setting. (a) Reflections provide heterogeneous cues for pose-aligned 360° panorama reconstruction. (b) Prior-based Reconstruction relies on privileged geometry priors, while (c) Conditional Generation may hallucinate inconsistent content. (d) *ReFlex* uses an *Observation-Consistent Estimator* and a *Reliability-Guided Refiner* to achieve faithful, realistic, and downstream-consistent recovery without privileged information.

ungrounded content, thus limiting their reliability for downstream geometric and semantic perception tasks.

The fundamental bottleneck in achieving such faithfulness lies in the intrinsic nature of the observation information for regions. A single reflective image provides only partial and spatially heterogeneous observation information of the surrounding environment. As illustrated in Fig. 1(a), the input captures direct observation evidence and reflection-mediated cues to out-of-frustum environment. The quality of these indirect cues varies substantially depending on surface properties and viewing conditions [4, 5]. This uneven distribution of observation information poses two coupled challenges for faithful panorama reconstruction. The model must effectively leverage direct observation evidence and strong reflection cues as reliable constraints, while also completing regions supported by weak or missing cues without deviating from the observed environment context.

To address these challenges, we propose *ReFlex*, an observation-constrained framework for faithful panorama reconstruction from a single reflective image, as illustrated in Fig. 1(d). To fully exploit the heterogeneous observation information in the input, ReFlex first employs an *Observation-Consistent Estimator* that uses hierarchical token prediction to preserve observation-supported structures from strong reflection cues while retaining weaker ones. Furthermore, we introduce signed routing to selectively propagate observation constraints based on estimated support relations. For regions remaining underconstrained due to sparse evidence, ReFlex incorporates a Reliability-Guided Refiner. This component estimates spatial reliability to guide a frozen diffusion model [8], enhancing uncertain areas while safeguarding observation-consistent structures. These modules enable ReFlex to recover scene structures grounded in direct and reliable observation information, while using the diffusion refinement to compensate for regions where evidence is missing.

To validate the effectiveness of ReFlex, we further construct *ReFlex-Bench*, a large-scale paired benchmark comprising 110,760 samples that encompass a wide array of scenes, reflective geometries, and observation regimes. Extensive experiments show that ReFlex achieves superior faithfulness without any privileged priors, maintains strong distributional realism and achieves the best performance on eight of nine downstream full-view perception protocols. Real-world zero-shot results further show that ReFlex recovers recognizable scene layouts and major structures, including regions beyond the camera frustum.

Our contributions are threefold. (i) We formulate faithful panorama reconstruction from a single reflective image as a controlled inverse problem constrained by heterogeneous observation information. (ii) We propose *ReFlex*, an observation-constrained reconstruction framework with an *Observation-Consistent Estimator* for faithful structure recovery and a *Reliability-Guided Refiner* for improving visual quality. (iii) We build *ReFlex-Bench* and demonstrate superior fidelity and downstream full-view perception without oracle test-time privileged inputs.

2 Related Work

2.1 Environment Estimation from Reflective Objects

Recovering environment information from reflective objects is rooted in inverse rendering, where image observations constrain geometry, reflectance, and illumination [12, 13]. This problem is inherently ill-posed because material and lighting are entangled, and surface reflectance often attenuates high-frequency illumination [14, 6]. To reduce this ambiguity, prior work has largely focused on illumination-centric tasks, such as radiometric relighting or material–lighting decomposition, often relying on privileged cues including normals, masks, and reflectance maps [15, 16, 17, 3], or on indirect proxy probes [18, 4]. Recent stochastic approaches further address material–illumination ambiguity through diffusion or posterior sampling [6, 19], but they still primarily model reflections as radiometric or appearance observations.

Beyond illumination estimation, another line of work exploits reflections as geometric or radiance-field evidence. Glossy objects have been interpreted as radiance-field cameras [1], while mirrors and glass have been used to provide additional constraints for scene reconstruction [2]. Reflection-aware neural rendering methods further model mirror or planar reflections explicitly, preventing reflected content from being incorrectly explained as physical scene geometry [20, 21, 22, 23]. However, these methods are typically tailored to specific reflective configurations, leaving open the question of how reflected content can serve as spatially structured evidence for dense, full-view scene perception. ReFlex addresses this gap by formulating reflections as heterogeneous image-level observations for pose-aligned 360° panorama reconstruction, combining observation-consistent estimation for structurally constrained regions with reliability-guided refinement in weakly constrained areas.

2.2 Generative Environment Inference

Generative models have long been used to infer HDR illumination or environment maps from narrow fields of view [24, 25], including spatially varying indoor lighting [26, 27] and generative text- or image-conditioned lighting [28, 9]. Diffusion models further improve panorama completion and generation, including NFoV-to-360 outpainting [29], distortion-aware completion [30], text-conditioned panorama synthesis [31], and cubemap-based generation [32]. Related diffusion-based lighting systems use generated chrome-ball probes for light estimation [10], estimate indoor illumination from a single image [33], or predict HDR environment lighting with video diffusion transformers [11]. Other diffusion-based inverse-rendering methods model illumination recovery as a posterior or stochastic inverse problem [19, 6].

Although these methods provide strong realism, directly using generative methods may hallucinate input-inconsistent geometry, layout, or out-of-view content, especially when reflection cues are spatially non-uniform. ReFlex therefore does not use diffusion as the primary reconstruction engine; it first reconstructs the panorama from direct and reflection-mediated evidence, then estimates spatial reliability to guide a frozen diffusion model toward weakly constrained regions while preserving observation-supported structures.

3 Method

Faithful panorama reconstruction from a reflective image is challenging because the available evidence is partial, spatially heterogeneous, and unevenly informative. The key difficulty is to extract valid reconstruction constraints from this heterogeneous evidence while avoiding localized quality degradation in weakly observed or unobserved regions. ReFlex addresses this challenge by progressively extracting and calibrating observation cues to recover observation-supported structures, while invoking generative priors only to compensate for residual ambiguity.

3.1 Problem Formulation and Overview

Problem Formulation. Given an RGB image with reflective objects $I \in \mathbb{R}^{H_I \times W_I \times 3}$, our goal is to reconstruct a pose-aligned equirectangular panorama $Y : \mathbb{S}^2 \to \mathbb{R}^3$ centered at the input camera. We formulate the image formation process as

$$I = \mathcal{F}_\xi(Y) + \epsilon, \tag{1}$$

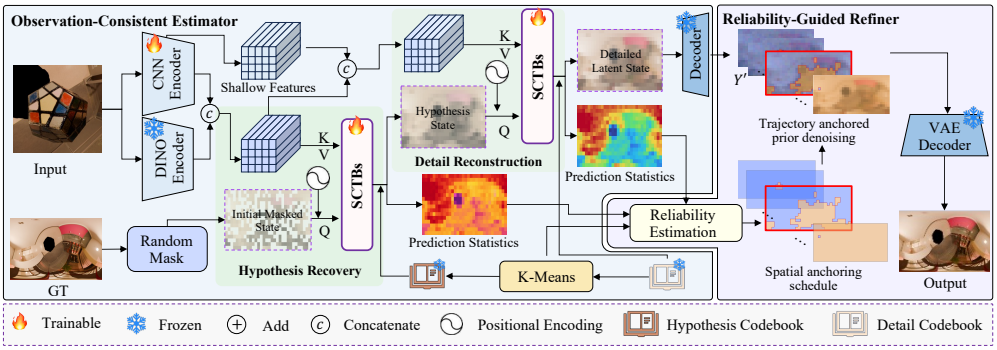

Figure 2: **Overview of ReFlex.** ReFlex first employs an Observation-Consistent Estimator (OCE) to stabilize dominant scene structures through hypothesis recovery and then reconstruct fine details, exploiting observation cues to improve structural fidelity. Both stages use stacked Support-Calibrated Transformer Blocks (SCTBs) to retrieve observation information and calibrate its propagation across panorama tokens. ReFlex then employs a Reliability-Guided Refiner (RGR) to improve the visual quality of underconstrained regions while preserving reliable structures. RGR uses the observation-grounded panorama Y' as a structural reference and estimates spatial reliability from OCE's prediction statistics, imposing stronger constraints on reliable predictions while allowing greater generative compensation in uncertain regions.

where $\mathcal{F}_\xi$ is a sample-dependent observation operator, ξ includes unknown camera, geometry, material, and visibility factors, and ϵ denotes noise. Since $\mathcal{F}_\xi$ is unknown and observations are incomplete, different regions are constrained to different degrees: directly visible or informative reflected regions provide structural constraints, while weakly observed or uncovered regions remain ambiguous.

Inspired by score-based diffusion models for inverse imaging [34], we conceptually formulate this non-uniform constraint as a balance between observation consistency and prior-based compensation:

$$Y_{\text{ideal}} = \arg\min_{Y} \mathcal{L}_{\text{obs}}(Y; I, a) + \lambda \mathcal{L}_{\text{prior}}(Y; 1-a), \tag{2}$$

where Y_{ideal} denotes an idealized reconstruction and $a : \mathbb{S}^2 \to [0, 1]$ is a conceptual reliability map. High a emphasizes observation consistency, whereas low a allows stronger prior-based compensation.

Framework Overview. ReFlex instantiates this formulation with two coupled modules:

$$\left(Y', C^{\text{oce}}\right) = \mathcal{E}_{\text{OCE}}(I), \qquad Y'' = \mathcal{R}_{\text{RGR}}\left(Y', C^{\text{oce}}\right). \tag{3}$$

As shown in Fig. 2, the Observation-Consistent Estimator E_{OCE} first extracts and integrates observation cues to recover observation-supported scene structures, yielding a panorama estimate Y' and prediction statistics C^{oce}. Since the observations are incomplete, we further employ a generative prior to improve content that remains insufficiently constrained. To keep this generative compensation consistent with the recovered scene structures, the Reliability-Guided Refiner R_{RGR} uses Y' as a structural reference and estimates prediction reliability at each location from C^{oce} to regulate the influence of the generative prior. It applies stronger structural constraints to more reliable predictions while allowing greater generative flexibility in uncertain regions.

3.2 Observation-Consistent Estimator

The *Observation-Consistent Estimator* aims to recover observation-supported panorama structures from direct-view evidence, reflection-mediated cues, and their induced relations across panorama tokens. The estimator operates in a robust latent space constructed from pretrained DINO [35] and VQGAN [36] representations: DINO extracts stable observation features, while VQGAN converts panorama reconstruction into a compact prediction task, improving stability and mitigating the ill-posedness of direct pixel-space recovery. The estimator combines two complementary designs. First, Support-Calibrated Transformer Blocks (SCTBs) retrieve observation information and calibrate panorama-token interactions through support-calibrated routing. Second, Hypothesis-to-Detail Latent Solving (H2D) organizes stacked SCTBs into two prediction stages, conditioning detail reconstruction on a coarser hypothesis-stage representation before recovering fine details.

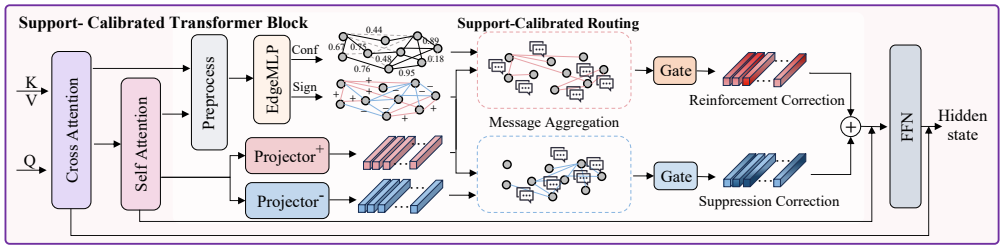

Figure 3: **Support-Calibrated Transformer Block (SCTB).** Cross-attention retrieves observation information, while self-attention integrates it across panorama tokens. Support-Calibrated Routing (SCR) reinforces observation-supported interactions and suppresses unsupported propagation through signed corrections.

Support-Calibrated Transformer Block. As shown in Fig. 3, a Support-Calibrated Transformer Block retrieves observation information from the input through cross-attention before integrating it across panorama tokens via self-attention. However, self-attention may propagate information along relations that are not sufficiently supported by the observations [37]. This is particularly problematic in regions with weak reflection cues, where attention-based generation may introduce subject leakage or poorly grounded interactions [38, 39]. To address this, the block incorporates Support-Calibrated Routing (SCR), which compares observation-support relations with internal self-attention propagation to suppress unsupported token-to-token information transfer.

Let $X^{\mathrm{sa}} \in \mathbb{R}^{N\times d}$ be self-attended panorama-token features, $A^{\mathrm{ca}} \in \mathbb{R}^{N\times L}$ the cross-attention from N panorama tokens to L observation tokens, and $A^{\mathrm{sa}} \in \mathbb{R}^{N\times N}$ the panorama self-attention map. SCR forms an observation-support relation map O and an internal propagation relation map P:

$$\tilde{A}^{\mathrm{ca}} = (A^{\mathrm{ca}})^{\gamma}, \qquad O = \tilde{A}^{\mathrm{ca}}(\tilde{A}^{\mathrm{ca}})^{\top}, \qquad P = A^{\mathrm{sa}}. \tag{4}$$

where O_{ij} measures the overlap of observation information attended by tokens i and j, P_{ij} denotes their self-attention-based propagation strength, and γ sharpens cross-attention evidence. Given an edge descriptor $\phi(O_{ij}, P_{ij}, x_i^{\mathrm{sa}}, x_j^{\mathrm{sa}})$, a shared edge MLP predicts a signed relation score:

$$\Delta_{ij} = f_\theta\big(\phi(O_{ij}, P_{ij}, x_i^{\mathrm{sa}}, x_j^{\mathrm{sa}})\big)\,, \tag{5}$$

where positive scores indicate observation-supported relations and negative scores indicate unsupported internal propagation.

SCR routes positive and negative edges separately:

$$R^{\pm} = \mathcal{N}(\mathrm{TopK}(\mathrm{ReLU}(\pm\Delta)^{\eta}))\,, \qquad m_i^{\pm} = \sum_j R_{ij}^{\pm} c_{ij}^{\pm} W_v^{\pm} x_j^{\mathrm{sa}}, \tag{6}$$

where TopK keeps the strongest row-wise edges, $\mathcal{N}$ normalizes each row by the row sum plus a small constant $\varepsilon_R > 0$, and returns a zero message when the pre-normalized row has zero mass. Here $\eta \geq 1$ sharpens confident edges, $c_{ij}^{\pm}$ are learned edge confidences, and $W_v^{\pm}$ are branch-specific value projections. The calibrated token feature is

$$x_i' = x_i^{\mathrm{sa}} + g_i^{+} W_o^{+} m_i^{+} - g_i^{-} W_o^{-} m_i^{-}\,, \tag{7}$$

where $g_i^{\pm}$ are token gates and $W_o^{\pm}$ are output projections. The positive branch reinforces evidence-consistent transfer, while the negative branch subtractively attenuates unsupported propagation.

Hypothesis-to-Detail Latent Solving. Directly predicting panorama latents from a single reflective image is highly underconstrained and may average incompatible structures. To make this prediction more tractable, H2D first solves in a compact hypothesis space to stabilize dominant scene structures, then recovers fine details conditioned on the hypothesis representation.

Let $\mathcal{D} = \{1, \ldots, K\}$ be the full detail codebook of the pretrained VQGAN tokenizer [36]. We partition $\mathcal{D}$ offline by k-means over codebook embeddings into Q disjoint groups $\{\mathcal{D}_q\}_{q=1}^{Q}$, and define the hypothesis vocabulary as $\mathcal{H} = \{1, \ldots, Q\}$. Each detail code $d \in \mathcal{D}_q$ is mapped to its hypothesis label q. A hybrid image encoder produces two observation memories: M^H for hypothesis recovery

and M^D for detail reconstruction. Both stages use stacked SCTBs for observation-conditioned token prediction.

For stage $s \in \{H, D\}$, masked token prediction is

$$p_t^s = F_s(M^s, B^s, z_t^s), \qquad z_{t+1}^s = \Gamma(z_t^s, p_t^s), \tag{8}$$

where z_t^s is the latent sequence at decoding step t, p_t^s is the token posterior predicted by solver F_s, B^s is optional stage context, and Γ commits confident tokens while leaving low-confidence positions revisable. At inference, the iterative commitment process runs for at most T_s steps for each stage $s \in \{H, D\}$, producing the final sequence $\hat{z}^s = z_{T_s}^s$.

In the hypothesis stage, $z_t^H \in \mathcal{H}^N$, $M^s = M^H$, and $B^s = \varnothing$. The reduced vocabulary suppresses detail-level variability and stabilizes dominant structures, including those carried by weak reflections. The hypothesis-stage hidden states are denoted by $R^H \in \mathbb{R}^{N \times d}$. In the detail stage, $z_t^D \in \mathcal{D}^N$, $M^s = M^D$, and $B^s = R^H$. The estimator reconstruction is

$$Y' = D_{\mathrm{vq}}(\hat{z}^D). \tag{9}$$

3.3 Reliability-Guided Refiner

Although Y' is grounded in direct-view and reflection-mediated observation information, weakly constrained, occluded, or uncovered regions may still lack valid evidence. The *Reliability-Guided Refiner* (RGR) estimates spatial reliability from the hierarchical token predictions of OCE and uses it to control a frozen diffusion model [8]. Rather than regenerating the panorama, RGR selectively releases the prior in low-reliability regions while anchoring high-reliability predictions to the observation-grounded reconstruction.

Reliability Scoring. Let R_n^H and R_n^D be the final hidden states of the hypothesis and detail stages at token n. We obtain the corresponding categorical predictive distributions:

$$\pi_n^H = \mathrm{softmax}(\mathrm{Head}_H(R_n^H)), \qquad \pi_n^D = \mathrm{softmax}(\mathrm{Head}_D(R_n^D)). \tag{10}$$

RGR computes reliability by jointly considering the sharpness of hierarchical token predictions and their consistency across the hypothesis and detail stages:

$$\hat{a}_n = \Phi(\pi_n^H, \pi_n^D, \{\mathcal{D}_q\}_{q=1}^Q) \in [0, 1]. \tag{11}$$

The function Φ combines prior-aware hypothesis concentration, within-group detail sharpness, ambiguity damping, and cross-stage agreement. Higher $\hat{a}_n$ indicates that OCE provides a more self-consistent and less ambiguous prediction for token n. We use $\hat{a}_n$ as a conservative reliability score for prior release. The full definition of Φ is provided in Appendix A.3.1.

Trajectory-Anchored Denoising. RGR converts the reliability map into a spatial anchoring schedule for diffusion refinement. Reliability is computed on the OCE token grid $\mathcal{T}$ and transferred to the diffusion latent grid $\mathcal{L}$ in log-SNR space. Given a T-step diffusion schedule $\{\bar{\alpha}_{\tau_s}\}_{s=1}^T$ [40], each latent location $\ell \in \mathcal{L}$ receives

$$\hat{s}_\ell = \arg\min_{s \in [T]} \left| \mathrm{logit}(\bar{\alpha}_{\tau_s}) - \mathcal{R}_{\mathcal{T} \to \mathcal{L}}(\mathrm{logit}(\mathrm{clip}(\hat{a}, \varepsilon_a, 1 - \varepsilon_a)))_\ell \right|, \qquad r_\ell = T - \hat{s}_\ell + 1. \tag{12}$$

Here $\hat{s}_\ell$ is the matched forward noising level and r_ℓ is the reverse-time anchoring duration. High-reliability locations remain anchored longer, while low-reliability locations are released earlier to the diffusion model.

We encode the OCE reconstruction as $x^{\mathrm{anc}} = E_{\mathrm{vae}}(Y')$ [8] and construct the anchor trajectory with fixed noise $\zeta \sim \mathcal{N}(0, I)$:

$$x_t^{\mathrm{anc}} = \sqrt{\bar{\alpha}_t}\, x^{\mathrm{anc}} + \sqrt{1 - \bar{\alpha}_t}\, \zeta. \tag{13}$$

At reverse step $t \to t - 1$, let x_{t-1}^{free} be the ordinary LoRA-adapted prior update and x_{t-1}^{anc} be the corresponding anchor state. With $s(t) = T - t + 1$ and maximum anchoring duration r_{ub}, RGR updates each location by

$$x_{t-1,\ell} = m_\ell^{(t)} x_{t-1,\ell}^{\mathrm{anc}} + (1 - m_\ell^{(t)}) x_{t-1,\ell}^{\mathrm{free}}, \qquad m_\ell^{(t)} = \mathbf{1}[s(t) \le \min(r_\ell, r_{\mathrm{ub}})]. \tag{14}$$

The final panorama is $Y'' = D_{\mathrm{vae}}(x_0)$. Through spatially varying anchoring, RGR uses the OCE reconstruction as a structural reference and allows greater generative flexibility in uncertain regions.

3.4 Training Objective

We optimize OCE and RGR in two separate stages. For OCE, the frozen VQGAN tokenizer encodes the ground-truth panorama Y into detail-token targets $z^{D,\star}$. Their codebook group assignments define the hypothesis targets $z^{H,\star}$, with $\star$ denoting ground-truth labels. Given randomly masked token sets $\mathcal{M}_H$ and $\mathcal{M}_D$, we minimize

$$\mathcal{L}_{\mathrm{OCE}} = \sum_{s\in\{H,D\}} \frac{\omega_s}{|\mathcal{M}_s|} \sum_{n\in\mathcal{M}_s} \mathrm{CE}(p_n^s, z_n^{s,\star}) \,, \tag{15}$$

where p_n^s denotes the predicted token distribution and ω_s weights the corresponding loss. The trainable OCE components are optimized jointly through this objective. The DINO encoder and VQ encoder/decoder remain frozen.

For RGR, we fix OCE and optimize only the LoRA parameters in the U-Net attention projections, keeping the diffusion backbone, VAE, and text encoder frozen. Given a clean panorama latent x_0, we sample a timestep t and noise $\epsilon \sim \mathcal{N}(0, I)$, and construct $x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$. The denoising objective is

$$\mathcal{L}_{\mathrm{TAD}} = \mathbb{E}_{x_0,t,\epsilon}\left[\left\|\epsilon_{\theta_0+\Delta\theta_{\mathrm{LoRA}}}(x_t, t) - \epsilon\right\|_2^2\right], \tag{16}$$

where θ_0 denotes the frozen diffusion weights and $\Delta\theta_{\mathrm{LoRA}}$ denotes the trainable LoRA parameters. At inference, the adapted prior is controlled by trajectory-anchored denoising. Further details are provided in Appendix B.

4 Experiments

4.1 Experimental Setup

Baselines and metrics. We compare ReFlex with oracle-assisted panorama estimation methods and conditional panorama generation methods. In RandBalls-Oracle, prior-based methods receive their required privileged annotations; in the full benchmark, normal-dependent baselines use MoGe-2 [41] estimated normals. We report PSNR, SSIM, LPIPS for reconstruction fidelity, FID and KID for distributional realism. We further test whether reconstructed panoramas preserve the geometry, semantics, and layout needed for full-view dense scene perception by evaluating downstream depth and normal estimation, semantic segmentation, and layout estimation with DA^2 [42], OOOPS [43], and LGTNet [44]. The metrics used for downstream task evaluation are described in Appendix D.

Benchmark. ReFlex-Bench contains paired reflective RGB inputs and pose-aligned 360° target panoramas rendered from 3,692 HDR environments collected from CGEES, HDRIMAPS, HDRI-Skies, Laval Indoor/Outdoor, and PolyHaven. For each environment, we render three regimes: RandBalls with dense and controlled reflection cues, Digital Gene [45] with complex reflective geometry and self-occlusion, and OmniObject3D [46] with scanned real-world objects. The benchmark spans controlled primitives, procedurally complex shapes, and real scanned assets. We sample camera pose and field of view, randomize illumination and material attributes, first split scenes 9:1 into training and test scenes, and carve out a validation subset from the training split, yielding an effective 8:1:1 train/validation/test split. Each sample is stored as a 1024×1024 RGB input paired with a 512×256 panorama target, as illustrated in Fig. 4.

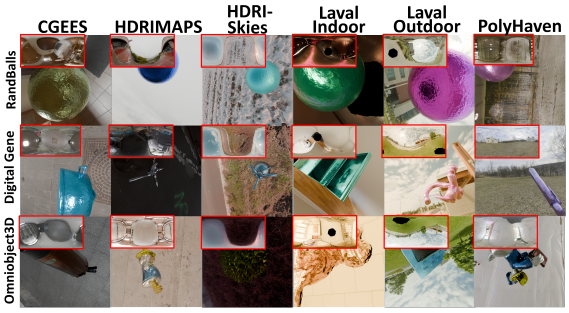

Figure 4: Dataset preview of reflective RGB inputs and paired pose-aligned target panoramas.

Implementation. ReFlex is trained with AdamW using a learning rate of 10^{-4} and a batch size of 2 for 80 epochs on one RTX 4090. OCE uses MaskGIT-style masked latent modeling with a cosine mask-ratio schedule, where the masking ratio is randomly sampled during training and progressively reduced during inference [47]. The masked-token objective is applied to both the hypothesis and detail stages. At inference, OCE performs eight MaskGIT-style iterations per stage, and RGR uses

Table 1: **Image-level recovery quality on ReFlex-Bench.** PSNR, SSIM, and LPIPS measure reconstruction fidelity; FID and KID measure distributional realism. Cue denotes additional test-time information beyond RGB: N = normal, M = region mask, and FoV = field of view. RandBalls-Oracle provides privileged annotations to cue-dependent methods, while full-benchmark columns use MoGe-2 estimated normals for normal-dependent baselines.

Method	Cue	RandBalls-Oracle					RandBalls			Digital Gene			OmniObject3D		
		PSNR↑	SSIM↑	LPIPS↓	FID↓	KID↓	PSNR↑	LPIPS↓	FID↓	PSNR↑	LPIPS↓	FID↓	PSNR↑	LPIPS↓	FID↓
Prior-based methods															
DeLightNet [17]	N/M	17.518	0.641	0.615	202.67	0.1640	15.723	0.683	257.46	14.184	0.722	308.98	13.729	0.741	301.78
Georgoulis et al. [5]	N/M	18.284	**0.657**	0.607	212.72	0.1837	16.488	0.640	224.67	14.014	0.687	202.96	14.140	0.603	224.08
DRMNet [6]	N/M	11.688	0.408	0.620	75.19	0.0531	11.117	0.608	93.09	10.594	0.633	160.75	–	–	–
Conditional generation methods															
LDMs Outpainting [8]	FoV/M	10.734	0.415	0.594	79.36	0.0465	10.734	0.594	79.36	8.481	0.673	98.61	8.454	0.677	99.73
LuxDiT [11]	none	16.688	0.585	0.368	**34.45**	**0.0127**	16.688	0.368	**34.45**	13.601	0.486	**43.02**	12.839	0.506	**43.72**
Our ReFlex-OCE	none	19.343	0.643	**0.314**	59.02	0.0323	19.343	**0.314**	59.02	**15.499**	0.445	79.14	14.802	0.473	92.50
Our ReFlex	none	**19.440**	0.654	0.316	42.27	0.0239	**19.440**	0.316	42.27	15.477	**0.444**	57.72	**14.810**	**0.467**	66.70

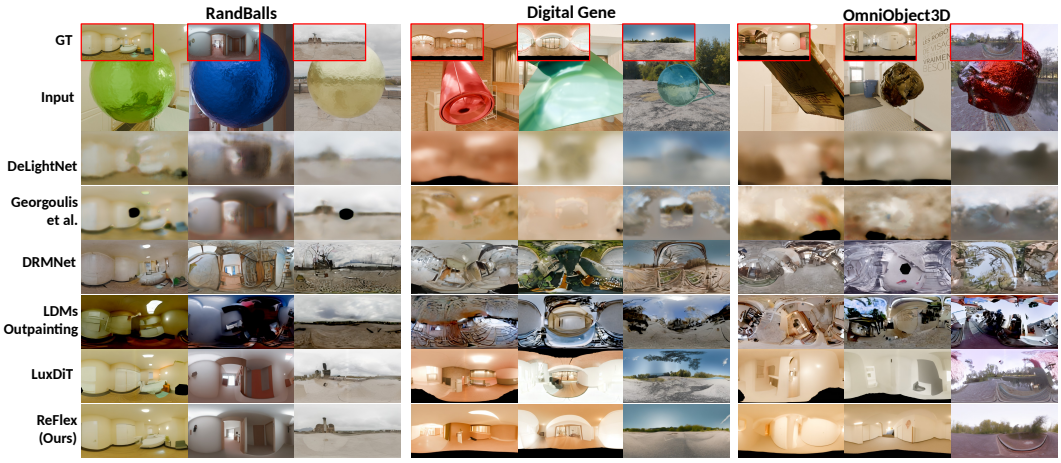

Figure 5: **Visual comparison on ReFlex-Bench.** Prior-based methods tend to erase fine structures in weakly constrained regions, whereas conditional generation methods may hallucinate structures that are inconsistent with the input observations. ReFlex preserves observation-supported geometry with OCE and refines uncertain regions with RGR. DRMNet on OmniObject3D is shown only as zero-shot visualization, where it is not applicable.

OCE's prediction statistics to guide generation refinement. Detailed configurations of all methods are provided in Appendix C, and complete downstream metric reports are provided in Appendix E.

4.2 Main Results

Image-level comparison. Tab. 1 evaluates ReFlex under both a controlled oracle setting and a practical non-oracle setting. On RandBalls-Oracle, methods relying on oracles receive privileged normals, masks, or field-of-view information, forming an upper-bound benchmark for prior-based estimation. Despite using only the RGB input, ReFlex matches or surpasses prior-based methods in reconstruction fidelity and achieves the lowest or near-lowest LPIPS, indicating stronger perceptual alignment with the paired ground truth. Meanwhile, ReFlex remains competitive with LuxDiT, a state-of-the-art generation method, in terms of FID and KID, indicating that its improved faithfulness does not come at a substantial cost to distributional realism. On the full ReFlex-Bench, oracle normals are replaced by MoGe-2 estimates for normal-dependent baselines to approximate practical test-time performance. ReFlex consistently achieves stronger fidelity across RandBalls, Digital Gene, and OmniObject3D, while maintaining competitive realism under complex reflective geometry and scanned real-world objects. These results demonstrate that ReFlex is robust to heterogeneous reflection cues and can recover faithful panoramas without privileged information.

Downstream full-view perception. Tab. 2 evaluates whether the recovered panoramas preserve scene information useful beyond image-level similarity. We feed the predicted panoramas from the

Table 2: **Downstream full-view perception.** We evaluate whether reconstructed panoramas support depth/normal estimation, segmentation, and layout estimation using DA^2, OOOPS, and LGTNet.

Method	RandBalls					Digital Gene					OmniObject3D				
	DA^2		OOOPS	LGTNet		DA^2		OOOPS	LGTNet		DA^2		OOOPS	LGTNet	
	AbsRel↓	Acc_{30° ↑	mIoU↑	3DIoU↑	δ_1 ↑	AbsRel↓	Acc_{30° ↑	mIoU↑	3DIoU↑	δ_1 ↑	AbsRel↓	Acc_{30° ↑	mIoU↑	3DIoU↑	δ_1 ↑
Georgoulis et al. [5]	1.378	0.368	0.210	0.146	0.319	1.734	0.339	0.042	0.170	0.347	1.610	0.363	0.005	0.157	0.340
DRMNet [6]	1.316	0.490	0.122	0.439	0.626	1.031	0.412	0.005	0.306	0.524	–	–	–	–	–
LDMs Outpainting [8]	1.254	0.463	0.152	0.328	0.511	1.426	0.401	0.033	0.304	0.524	1.528	0.396	0.030	0.313	0.519
LuxDiT [11]	0.769	0.401	0.385	0.355	0.543	1.128	0.438	0.280	0.320	0.499	1.495	0.428	**0.295**	0.333	0.517
Our ReFlex	**0.517**	**0.607**	**0.434**	**0.495**	**0.688**	**0.685**	**0.538**	**0.288**	**0.356**	**0.550**	**0.893**	**0.516**	0.261	**0.367**	**0.552**

practical non-oracle setting into off-the-shelf full-view perception models and evaluate depth/normal estimation, semantic segmentation, and indoor layout estimation, which jointly measure global geometry, semantic appearance, and spatial layout.

ReFlex achieves the best performance on eight out of nine downstream protocols. For depth estimation, it reduces AbsRel by 32.8%, 33.6%, and 40.3% over the strongest baselines on RandBalls, Digital Gene, and OmniObject3D, respectively; for normal estimation, it improves accuracy by 23.9%, 22.8%, and 20.6%. It also achieves the best mIoU on RandBalls and Digital Gene, and improves 3DIoU by 12.8%, 11.3%, and 10.2% for layout estimation.

The visualizations in Fig. 5 further support the trend of improved downstream consistency. Prior-based methods often smooth weakly constrained structures, whereas conditional generation methods may hallucinate plausible but input-inconsistent content. ReFlex instead preserves observation-supported structures while refining uncertain regions, yielding panoramas that better support downstream geometry, layout, and semantic understanding. The only exception is semantic segmentation on OmniObject3D, where LuxDiT obtains slightly higher mIoU, suggesting that generative models may produce semantically typical content without consistently preserving geometry or layout.

4.3 Real-World Zero-Shot Evaluation

To assess how ReFlex transfers beyond synthetic rendering, we further collect a paired real-world evaluation set using a Canon EOS 77D for reflective RGB inputs and an Insta360 X5 for panorama references. The set contains 208 inputs from 10 indoor/outdoor scenes and 14 reflector configurations, including colored spheres and everyday metalware with diverse geometries. No real image is used for training, validation, hyperparameter tuning, or checkpoint selection.

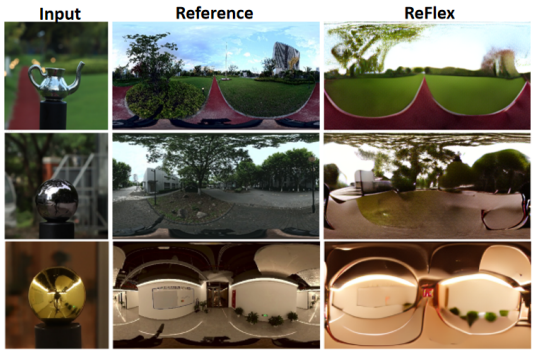

Figure 6: **Real-world zero-shot reconstruction.** Despite being trained only on synthetic data, ReFlex retains recognizable global scene organization and out-of-view structures, including dominant layout, horizon, and large-scale structural boundaries.

Because the reflective inputs and panorama references are captured with different cameras, we first estimate their relative orientation from correspondences in the directly visible background and then apply a global photometric calibration to compensate for exposure and color-response differences. Full acquisition, geometric/photometric alignment, and residual-error analyses are provided in Appendix H.

Across all 208 real-world samples, ReFlex achieves 12.751 ± 1.582 dB PSNR, 0.362 ± 0.105 SSIM, and 0.678 ± 0.093 LPIPS under a strict zero-shot setting. Compared with the synthetic inputs in Fig. 4, real photographs exhibit additional camera-induced effects, including background defocus blur and sensor noise. Despite this substantial sim-to-real gap, Fig. 6 shows that ReFlex recovers the coarse layout and major structures of the surrounding scene, including regions beyond the direct field of view, in broad agreement with the panorama references. Although challenging cases still exhibit smoothing and fine-detail loss, these qualitative results demonstrate non-trivial zero-shot transfer and provide promising evidence of the applicability of observation-grounded panorama reconstruction to real-world reflective images. These results further suggest that reflections in a single RGB image

can provide meaningful observation cues for recovering scene structures beyond the camera frustum. Additional analyses of score distributions, reflector- and scene-wise performance are provided in Appendix I.

4.4 Ablation Studies

Tab. 3 summarizes the contributions of architecture design, reliability-guided refinement, and input evidence. Starting from the VQ tokenizer baseline, DINO improves PSNR by 7.9% and reduces LPIPS by 17.1%, while H2D reduces FID by 47.1%, showing that robust observation encoding and hypothesis-to-detail solving improve complementary aspects of reconstruction. Combining DINO and H2D yields larger gains, improving PSNR by 24.3% and reducing LPIPS/FID by 40.7%/61.2% over the baseline. Full OCE further improves perceptual fidelity through support-calibrated routing, and adding RGR reduces FID by 28.4% with negligible changes in PSNR and LPIPS. This confirms the complementary roles of the two modules: OCE improves observation-consistent reconstruction, while RGR improves realism in uncertain regions.

Table 3: **Compact ablation studies on RandBalls.** PSNR and LPIPS measure fidelity, while FID measures realism.

Study	Variant	PSNR↑	LPIPS↓	FID↓
Arch.	VQ tokenizer baseline	15.192	0.551	176.89
	+ DINO encoder	16.389	0.457	125.04
	+ H2D solving	16.524	0.427	93.57
	+ DINO + H2D	18.877	0.327	68.60
	Full OCE, w/o RGR	19.343	**0.314**	59.02
	Full ReFlex	**19.440**	0.316	**42.27**
RGR	OCE only	19.343	**0.314**	59.02
	Reliability-guided	**19.440**	0.316	42.27
	Uniform release	19.130	0.319	**39.46**
	Entropy-based	18.993	0.324	40.78
Input	w/o reflective obj.	13.679	0.634	145.77
	w/o background	16.078	0.447	93.55
	Full input	**19.343**	**0.314**	**59.02**

The refinement and input ablations further clarify the role of reliability guidance and reflection cues. The uniform variant obtains the lowest FID but worsens PSNR and LPIPS, indicating that less constrained generation refinement improves distributional realism at the cost of input consistency. The entropy-based variant is a simple confidence-guided ablation with a similar limitation: it provides coarser control, whereas our reliability signal enables finer-grained refinement. The proposed reliability-guided refinement therefore provides a better fidelity–realism trade-off. Finally, removing the reflective object causes PSNR to drop by 29.3%, LPIPS to more than double, and FID to increase by 147.0%, while removing the visible background causes a smaller degradation. This suggests that ReFlex does not merely rely on dataset priors or visible-background extrapolation; reflection-mediated evidence provides the dominant out-of-frustum constraints for faithful panorama recovery.

5 Conclusion

In this work, we present ReFlex, an observation-constrained framework for faithful panorama reconstruction from a single reflective RGB image. ReFlex treats reflected content as heterogeneous observation cues in a controlled inverse problem. Its Observation-Consistent Estimator integrates direct and reflection-mediated evidence through hierarchical token prediction and signed routing, while its Reliability-Guided Refiner uses spatial reliability to improve uncertain regions with a frozen diffusion model without overwriting observation-supported structures. We demonstrate that ReFlex improves reconstruction faithfulness, maintains competitive realism, and benefits downstream full-view perception. Zero-shot real-world results further suggest that reflections in a single image can serve as practical cues for recovering scene content beyond the camera frustum.

Broader Impact. By recovering out-of-frustum scene content from reflections, ReFlex may benefit applications that require broader scene awareness under constrained sensing conditions, such as robotic navigation, embodied perception, and scene understanding. It may also facilitate full-view perception without dense capture or privileged test-time inputs.

Limitations. ReFlex-Bench is constructed in physically based simulation. Although our zero-shot real-world evaluation shows that ReFlex can recover coarse scene layouts and major structures, challenging cases still exhibit smoothing and fine-detail loss. Sensor noise, blur, and complex material responses may contribute to the remaining sim-to-real gap. Bridging this gap remains an important direction for future research.

Acknowledgments and Disclosure of Funding

This research received no external funding, and the authors declare that they have no competing interests.

References

[1] Kushagra Tiwary, Akshat Dave, Nikhil Behari, Tzofi Klinghoffer, Ashok Veeraraghavan, and Ramesh Raskar. Orca: Glossy objects as radiance-field cameras. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20773–20782, 2023.

[2] Thomas Whelan, Michael Goesele, Steven J Lovegrove, Julian Straub, Simon Green, Richard Szeliski, Steven Butterfield, Shobhit Verma, Richard A Newcombe, M Goesele, et al. Reconstructing scenes with mirror and glass surfaces. *ACM Trans. Graph.*, 37(4):102, 2018.

[3] Stamatios Georgoulis, Konstantinos Rematas, Tobias Ritschel, Efstratios Gavves, Mario Fritz, Luc Van Gool, and Tinne Tuytelaars. Reflectance and natural illumination from single-material specular objects using deep learning. *IEEE transactions on pattern analysis and machine intelligence*, 40(8):1932–1947, 2017.

[4] Hong-Xing Yu, Samir Agarwala, Charles Herrmann, Richard Szeliski, Noah Snavely, Jiajun Wu, and Deqing Sun. Accidental light probes. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12521–12530. IEEE, 2023.

[5] Stamatios Georgoulis, Konstantinos Rematas, Tobias Ritschel, Mario Fritz, Tinne Tuytelaars, and Luc Van Gool. What is around the camera? *Proceedings ICCV 2017*, 2017:5170–5178, 2017.

[6] Yuto Enyo and Ko Nishino. Diffusion reflectance map: Single-image stochastic inverse rendering of illumination and reflectance. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11873–11883. IEEE Computer Society, 2024.

[7] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[8] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[9] Guangcong Wang, Yinuo Yang, Chen Change Loy, and Ziwei Liu. Stylelight: Hdr panorama generation for lighting estimation and editing. In *European conference on computer vision*, pages 477–492. Springer, 2022.

[10] Pakkapon Phongthawee, Worameth Chinchuthakun, Nontaphat Sinsunthithet, Varun Jampani, Amit Raj, Pramook Khungurn, and Supasorn Suwajanakorn. Diffusionlight: Light probes for free by painting a chrome ball. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 98–108. IEEE Computer Society, 2024.

[11] Ruofan Liang, Kai He, Zan Gojcic, Igor Gilitschenski, Sanja Fidler, Nandita Vijaykumar, and Zian Wang. Luxdit: Lighting estimation with video diffusion transformer. *arXiv preprint arXiv:2509.03680*, 2025.

[12] Stephen Robert Marschner. *Inverse rendering for computer graphics*. Cornell University, 1998.

[13] Ravi Ramamoorthi and Pat Hanrahan. A signal-processing framework for inverse rendering. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 117–128, 2001.

[14] Stephen Lombardi and Ko Nishino. Reflectance and natural illumination from a single image. In *European Conference on Computer Vision*, pages 582–595. Springer, 2012.

[15] Konstantinos Rematas, Tobias Ritschel, Mario Fritz, Efstratios Gavves, and Tinne Tuytelaars. Deep reflectance maps. In *Proceedings of the IEEE Conference on computer vision and pattern recognition*, pages 4508–4516, 2016.

[16] Maxim Maximov, Laura Leal-Taixé, Mario Fritz, and Tobias Ritschel. Deep appearance maps. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8729–8738, 2019.

[17] Stamatios Georgoulis, Konstantinos Rematas, Tobias Ritschel, Mario Fritz, Luc Van Gool, and Tinne Tuytelaars. Delight-net: Decomposing reflectance maps into specular materials and natural illumination. *arXiv preprint arXiv:1603.08240*, 2016.

[18] Tristan Swedish, Connor Henley, and Ramesh Raskar. Objects as cameras: Estimating high-frequency illumination from shadows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2593–2602, 2021.

[19] Linjie Lyu, Ayush Tewari, Marc Habermann, Shunsuke Saito, Michael Zollhöfer, Thomas Leimkühler, and Christian Theobalt. Diffusion posterior illumination for ambiguity-aware inverse rendering. *arXiv preprint arXiv:2310.00362*, 2023.

[20] Junyi Zeng, Chong Bao, Rui Chen, Zilong Dong, Guofeng Zhang, Hujun Bao, and Zhaopeng Cui. Mirror-nerf: Learning neural radiance fields for mirrors with whitted-style ray tracing. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 4606–4615, 2023.

[21] Leif Van Holland, Ruben Bliersbach, Jan U Müller, Patrick Stotko, and Reinhard Klein. Tramnerf: Tracing mirror and near-perfect specular reflections through neural radiance fields. In *Computer Graphics Forum*, volume 43, page e15163. Wiley Online Library, 2024.

[22] Chen Gao, Yipeng Wang, Changil Kim, Jia-Bin Huang, and Johannes Kopf. Planar reflection-aware neural radiance fields. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–10, 2024.

[23] Jing Wu, Zirui Wang, Iro Laina, and Victor Adrian Prisacariu. Reflect3r: Single-view 3d stereo reconstruction aided by mirror reflections. *arXiv preprint arXiv:2509.20607*, 2025.

[24] Marc-André Gardner, Kalyan Sunkavalli, Ersin Yumer, Xiaohui Shen, Emiliano Gambaretto, Christian Gagné, and Jean-François Lalonde. Learning to predict indoor illumination from a single image. *arXiv preprint arXiv:1704.00090*, 2017.

[25] Chloe LeGendre, Wan-Chun Ma, Graham Fyffe, John Flynn, Laurent Charbonnel, Jay Busch, and Paul Debevec. Deeplight: Learning illumination for unconstrained mobile mixed reality. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5918–5928, 2019.

[26] Mathieu Garon, Kalyan Sunkavalli, Sunil Hadap, Nathan Carr, and Jean-François Lalonde. Fast spatially-varying indoor lighting estimation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6901–6910. IEEE, 2019.

[27] Shuran Song and Thomas Funkhouser. Neural illumination: Lighting prediction for indoor environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6918–6926, 2019.

[28] Zhaoxi Chen, Guangcong Wang, and Ziwei Liu. Text2light: Zero-shot text-driven hdr panorama generation. *arXiv preprint arXiv:2209.09898*, 2022.

[29] Tianhao Wu, Chuanxia Zheng, and Tat-Jen Cham. Panodiffusion: 360-degree panorama outpainting via diffusion. In *The Twelfth International Conference on Learning Representations*.

[30] Dian Zheng, Cheng Zhang, Xiao-Ming Wu, Cao Li, Chengfei Lv, Jian-Fang Hu, and Wei-Shi Zheng. Panorama generation from nfov image done right. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21610–21619, 2025.

[31] Cheng Zhang, Qianyi Wu, Camilo Cruz Gambardella, Xiaoshui Huang, Dinh Phung, Wanli Ouyang, and Jianfei Cai. Taming stable diffusion for text to 360 panorama image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6347–6357, 2024.

[32] Nikolai Kalischek, Michael Oechsle, Fabian Manhardt, Philipp Henzler, Konrad Schindler, and Federico Tombari. Cubediff: Repurposing diffusion-based image models for panorama generation. In *The Thirteenth International Conference on Learning Representations*, 2025.

[33] Shiyuan Shen, Zhongyun Bao, Wenju Xu, and Chunxia Xiao. Illumidiff: Indoor illumination estimation from a single image with diffusion model. *IEEE transactions on visualization and computer graphics*, 2025.

[34] Berthy T Feng, Jamie Smith, Michael Rubinstein, Huiwen Chang, Katherine L Bouman, and William T Freeman. Score-based diffusion models as principled priors for inverse imaging. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10520–10531, 2023.

[35] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

[36] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12868–12878. IEEE, 2021.

[37] Bingyan Liu, Chengyu Wang, Tingfeng Cao, Kui Jia, and Jun Huang. Towards understanding cross and self-attention in stable diffusion for text-guided image editing. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7817–7826. IEEE Computer Society, 2024.

[38] Quynh Phung, Songwei Ge, and Jia-Bin Huang. Grounded text-to-image synthesis with attention refocusing. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7932–7942. IEEE Computer Society, 2024.

[39] Omer Dahary, Or Patashnik, Kfir Aberman, and Daniel Cohen-Or. Be yourself: Bounded attention for multi-subject text-to-image generation. In *European Conference on Computer Vision*, pages 432–448. Springer, 2024.

[40] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[41] Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. Moge-2: Accurate monocular geometry with metric scale and sharp details. *arXiv preprint arXiv:2507.02546*, 2025.

[42] Haodong Li, Wangguangdong Zheng, Jing He, Yuhao Liu, Xin Lin, Xin Yang, Ying-Cong Chen, and Chunchao Guo. DA^2: Depth Anything in Any Direction. *arXiv preprint arXiv:2509.26618*, 2025.

[43] Junwei Zheng, Ruiping Liu, Yufan Chen, Kunyu Peng, Chengzhi Wu, Kailun Yang, Jiaming Zhang, and Rainer Stiefelhagen. Open panoramic segmentation. In *European Conference on Computer Vision*, pages 164–182. Springer, 2024.

[44] Zhigang Jiang, Zhongzheng Xiang, Jinhua Xu, and Ming Zhao. Lgt-net: Indoor panoramic room layout estimation with geometry-aware transformer network. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1644–1653. IEEE, 2022.

[45] Jianhua Sun and Cewu Lu. Digital gene: Learning about the physical world through analytic concepts. *arXiv preprint arXiv:2504.04170*, 2025.

[46] Tong Wu, Jiarui Zhang, Xiao Fu, Yuxin Wang, Jiawei Ren, Liang Pan, Wayne Wu, Lei Yang, Jiaqi Wang, Chen Qian, et al. Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 803–814, 2023.

[47] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11305–11315. IEEE, 2022.

A Additional Method Details

This section gives implementation and derivation details for the three algorithmic components of ReFlex: Hypothesis-to-Detail Latent Solving (H2D), Support-Calibrated Routing (SCR), and Trajectory-Anchored Denoising (TAD). The training losses are separated into Appendix B to keep the method derivations distinct from optimization details.

A.1 Hypothesis-to-Detail Latent Solving

A.1.1 Detail Codebook and Hypothesis Targets

ReFlex performs observation-constrained reconstruction in a discrete latent space. Let $\mathcal{D} = \{1, \dots, K\}$ denote the full detail codebook of the panorama tokenizer. We partition $\mathcal{D}$ into Q disjoint hypothesis groups:

$$\mathcal{D} = \bigcup_{q=1}^{Q} \mathcal{D}_q, \qquad \mathcal{D}_i \cap \mathcal{D}_j = \varnothing \text{ for } i \neq j, \qquad |\mathcal{D}_q| = m_q, \qquad \sum_{q=1}^{Q} m_q = K. \tag{17}$$

The reduced hypothesis vocabulary is $\mathcal{H} = \{1, \dots, Q\}$. For a detail code $d \in \mathcal{D}$, define its hypothesis assignment by

$$\psi(d) = q \quad \text{if} \quad d \in \mathcal{D}_q. \tag{18}$$

Given a ground-truth detail latent sequence $z^{D,\star} \in \mathcal{D}^N$, the corresponding hypothesis target is

$$z_n^{H,\star} = \psi(z_n^{D,\star}), \qquad n = 1, \dots, N. \tag{19}$$

A.1.2 Iterative Token Commitment

For stage $s \in \{H, D\}$, H2D predicts token posteriors and updates the latent sequence as

$$p_t^s = F_s(M^s, B^s, z_t^s), \qquad z_{t+1}^s = \Gamma(z_t^s, p_t^s). \tag{20}$$

Here M^s is the observation memory, B^s is the auxiliary stage context, and z_t^s is the current latent sequence. Specifically, $B^H = \varnothing$ for hypothesis recovery and $B^D = R^H$ for detail reconstruction, where R^H denotes hypothesis-stage hidden states. The confidence of token n at iteration t is

$$\omega_{t,n}^s = \max_k p_{t,n,k}^s. \tag{21}$$

Let Ω_t^s be the positions selected by the confidence schedule. The commitment operator is

$$z_{t+1,n}^s = \begin{cases} \arg\max_k p_{t,n,k}^s, & n \in \Omega_t^s, \\ z_{t,n}^s, & n \notin \Omega_t^s. \end{cases} \tag{22}$$

Thus, high-confidence predictions are committed while low-confidence positions remain revisable.

A.2 Support-Calibrated Routing

A.2.1 Observation-Support Relation Map and Edge Descriptor

SCR reconciles self-attention propagation with observation-support relations. Let $X^{\mathrm{sa}} \in \mathbb{R}^{N \times d}$ be token features after self-attention, $A^{\mathrm{ca}} \in \mathbb{R}^{N \times L}$ cross-attention from panorama tokens to observation tokens, and $A^{\mathrm{sa}} \in \mathbb{R}^{N \times N}$ self-attention among panorama tokens. The observation-support relation map and internal propagation relation map are

$$\tilde{A}^{\mathrm{ca}} = (A^{\mathrm{ca}})^{\gamma}, \qquad O = \tilde{A}^{\mathrm{ca}} (\tilde{A}^{\mathrm{ca}})^{\top}, \qquad P = A^{\mathrm{sa}}. \tag{23}$$

Here O_{ij} measures shared observation information between panorama tokens i and j, P_{ij} measures their internal propagation strength, and γ sharpens the cross-attention evidence.

For token pair (i, j), SCR builds the edge descriptor

$$\phi_{ij} = \Big[\ell(O_{ij}),\ \ell(\beta P_{ij}),\ \ell(O_{ij}) - \ell(\beta P_{ij}),\ \big|\ell(O_{ij}) - \ell(\beta P_{ij})\big|,\ u_i,\ u_j,\ u_i - u_j \Big], \tag{24}$$

where

$$\ell(x) = \log(x + \varepsilon_{\log}), \qquad u_i = W_u \mathrm{LN}(x_i^{\mathrm{sa}}). \tag{25}$$

Here $\beta > 0$ is a learnable scale, W_u is a learnable projection, and $\mathrm{LN}(\cdot)$ denotes layer normalization. A shared edge MLP predicts a signed relation score:

$$\Delta_{ij} = f_\theta(\phi_{ij}). \tag{26}$$

Positive and negative scores select additive and subtractive routing branches, respectively.

A.2.2 Signed Sparse Routing and Residual Update

The positive and negative routing matrices are constructed from row-wise sparse scores:

$$S_{ij}^{\pm} = \mathrm{TopK}\left(\mathrm{ReLU}(\pm\Delta_{ij})^{\eta}\right), \qquad R_{ij}^{\pm} = \frac{S_{ij}^{\pm}}{\sum_{j'} S_{ij'}^{\pm} + \varepsilon_R}. \tag{27}$$

Here $\eta \geq 1$, TopK keeps the strongest row-wise edges, and $\varepsilon_R > 0$ prevents division by zero. If a row has zero mass before normalization, the corresponding positive or negative message is defined to be the zero vector. Branch-specific confidence weights are

$$c_{ij}^{\pm} = \sigma\left(f_c^{\pm}(\phi_{ij})\right), \tag{28}$$

and the positive and negative messages are

$$m_i^{\pm} = \sum_j R_{ij}^{\pm} c_{ij}^{\pm} W_v^{\pm} x_j^{\mathrm{sa}}. \tag{29}$$

With token gates

$$g_i^{\pm} = \sigma(f_g^{\pm}(x_i^{\mathrm{sa}})), \tag{30}$$

SCR applies the signed residual correction

$$x_i' = x_i^{\mathrm{sa}} + g_i^{+} W_o^{+} m_i^{+} - g_i^{-} W_o^{-} m_i^{-}. \tag{31}$$

The positive branch enhances observation-supported propagation, while the negative branch subtractively suppresses unsupported propagation.

A.3 Trajectory-Anchored Denoising

A.3.1 Reliability-Motivated Calibration from Hierarchical Posteriors

Trajectory-Anchored Denoising (TAD) uses the hierarchical predictions of OCE to decide where a frozen diffusion model should be released. This section separates four components that play different roles: (i) an exact categorical identity under an oracle clean-code model; (ii) a prior-aware inference-time reliability surrogate; (iii) an empirical cross-stage agreement calibration gate; and (iv) a monotone release projection from reliability to diffusion reverse time. Only the first component is an exact identity. The remaining components are conservative calibration mechanisms used at inference time; they are not claimed to be unbiased estimators of a diffusion posterior or probabilistically equivalent to the Gaussian diffusion forward process.

Categorical reference priors. Let the detail codebook be $\mathcal{D} = \{1, \ldots, K\}$ and let it be partitioned into disjoint hypothesis groups

$$\mathcal{D} = \bigcup_{q=1}^{Q} \mathcal{D}_q, \qquad \mathcal{D}_i \cap \mathcal{D}_j = \varnothing \quad (i \neq j), \qquad |\mathcal{D}_q| = m_q, \qquad \sum_{q=1}^{Q} m_q = K. \tag{32}$$

The uniform codebook prior over detail codes is

$$u_K(d) = \frac{1}{K}, \qquad d \in \mathcal{D}. \tag{33}$$

It induces a generally non-uniform codebook prior over hypothesis groups,

$$\nu_q = \sum_{d \in \mathcal{D}_q} u_K(d) = \frac{m_q}{K}, \qquad \sum_{q=1}^{Q} \nu_q = 1. \tag{34}$$

For each group q, define the group-uniform distribution over detail codes:

$$s_q(d) = \begin{cases} \frac{1}{m_q}, & d \in \mathcal{D}_q, \\ 0, & d \notin \mathcal{D}_q. \end{cases} \tag{35}$$

For a clean detail code d^*, let e_{d^*} be the corresponding one-hot detail distribution and let $q^* = \psi(d^*)$ be its hypothesis group.

Exact oracle identity. We first state an oracle categorical identity. Assume a two-level symmetric categorical corruption model:

$$p = (1 - \lambda_H)u_K + \lambda_H(1 - \lambda_D)s_{q^*} + \lambda_H\lambda_D e_{d^*}, \tag{36}$$

where $\lambda_H \in [0, 1]$ measures retention of the correct hypothesis group and $\lambda_D \in [0, 1]$ measures retention of the correct detail code conditioned on that group. Subtracting u_K gives

$$p - u_K = \lambda_H(s_{q^*} - u_K) + \lambda_H\lambda_D(e_{d^*} - s_{q^*}). \tag{37}$$

Claim 1: Orthogonality. Let

$$a_{q^*} = s_{q^*} - u_K, \qquad b_{d^*} = e_{d^*} - s_{q^*}. \tag{38}$$

Then

$$\langle a_{q^*}, b_{d^*} \rangle = 0. \tag{39}$$

Proof. Inside $\mathcal{D}_{q^*}$, a_{q^*} is constant:

$$a_{q^*}(d) = \frac{1}{m_{q^*}} - \frac{1}{K}, \qquad d \in \mathcal{D}_{q^*}. \tag{40}$$

The vector b_{d^*} has zero sum inside $\mathcal{D}_{q^*}$:

$$\sum_{d \in \mathcal{D}_{q^*}} b_{d^*}(d) = \left(1 - \frac{1}{m_{q^*}}\right) + (m_{q^*} - 1)\left(-\frac{1}{m_{q^*}}\right) = 0. \tag{41}$$

Outside $\mathcal{D}_{q^*}$, $b_{d^*}(d) = 0$. Thus,

$$\langle a_{q^*}, b_{d^*} \rangle = \left(\frac{1}{m_{q^*}} - \frac{1}{K}\right) \sum_{d \in \mathcal{D}_{q^*}} b_{d^*}(d) = 0. \tag{42}$$

□

Claim 2: Component norms. The hypothesis and detail components satisfy

$$\|s_{q^*} - u_K\|_2^2 = \frac{1}{m_{q^*}} - \frac{1}{K} = \frac{K - m_{q^*}}{K m_{q^*}}, \tag{43}$$

$$\|e_{d^*} - s_{q^*}\|_2^2 = 1 - \frac{1}{m_{q^*}} = \frac{m_{q^*} - 1}{m_{q^*}}. \tag{44}$$

Proof. For the hypothesis component,

$$\|s_{q^*} - u_K\|_2^2 = m_{q^*}\left(\frac{1}{m_{q^*}} - \frac{1}{K}\right)^2 + (K - m_{q^*})\left(\frac{1}{K}\right)^2 \tag{45}$$

$$= \frac{1}{m_{q^*}} - \frac{1}{K}. \tag{46}$$

For the detail component, inside $\mathcal{D}_{q^*}$ one entry equals $1 - \frac{1}{m_{q^*}}$ and the remaining $m_{q^*} - 1$ entries equal $-\frac{1}{m_{q^*}}$. Therefore,

$$\|e_{d^*} - s_{q^*}\|_2^2 = \left(1 - \frac{1}{m_{q^*}}\right)^2 + (m_{q^*} - 1)\left(\frac{1}{m_{q^*}}\right)^2 \tag{47}$$

$$= 1 - \frac{1}{m_{q^*}}. \tag{48}$$

□

Claim 3: Oracle categorical signal identity. Define the normalized categorical signal coefficient

$$\bar{\alpha}^{\mathrm{cat}} = \frac{K}{K - 1}\|p - u_K\|_2^2. \tag{49}$$

Under Eq. (36),

$$\bar{\alpha}^{\mathrm{cat}} = \lambda_H^2 A_{q^*} + \lambda_H^2 \lambda_D^2 B_{q^*}, \tag{50}$$

where

$$A_q = \frac{K - m_q}{(K-1)m_q}, \qquad B_q = \frac{K(m_q - 1)}{(K-1)m_q} = 1 - A_q. \tag{51}$$

Proof. Using Eq. (37),

$$p - u_K = \lambda_H a_{q^*} + \lambda_H \lambda_D b_{d^*}. \tag{52}$$

By Claim 1,

$$\|p - u_K\|_2^2 = \lambda_H^2 \|a_{q^*}\|_2^2 + \lambda_H^2 \lambda_D^2 \|b_{d^*}\|_2^2. \tag{53}$$

Substituting Eqs. (43) and (44), then multiplying by $K/(K-1)$, gives Eq. (50). □

This identity is exact only in the oracle setting where the clean group q^* and clean detail code d^* exist under the symmetric categorical model. At inference time, these clean variables are unknown, and OCE provides only hierarchical predictive distributions. We therefore use the identity as motivation for a prior-aware reliability surrogate, not as an exact posterior estimator.

Inference-time hierarchical posteriors. At token n, the hypothesis and detail stages produce categorical predictive distributions

$$\pi_n^H = \mathrm{softmax}\big(\mathrm{Head}_H(R_n^H)\big), \qquad \pi_n^D = \mathrm{softmax}\big(\mathrm{Head}_D(R_n^D)\big), \tag{54}$$

where Head_H and Head_D map hidden states to logits over $\mathcal{H}$ and $\mathcal{D}$, respectively. We do not introduce additional temperature calibration in the reliability estimator. This keeps the reliability score tied to the predictive distributions produced by the trained OCE model and avoids adding post-hoc calibration hyperparameters to the prior-release rule.

The detail posterior is aggregated into the hypothesis space:

$$\mu_{n,q} = \sum_{d \in \mathcal{D}_q} \pi_{n,d}^D, \qquad q = 1, \ldots, Q. \tag{55}$$

We then fuse the direct hypothesis posterior and the aggregated detail posterior using a symmetric product-of-experts form:

$$\eta_{n,q} = \sqrt{(\pi_{n,q}^H + \varepsilon)(\mu_{n,q} + \varepsilon)}, \qquad w_{n,q} = \frac{\eta_{n,q}}{\sum_j \eta_{n,j}}. \tag{56}$$

This symmetric fusion is equivalent to setting the fusion exponent to $1/2$ for both stages. It assigns high belief only to groups supported by both the hypothesis prediction and the detail prediction aggregated to the hypothesis space. Thus, disagreement between the two stages is suppressed before computing reliability.

The fused belief w_n is not used as a clean-group posterior directly, because larger hypothesis groups can receive non-negligible mass simply by containing more detail codes. We therefore center w_n against a reference hypothesis prior. By default, we use the codebook-size prior

$$\nu_q = \frac{m_q}{K}, \qquad q = 1, \ldots, Q, \tag{57}$$

which corresponds to a uniform prior over detail codes. If empirical code usage is available, ν_q can instead be set to the training-set frequency of group q without changing the derivation below. The following prior-aware responsibility uses this reference prior before forming an inference-time reliability surrogate.

Prior-aware hypothesis responsibility. For each hypothesis group, define

$$D_q = \|e_q - \nu\|_2^2 = 1 - 2\nu_q + \sum_{j=1}^{Q} \nu_j^2. \tag{58}$$

We compute group-prior-centered positive excess scores

$$b_{n,q} = \frac{[w_{n,q} - \nu_q]_+}{\sqrt{D_q} + \varepsilon}, \qquad [x]_+ = \max(x, 0), \tag{59}$$

and let $Z_n = \sum_j b_{n,j}$. When $Z_n > \varepsilon_\rho$, we normalize them into a soft clean-group responsibility:

$$\rho_{n,q} = \frac{b_{n,q}}{Z_n}. \tag{60}$$

When $Z_n \le \epsilon_\rho$, the total positive prior-centered excess does not exceed the numerical threshold; in this case, we set the subsequent reliability contribution of token n to zero. This group-prior-centered responsibility replaces the hard $\arg\max_q w_{n,q}$ rule. It explicitly accounts for non-uniform reference prior mass and avoids treating large groups as more reliable merely because they contain more detail codes.

Under the prior-aware one-spike coarse model with $\lambda_H > 0$

$$w_n = (1 - \lambda_H)\nu + \lambda_H e_{q^*}, \tag{61}$$

we have

$$w_{n,q^*} - \nu_{q^*} = \lambda_H(1 - \nu_{q^*}) > 0, \qquad w_{n,q} - \nu_q = -\lambda_H \nu_q < 0 \quad (q \neq q^*). \tag{62}$$

Therefore, when $Z_n > \epsilon_\rho$, $\rho_{n,q^*} = 1$ and $\rho_{n,q} = 0$ for $q \neq q^*$ in the noiseless ideal case.

A hard group-prior-centered selector

$$\hat{q}_n = \arg\max_q (w_{n,q} - \nu_q) \tag{63}$$

recovers q^* under bounded perturbation. Specifically, if

$$w_n = (1 - \lambda_H)\nu + \lambda_H e_{q^*} + \epsilon_n, \qquad \|\epsilon_n\|_\infty \le \delta, \tag{64}$$

then the pairwise centered margin between q^* and any non-target group q is

$$\lambda_H(1 - \nu_{q^*} + \nu_q). \tag{65}$$

Thus, a sufficient recovery condition is

$$\lambda_H \min_{q \neq q^*}(1 - \nu_{q^*} + \nu_q) > 2\delta. \tag{66}$$

For tokens with $Z_n > \epsilon_\rho$, the true group receives the largest soft responsibility in Eq. (60) when

$$\frac{\lambda_H(1 - \nu_{q^*}) - \delta}{\sqrt{D_{q^*}} + \varepsilon} > \max_{q \neq q^*} \frac{[-\lambda_H \nu_q + \delta]_+}{\sqrt{D_q} + \varepsilon}. \tag{67}$$

The positive-part term appears because non-target groups have negative centered scores in the noiseless group-prior-centered model. They contribute to the soft responsibility only when perturbations are large enough to overcome their group-prior-centered suppression term, i.e., when $\delta > \lambda_H \nu_q$. When $\delta \le \lambda_H \nu_q$, that non-target group contributes zero positive mass.

Hypothesis retention surrogate. Using the group-prior-centered responsibility, we estimate the squared hypothesis retention by the method-of-moments surrogate

$$\hat{\lambda}^2_{H,n} = \begin{cases} \mathrm{clip}\left(\dfrac{\|w_n - \nu\|_2^2}{\sum_{q=1}^{Q} \rho_{n,q} D_q + \varepsilon}, 0, 1 \right), & Z_n > \varepsilon_\rho, \\ 0, & Z_n \le \varepsilon_\rho. \end{cases} \tag{68}$$

Under the noiseless one-spike model in Eq. (61), if $Z_n > \epsilon_\rho$, then $\rho_{n,q^*} = 1$ and

$$\|w_n - \nu\|_2^2 = \lambda_H^2 D_{q^*}. \tag{69}$$

Therefore, when numerical stabilization and thresholding are omitted, Eq. (68) recovers λ_H^2 exactly in the ideal coarse model. Outside this model, it should be interpreted as a prior-aware reliability surrogate.

To suppress flat or multi-modal coarse beliefs, we optionally use an ambiguity damping factor

$$\chi_n = \begin{cases} 1 - \dfrac{H(\rho_n)}{\log Q}, & Z_n > \varepsilon_\rho, \\ 0, & Z_n \le \varepsilon_\rho, \end{cases} \qquad H(\rho_n) = -\sum_{q=1}^{Q} \rho_{n,q} \log(\rho_{n,q} + \varepsilon). \tag{70}$$

When $Z_n > \epsilon_\rho$, ρ_n is a valid probability distribution and $H(\rho_n)$ is a numerically stabilized approximation to categorical entropy. When $Z_n \le \varepsilon_\rho$, the token has insufficient positive prior-centered evidence under the numerical threshold and is assigned zero reliability contribution. This term is not part of the exact oracle identity. It is a conservative inference-time safeguard: when responsibility is diffuse or absent, the token is treated as less reliable.

Within-group detail retention surrogate. For each group q, compute a numerically stabilized approximation to the conditional detail distribution

$$r_{n,d|q} = \frac{\pi^D_{n,d}}{\mu_{n,q} + \varepsilon}, \qquad d \in \mathcal{D}_q. \tag{71}$$

For $m_q > 1$, define

$$\hat{\lambda}^2_{D,n,q} = \mathrm{clip}\left(\frac{m_q}{m_q - 1} \sum_{d \in \mathcal{D}_q} \left(r_{n,d|q} - \frac{1}{m_q} \right)^2, 0, 1 \right). \tag{72}$$

For $m_q = 1$, the coefficient B_q in Eq. (51) is zero, so the value of $\hat{\lambda}^2_{D,n,q}$ does not affect the final categorical score; in implementation we set it to one.

The estimator in Eq. (72) is a method-of-moments plug-in estimator, not a maximum-likelihood estimator. Indeed, if the group-conditional posterior follows a symmetric mixture

$$r_{\cdot|q} = (1 - \lambda_D) u_q + \lambda_D e_{d^*}, \qquad u_q(d) = \frac{1}{m_q},\ d \in \mathcal{D}_q, \tag{73}$$

then

$$r_{\cdot|q} - u_q = \lambda_D (e_{d^*} - u_q), \tag{74}$$

and therefore

$$\| r_{\cdot|q} - u_q \|_2^2 = \lambda_D^2 \left(1 - \frac{1}{m_q} \right). \tag{75}$$

Solving for λ_D^2 yields Eq. (72).

Raw categorical reliability surrogate. Combining the oracle identity with inference-time responsibilities gives a posterior-averaged surrogate:

$$\hat{\tilde{\alpha}}^{\mathrm{cat}}_{n,\mathrm{raw}} = \mathrm{clip}\left(\hat{\lambda}^2_{H,n} \sum_{q=1}^{Q} \rho_{n,q} \left(A_q + B_q \hat{\lambda}^2_{D,n,q} \right), 0, 1 \right). \tag{76}$$

This is not an exact identity because q^* and d^* are unknown at inference time. It is a prior-aware reliability surrogate motivated by Eq. (50). With ambiguity damping, we use

$$\hat{\tilde{\alpha}}^{\mathrm{cat}}_{n,\mathrm{amb}} = \chi_n \hat{\tilde{\alpha}}^{\mathrm{cat}}_{n,\mathrm{raw}}. \tag{77}$$

Cross-stage agreement gate. The categorical surrogate above measures group-prior-centered concentration and within-group detail sharpness. However, a token may still be unreliable when the hypothesis and detail stages concentrate on different hypothesis groups. We therefore apply a conservative cross-stage agreement gate.

This gate is not part of the oracle categorical identity in Eq. (50). It is an inference-time safeguard for prior release: when the two stages disagree, the location is treated as less reliable even if one stage is individually sharp. Since Eq. (54) uses the native predictive distributions without temperature calibration, the agreement can be computed directly in the shared hypothesis space:

$$\kappa_n = \mathrm{clip}\left(\frac{\sum_{q=1}^{Q} \pi^H_{n,q} \mu_{n,q} - \sum_{q=1}^{Q} \nu_q^2}{1 - \sum_{q=1}^{Q} \nu_q^2 + \varepsilon}, 0, 1 \right). \tag{78}$$

The subtraction of $\sum_q \nu_q^2$ removes the agreement expected from the reference hypothesis prior alone. If the two stages agree only at the level of this reference prior, then $\sum_q \pi^H_{n,q} \mu_{n,q} \approx \sum_q \nu_q^2$ and $\kappa_n \approx 0$. If both stages concentrate on the same hypothesis group, then $\sum_q \pi^H_{n,q} \mu_{n,q} \approx 1$

and $\kappa_n \approx 1$. Thus, κ_n acts as a prior-normalized agreement gate rather than as an independently calibrated probability.

The final reliability score is

$$\hat{a}_n = \mathrm{clip}\left(\kappa_n \hat{\bar{\alpha}}^{\mathrm{cat}}_{n,\mathrm{amb}}, 0, 1\right). \tag{79}$$

Equivalently, Eq. (79) instantiates the reliability function $\Phi(\pi_n^H, \pi_n^D, \{\mathcal{D}_q\}_{q=1}^Q)$ used by RGR in the main paper. The score $\hat{a}_n$ should be interpreted as a conservative reliability estimate for prior release, not as an exact posterior probability. It is designed to be high only when the fused prediction is group-prior-centered, detail-sharp, and cross-stage consistent.

A.3.2 Monotone Release Projection

TAD uses the final reliability map to determine how long each latent location should remain anchored to the OCE reconstruction. This projection is monotone: locations with higher reliability are anchored for no fewer reverse steps. It is not a probabilistic equivalence between categorical purity and Gaussian diffusion signal-to-noise ratio.

Let $\mathcal{T}$ be the OCE token grid and $\mathcal{L}$ be the diffusion latent grid. We first clamp the reliability to avoid numerical divergence:

$$\tilde{a}_n = \mathrm{clip}(\hat{a}_n, \varepsilon_a, 1 - \varepsilon_a), \qquad 0 < \varepsilon_a < \frac{1}{2}. \tag{80}$$

We then form a logit reliability map and resample it to the diffusion latent grid:

$$g_\ell = \mathcal{R}_{\mathcal{T}\to\mathcal{L}}\left(\mathrm{logit}(\tilde{a})\right)_\ell. \tag{81}$$

Let $\{\bar{\alpha}_{\tau_s}\}_{s=1}^T$ be the noise schedule used by the frozen diffusion model, ordered such that $\bar{\alpha}_{\tau_s}$ decreases with the forward noising index s. We match each latent location to the closest schedule point in logit space:

$$\hat{s}_\ell = \arg\min_{s\in[T]} \left|\mathrm{logit}(\bar{\alpha}_{\tau_s}) - g_\ell\right|. \tag{82}$$

Because $\mathrm{logit}(\cdot)$ is strictly increasing on $(0,1)$ and $\bar{\alpha}_{\tau_s}$ is monotone in s, the mapping from the resampled logit reliability g_ℓ to the matched index $\hat{s}_\ell$ is monotone up to schedule discretization ties. That is, after resampling to the diffusion latent grid, if $g_i \geq g_j$, then the matched forward noising indices satisfy $\hat{s}_i \leq \hat{s}_j$ up to one schedule bin. Thus, higher-reliability latent locations are assigned no shorter reverse-time anchoring durations after projecting $\hat{s}_\ell$ to r_ℓ.

Anchoring duration. We use a fully anchored reliability-preserving policy by default:

$$r_\ell = T - \hat{s}_\ell + 1. \tag{83}$$

Under the uncapped policy, maximally reliable locations remain anchored throughout the entire reverse trajectory. Less reliable locations receive shorter anchoring durations and are therefore exposed earlier to the diffusion model. With the practical cap r_{ub}, the effective duration is saturated at $\min(r_\ell, r_{\mathrm{ub}})$. This keeps high-confidence regions close to the OCE trajectory for longer while still allowing the prior to refine weakly constrained regions.

A.3.3 Trajectory-Anchored Reverse Update

Let Y' be the OCE reconstruction and let

$$x^{\mathrm{anc}} = E_{\mathrm{vae}}(Y') \tag{84}$$

be its latent representation. We sample a fixed noise

$$\zeta \sim \mathcal{N}(0, I) \tag{85}$$

and define the anchor trajectory

$$x_t^{\mathrm{anc}} = \sqrt{\bar{\alpha}_t}\, x^{\mathrm{anc}} + \sqrt{1 - \bar{\alpha}_t}\, \zeta. \tag{86}$$

At reverse step $t \to t-1$, let x_{t-1}^{free} denote the ordinary reverse update of the LoRA-adapted diffusion model with frozen base weights, and let x_{t-1}^{anc} denote the corresponding anchor state from Eq. (86). Let

$$s(t) = T - t + 1 \tag{87}$$

be the number of executed reverse steps, and let r_{ub} cap the maximum anchoring duration. The spatial anchoring mask is

$$m_\ell^{(t)} = \mathbf{1}\left[s(t) \le \min(r_\ell, r_{\mathrm{ub}})\right], \tag{88}$$

broadcast over latent channels. The reverse update is

$$x_{t-1,\ell} = m_\ell^{(t)} x_{t-1,\ell}^{\mathrm{anc}} + \left(1 - m_\ell^{(t)}\right) x_{t-1,\ell}^{\mathrm{free}}. \tag{89}$$

Therefore, a location with larger r_ℓ remains anchored for more reverse steps, while a location with smaller r_ℓ is anchored for fewer reverse steps and is exposed earlier to the learned prior. After all reverse steps, the final panorama is decoded as

$$Y'' = D_{\mathrm{vae}}(x_0). \tag{90}$$

Interpretation. The role of TAD is not to convert a categorical posterior into a Gaussian posterior. Rather, it uses a prior-aware, cross-stage calibrated reliability score to impose a monotone release schedule: observation-reliable regions stay close to the OCE trajectory for longer, while unreliable regions are exposed earlier to the frozen diffusion model. This implements the intended division of labor between observation-grounded reconstruction and generative completion.

B Training Objectives and Implementation Details

This section summarizes the losses used for observation-constrained latent solving and reliability-guided prior adaptation.

Observation-Consistent Estimator is trained with standard masked-token cross-entropy losses in the hypothesis and detail spaces. During training, we sample masked token sets $\mathcal{M}_H$ and $\mathcal{M}_D$ for the hypothesis and detail stages, respectively. The model predicts the masked targets in a single forward pass, while iterative token commitment is used only for MaskGIT-style inference.

For the hypothesis stage, let p_n^H denote the predicted categorical distribution over the hypothesis vocabulary $\mathcal{H}$ at token n. The hypothesis loss is

$$\mathcal{L}_H = \frac{1}{|\mathcal{M}_H|} \sum_{n \in \mathcal{M}_H} \mathrm{CE}\left(p_n^H, z_n^{H,\star}\right). \tag{91}$$

For the detail stage, let p_n^D denote the predicted categorical distribution over the detail codebook $\mathcal{D}$ at token n. The detail loss is

$$\mathcal{L}_D = \frac{1}{|\mathcal{M}_D|} \sum_{n \in \mathcal{M}_D} \mathrm{CE}\left(p_n^D, z_n^{D,\star}\right). \tag{92}$$

The reconstruction solver is optimized by

$$\mathcal{L}_{\mathrm{OCE}} = \omega_H \mathcal{L}_H + \omega_D \mathcal{L}_D. \tag{93}$$

SCR is trained end-to-end through the same masked-token objectives.

For RGR, the diffusion backbone is frozen and only LoRA parameters are updated. Given a clean latent x_0, timestep t, noise $\epsilon \sim \mathcal{N}(0, I)$, and

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, \tag{94}$$

the LoRA-adapted denoiser is trained with

$$\mathcal{L}_{\mathrm{TAD}} = \mathbb{E}_{x_0,t,\epsilon}\left[\left\|\epsilon - \epsilon_{\theta_0 + \Delta\theta_{\mathrm{LoRA}}}(x_t, t)\right\|_2^2\right], \tag{95}$$

where θ_0 denotes frozen base-prior weights, and $\Delta\theta_{\mathrm{LoRA}}$ denotes trainable LoRA parameters. At inference time, the adapted prior is used only through the trajectory-anchored update in Eq. (89).

C Baseline Configuration and Comparison Protocol

A fully homogeneous head-to-head comparison is not possible in our setting because the candidate baselines differ in both methodological principle and test-time information requirements. We therefore enforce fairness along three dimensions that can be standardized across methods: *benchmark*, *output space*, and *evaluation protocol*. All methods are trained or adapted on the same benchmark, evaluated on the same scene-level test split, and compared in the same pose-aligned equirectangular panorama space at the same spatial resolution. The main paper groups methods into *Prior-based Methods*, *Conditional Generation Methods*, and *Reconstruction*, because these families differ fundamentally in what they assume and how they use auxiliary information.

Unified benchmark and split. All compared methods are trained, fine-tuned, or otherwise adapted on the same benchmark described in the main paper, where each sample pairs a reflective RGB image with a pose-aligned equirectangular panorama. The benchmark is split at the HDR-scene level to avoid scene leakage between training and testing. Unless otherwise stated, all methods use the same training split and are evaluated on the same held-out test split. This standardization ensures that performance differences are not caused by access to different environments, object instances, or test scenes.

Unified output and evaluation space. To make image-level metrics directly comparable, all methods are evaluated in the same pose-aligned equirectangular panorama space at resolution 512×256. When a baseline natively predicts a different representation, we convert its output to this common target space before metric computation. For downstream evaluation, DA^2 is evaluated on the same pose-aligned test split used for image-level recovery, since it can process non-canonical equirectangular panoramas. OOOPS and LGTNet assume canonical gravity-aligned inputs, so we additionally evaluate them on gravity-aligned test samples, with the same alignment applied to both predictions and ground truth. All downstream metrics are computed on the full panorama. In this sense, fairness is enforced at the level of the final output representation and evaluation protocol, even when native model parameterizations differ.

Method-specific adaptation. Although the benchmark and evaluation space are unified, several baselines require task-specific adaptation because their native formulations do not directly match pose-aligned panorama recovery from a single RGB image with reflections.

Reconstruction. ReFlex operates directly from a single RGB image with reflections and does not require privileged information at inference time. It therefore defines the target problem setting studied in the paper.

Conditional Generation Methods. LDMs Outpainting and LuxDiT both rely on generative priors, but differ in how they are conditioned. LDMs Outpainting is not natively formulated for reflection-guided panorama recovery, so we provide FoV and mask information and construct a partially observed equirectangular canvas as its conditioning input. This gives LDMs Outpainting access to explicit geometric visibility information that is unavailable to ReFlex in the reconstruction setting. LuxDiT, in contrast, accepts RGB input directly and requires no privileged information, but its raw panorama prediction is centered opposite to the input pinhole viewing direction. We therefore circularly shift the raw prediction by half the horizontal extent so that it matches the pose convention used in our benchmark. In addition, LuxDiT produces both an LDR structure branch and a log-intensity branch; since our target is an LDR panorama, we supervise and fine-tune only the LDR structure branch.

Prior-based methods. DeLightNet, Georgoulis et al., and DRMNet belong to a reflector-centered family of methods that rely on reflectance-map-style inputs rather than a raw RGB image with reflections alone. In our benchmark, the required oracle signals are derived from renderer-side ground truth. For DeLightNet and Georgoulis et al., RandBalls-Oracle provides renderer-side normals, and masks. For DRMNet, RandBalls-Oracle provides renderer-side normals and masks. In the full ReFlex-Bench evaluation, oracle normals are not used; normal-dependent baselines instead use MoGe-2 estimated normals, while masks and field-of-view inputs are provided only when required by the corresponding baseline protocol. Because these methods operate in a native reflector-centered parameterization rather than directly in a pose-aligned equirectangular panorama space, their predictions are projected into the same equirectangular coordinate system before evaluation. For models whose original outputs are HDR-oriented, we adapt the output supervision as needed so that all methods are compared in the same LDR panorama space.

Training protocol and convergence. For open-source baselines, we retain the training recipe provided by the original implementation whenever possible and adapt only those components necessary to match our benchmark output space. For reimplemented baselines, we use the same optimizer family and learning-rate scale as our own training setup for consistency. Lightweight or earlier baselines are trained from scratch on our benchmark, whereas larger pretrained models such as LuxDiT are adapted using the fine-tuning strategy most consistent with their original synthetic-data setup; in our case, this is implemented with a LoRA-based adaptation. To avoid favoring one method through an arbitrary update budget, we determine convergence by early stopping on the validation loss: if the validation loss does not improve for three consecutive evaluations, training is terminated and the best checkpoint is retained.

D Evaluation Metrics for Downstream Tasks

To evaluate whether the generated panoramas preserve information that is useful for downstream perception, we assess their performance under several task-specific metrics. Given a generated panorama and its corresponding reference panorama, both are processed by the same downstream estimator. The resulting task outputs are then compared in the corresponding prediction space. This evaluation protocol focuses on downstream consistency rather than direct low-level image similarity. We consider depth or distance estimation, surface normal estimation, room layout estimation, and semantic segmentation consistency.

D.1 Depth / Distance Estimation

Let d_i denote the reference distance at pixel i, and let $\hat{d}_i$ denote the predicted distance obtained from the generated panorama. The set of valid evaluation pixels is denoted by Ω, with $N = |\Omega|$. Since monocular geometric predictors may produce distances up to an unknown global scale, the predicted distance map is first aligned to the reference by a positive scale-only least-squares factor:

$$s^\star = \arg\min_{s>0} \sum_{i\in\Omega} \left(s\hat{d}_i - d_i\right)^2. \tag{96}$$

All distance metrics are then computed using the aligned prediction $s^\star \hat{d}_i$.

Absolute Relative Error. The absolute relative error measures the average relative discrepancy between the aligned prediction and the reference:

$$\mathrm{AbsRel} = \frac{1}{N} \sum_{i\in\Omega} \frac{\left|s^\star \hat{d}_i - d_i\right|}{\max(d_i, \varepsilon_d)}. \tag{97}$$

Here $\varepsilon_d > 0$ avoids division by zero at invalid or near-zero reference distances; pixels with invalid distances are excluded from Ω. This metric normalizes the absolute error by the reference distance and is therefore sensitive to proportional errors. Lower values indicate better distance consistency.

Root Mean Squared Error. The root mean squared error measures the Euclidean discrepancy between the aligned prediction and the reference:

$$\mathrm{sRMSE} = \sqrt{\frac{1}{N} \sum_{i\in\Omega} \left(s^\star \hat{d}_i - d_i\right)^2}. \tag{98}$$

Compared with the absolute relative error, RMSE penalizes large local deviations more strongly due to the squared residual. Lower values indicate more accurate geometric reconstruction.

Scale Factor. The estimated scale factor $s^\star$ can also be summarized across samples. This quantity is not itself an accuracy metric; instead, it reflects the global multiplicative correction required to align the predicted distances with the reference distances.

D.2 Surface Normal Estimation

Let $\mathbf{n}_i \in \mathbb{R}^3$ be the reference unit surface normal at pixel i, and let $\hat{\mathbf{n}}_i \in \mathbb{R}^3$ be the corresponding predicted unit normal. The per-pixel angular error is computed as

$$\theta_i = \arccos\left(\mathrm{clip}\left(\hat{\mathbf{n}}_i^\top \mathbf{n}_i, -1, 1\right)\right) \cdot \frac{180}{\pi}, \tag{99}$$

where θ_i is measured in degrees. The clipping operation avoids numerical instability when the dot product is slightly outside the valid range $[-1, 1]$.

Mean Angular Error. The mean angular error is defined as

$$\text{MeanAngularError} = \frac{1}{N} \sum_{i \in \Omega} \theta_i. \tag{100}$$

It measures the average directional discrepancy between predicted and reference normals. Lower values indicate better surface orientation consistency.

Median Angular Error. The median angular error is defined as

$$\text{MedianAngularError} = \text{median}_{i \in \Omega} \left(\theta_i \right). \tag{101}$$

Compared with the mean angular error, the median angular error is more robust to outliers and better reflects the typical per-pixel normal estimation quality.

Threshold Accuracy. Normal accuracy under an angular threshold τ is computed as

$$\text{Acc}_\tau = \frac{1}{N} \sum_{i \in \Omega} \mathbf{1} \left(\theta_i < \tau \right), \tag{102}$$

where $\mathbf{1}(\cdot)$ is the indicator function. We report accuracies at $\tau \in \{11.25^\circ, 22.5^\circ, 30^\circ\}$. Higher values indicate that a larger fraction of predicted normals fall within the specified angular tolerance.

D.3 Room Layout Estimation

For room layout estimation, the generated and reference panoramas are passed through the same layout estimator. Let $\mathcal{L}$ denote the reference layout and $\hat{\mathcal{L}}$ denote the layout inferred from the generated panorama. The predicted and reference layouts are compared using both volumetric and image-space geometric criteria.

3D Intersection over Union. Let V and $\widehat{V}$ denote the enclosed room volumes represented by L and $\widehat{L}$ in a common coordinate system. The 3D Intersection over Union measures their volumetric overlap:

$$\text{3DIoU} = \frac{\text{Vol}(\widehat{V} \cap V)}{\text{Vol}(\widehat{V} \cup V)}. \tag{103}$$

Higher 3DIoU indicates stronger agreement between the room volumes inferred from the reconstructed and reference panoramas.

2D Intersection over Union. Let P and $\widehat{P}$ denote the floor-plan polygons associated with L and $\widehat{L}$ on the ground plane. The 2D Intersection over Union is

$$\text{2DIoU} = \frac{\text{Area}(\widehat{P} \cap P)}{\text{Area}(\widehat{P} \cup P)}. \tag{104}$$

This metric measures agreement between the inferred floor-plan shapes in the top-down plane.

Corner Error. Let $\{(\mathbf{c}_k, \widehat{\mathbf{c}}_k)\}_{k=1}^{K_c}$ denote the corresponding layout-corner pairs established by the evaluator. The corner coordinates are expressed in pixels on a common equirectangular evaluation grid of width W_e and height H_e. We normalize the mean Euclidean corner distance by the image diagonal and report it as a percentage:

$$\text{CE} = \frac{100}{K_c \sqrt{W_\text{e}^2 + H_\text{e}^2}} \sum_{k=1}^{K_c} \left\| \widehat{\mathbf{c}}_k - \mathbf{c}_k \right\|_2. \tag{105}$$

Lower CE indicates closer agreement between the corresponding layout corners in the panorama image plane.

Layout RMSE. We convert L and $\widehat{L}$ into layout-induced depth maps using the same geometric conversion and camera-height convention. Let d_i^{lay} and $\widehat{d}_i^{\mathrm{lay}}$ denote the resulting reference-induced and reconstructed-image-induced depths at pixel i. Let Ω_{lay} be the evaluation pixel set specified by the evaluator, with $N_{\mathrm{lay}} = |\Omega_{\mathrm{lay}}|$. The layout depth RMSE is

$$\mathrm{RMSE} = \sqrt{\frac{1}{N_{\mathrm{lay}}} \sum_{i \in \Omega_{\mathrm{lay}}} \left(\widehat{d}_i^{\mathrm{lay}} - d_i^{\mathrm{lay}}\right)^2}. \tag{106}$$

Lower RMSE indicates stronger agreement between the depth maps induced by the two inferred room layouts.

Threshold Accuracy. Using the same layout-induced depth maps and evaluation pixels, we compute the fraction of locations whose multiplicative depth discrepancy is below 1.25:

$$\delta_1 = \frac{1}{N_{\mathrm{lay}}} \sum_{i \in \Omega_{\mathrm{lay}}} \mathbf{1}\left[\max\left(\frac{\widehat{d}_i^{\mathrm{lay}} + \epsilon_{\mathrm{lay}}}{d_i^{\mathrm{lay}} + \epsilon_{\mathrm{lay}}}, \frac{d_i^{\mathrm{lay}} + \epsilon_{\mathrm{lay}}}{\widehat{d}_i^{\mathrm{lay}} + \epsilon_{\mathrm{lay}}}\right) < 1.25\right]. \tag{107}$$

Here $\epsilon_{\mathrm{lay}} > 0$ is a small numerical-stability constant. Higher δ_1 indicates that a larger fraction of the layout-induced depths agree within the specified tolerance.

D.4 Semantic Segmentation Consistency

For semantic segmentation, the evaluation is conducted as a proxy downstream consistency test. A fixed semantic segmentation model is applied to both the generated and reference panoramas, and the resulting semantic predictions are compared. Therefore, the reported metrics measure semantic consistency under the downstream segmentation model, rather than absolute segmentation accuracy against human annotations.

Let C denote the number of semantic classes. For each class c, let TP_c, FP_c, and FN_c denote the numbers of true-positive, false-positive, and false-negative pixels, respectively, when comparing the segmentation induced by the generated panorama with the segmentation induced by the reference panorama. Classes with $TP_c + FP_c + FN_c = 0$ are absent from both predictions and are ignored in the class average. Let $\mathcal{C}_{\mathrm{valid}}$ denote the remaining valid class set.

Mean Intersection over Union. The Intersection over Union for class c is defined as

$$\mathrm{IoU}_c = \frac{TP_c}{TP_c + FP_c + FN_c}. \tag{108}$$

The mean Intersection over Union is then computed by averaging over all semantic classes:

$$\mathrm{mIoU} = \frac{1}{|\mathcal{C}_{\mathrm{valid}}|} \sum_{c \in \mathcal{C}_{\mathrm{valid}}} \mathrm{IoU}_c. \tag{109}$$

A higher mIoU indicates stronger agreement between the generated-induced and reference-induced semantic regions.

Dice Coefficient. The Dice coefficient for class c is defined as

$$\mathrm{Dice}_c = \frac{2TP_c}{2TP_c + FP_c + FN_c}. \tag{110}$$

The final Dice score is obtained by averaging over all classes:

$$\mathrm{Dice} = \frac{1}{|\mathcal{C}_{\mathrm{valid}}|} \sum_{c \in \mathcal{C}_{\mathrm{valid}}} \mathrm{Dice}_c. \tag{111}$$

The Dice coefficient emphasizes region overlap and is closely related to the F1 score. Higher Dice values indicate better semantic consistency between the two segmentation outputs.

D.5 Metric Directionality

For error-based metrics, including AbsRel, RMSE, mean angular error, median angular error, and corner error, lower values indicate better downstream consistency. For overlap- or accuracy-based metrics, including 2DIoU, 3DIoU, normal threshold accuracy, δ_1, mIoU, and Dice, higher values indicate better performance. These metrics evaluate whether generated panoramas preserve geometric structure, surface orientation, room layout annotations, and semantic content in downstream perception pipelines.

E Extended Downstream Metric Tables

Table 4: Full downstream probing on full-view panoramas on RandBalls. DA^2 is evaluated on the original pose-aligned test split, while OOOPS and LGTNet are evaluated on gravity-aligned samples matching their canonical input assumptions. DA^2 evaluates dense geometry, OOOPS evaluates semantic consistency, and LGTNet evaluates indoor layout on Laval Indoor. "–" means not evaluated.

	DA^2: depth / normal							OOOPS: semantics		LGTNet: layout				
Method	AbsRel↓	sRMSE↓	Mean↓	Median↓	$\text{Acc}_{11.25^\circ}$ ↑	$\text{Acc}_{22.5^\circ}$ ↑	Acc_{30° ↑	mIoU↑	Dice↑	3DIoU↑	2DIoU↑	CE↓	RMSE↓	δ_1 ↑
Georgoulis et al.	1.3781	0.2099	40.51	37.76	0.0925	0.2456	0.3681	0.2098	0.3073	0.1463	0.1557	7.4551	1.2915	0.3186
LDMs Outpainting	1.2540	0.2064	37.74	32.47	0.1347	0.3388	0.4627	0.1517	0.2272	0.3279	0.3589	4.5457	0.9338	0.5113
LuxDiT	0.7690	0.2083	39.91	36.26	0.1083	0.2798	0.4006	0.3846	0.5235	0.3545	0.3857	4.3768	0.8711	0.5434
DRMNet	1.3163	0.1953	35.67	30.63	0.1545	0.3647	0.4901	0.1223	0.1916	0.4385	0.4759	3.2563	0.7247	0.6255
ReFlex	**0.5172**	**0.1712**	**30.89**	**23.18**	**0.2318**	**0.4876**	**0.6070**	**0.4344**	**0.5778**	**0.4948**	**0.5301**	**2.9661**	**0.6595**	**0.6879**

Table 5: Full downstream probing on full-view panoramas on Digital Gene. DA^2 is evaluated on the original pose-aligned test split, while OOOPS and LGTNet are evaluated on gravity-aligned samples matching their canonical input assumptions. DA^2 evaluates dense geometry, OOOPS evaluates semantic consistency, and LGTNet evaluates indoor layout on Laval Indoor. "–" means not evaluated.

	DA^2: depth / normal							OOOPS: semantics		LGTNet: layout				
Method	AbsRel↓	sRMSE↓	Mean↓	Median↓	$\text{Acc}_{11.25^\circ}$ ↑	$\text{Acc}_{22.5^\circ}$ ↑	Acc_{30° ↑	mIoU↑	Dice↑	3DIoU↑	2DIoU↑	CE↓	RMSE↓	δ_1 ↑
Georgoulis et al.	1.7340	0.2170	42.07	39.31	0.0717	0.2169	0.3391	0.0424	0.0768	0.1703	0.1797	6.3930	1.2677	0.3469
LDMs Outpainting	1.4262	0.2184	40.60	36.57	0.1017	0.2808	0.4007	0.0332	0.0613	0.3035	0.3306	4.8389	0.9754	0.5239
LuxDiT	1.1275	0.2060	38.66	34.13	0.1323	0.3195	0.4381	0.2804	0.4080	0.3200	0.3487	4.6988	0.9464	0.4989
DRMNet	1.0306	0.2102	39.13	35.43	0.1065	0.2867	0.4124	0.0045	0.0088	0.3063	0.3331	4.4751	0.9503	0.5236
Ours	**0.6853**	**0.1936**	**34.68**	**27.44**	**0.1859**	**0.4179**	**0.5380**	**0.2882**	**0.4117**	**0.3559**	**0.3869**	**4.0646**	**0.9027**	**0.5503**

Table 6: Full downstream probing on full-view panoramas on OmniObject3D. DA^2 is evaluated on the original pose-aligned test split, while OOOPS and LGTNet are evaluated on gravity-aligned samples matching their canonical input assumptions. DA^2 evaluates dense geometry, OOOPS evaluates semantic consistency, and LGTNet evaluates indoor layout on Laval Indoor. "–" means not evaluated.

	DA^2: depth / normal							OOOPS: semantics		LGTNet: layout				
Method	AbsRel↓	sRMSE↓	Mean↓	Median↓	$\text{Acc}_{11.25^\circ}$ ↑	$\text{Acc}_{22.5^\circ}$ ↑	Acc_{30° ↑	mIoU↑	Dice↑	3DIoU↑	2DIoU↑	CE↓	RMSE↓	δ_1 ↑
Georgoulis et al.	1.6100	0.2106	40.88	38.05	0.0891	0.2406	0.3628	0.0053	0.0103	0.1570	0.1674	7.4617	1.3083	0.3401
LDMs Outpainting	1.5281	0.2187	40.84	36.81	0.0982	0.2760	0.3964	0.0304	0.0551	0.3132	0.3422	4.5563	0.9619	0.5187
LuxDiT	1.4948	0.2069	39.22	34.78	0.1233	0.3075	0.4276	**0.2948**	**0.4286**	0.3332	0.3599	4.6594	0.9129	0.5173
DRMNet	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Ours	**0.8927**	**0.1959**	**35.56**	**28.90**	**0.1716**	**0.3951**	**0.5164**	0.2606	0.3802	**0.3666**	**0.3987**	**4.1086**	**0.8717**	**0.5519**

F Statistical Significance of RGR Refinement

To assess the statistical significance of the improvements introduced by the Reliability-Guided Refiner (RGR), we fix the same trained Observation-Consistent Estimator (OCE) checkpoint and independently train RGR with three different random seeds. This setting isolates the variability introduced by Stage-2 training while maintaining a common OCE reference.

As shown in Tab. 7, all three RGR runs consistently reduce FID and KID relative to the fixed OCE baseline. In particular, the mean FID decreases from 59.02 to 45.40, corresponding to a 23.1% reduction. A two-sided one-sample t-test on the seed-level FID differences relative to the fixed OCE baseline yields $t(2) = -7.83$ and $p = 0.016$, with a 95% confidence interval of $[-21.11, -6.14]$ for the FID difference. This indicates a statistically significant reduction in FID under the fixed-OCE evaluation setting.

Table 7: **Statistical evaluation of RGR across three training seeds.** All RGR runs share the same fixed OCE checkpoint. The OCE-only row reports the prediction before refinement. Mean ± standard deviation is computed over the three independently trained RGR runs.

Setting	PSNR↑	SSIM↑	LPIPS↓	FID↓	KID↓
OCE only	19.343	0.643	0.314	59.02	0.0323
RGR, seed 42	19.440	0.654	0.316	42.27	0.0239
RGR, seed 43	19.343	0.653	0.314	48.28	0.0288
RGR, seed 44	19.363	0.653	0.315	45.64	0.0267
RGR, mean ± std	19.382 ± 0.051	0.653 ± 0.001	0.315 ± 0.001	45.40 ± 3.01	0.0265 ± 0.0025

KID also decreases across all three runs, from 0.0323 to 0.0265 on average. Meanwhile, SSIM consistently improves, while PSNR and LPIPS remain approximately unchanged. These results provide evidence that RGR improves distributional realism across the evaluated training seeds while preserving the fidelity established by OCE.

G Failure Case Analysis

Failure Patterns across Regimes. Figure 7 shows that reconstruction failures occur across different regimes, but their manifestations vary with the available reflection evidence. In RandBalls, where the scene exhibits relatively regular geometric structures, failures mainly involve the loss of thin structures and local geometric distortions. The reconstruction preserves the overall appearance of the interior but alters the arrangement of railings and roof supports. In contrast, Digital Gene and OmniObject3D pose greater reconstruction challenges due to complex reflector geometries and limited observation coverage. In these cases, ReFlex preserves the dominant scene appearance, although some fine structures are simplified or omitted when the available reflection cues provide insufficient geometric constraints. These cases also highlight a limitation of observation-grounded reconstruction: when the input provides insufficient evidence, ReFlex may produce simplified structures rather than synthesize visually plausible but potentially unsupported details.

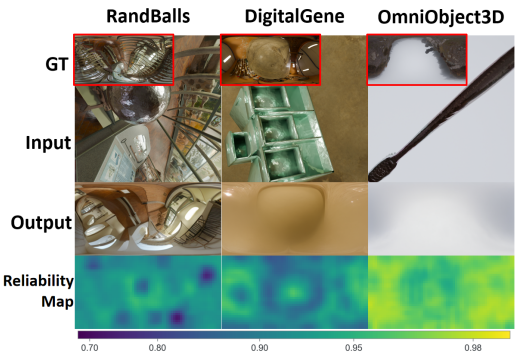

Figure 7: **Representative failure cases across three regimes.** Each column shows the input with an inset ground-truth panorama (red border), the ReFlex reconstruction, and the corresponding reliability map.

Limitations of reliability guidance. Reliability measures internal hierarchical consistency rather than guaranteed correctness. A shared confident error across stages can keep an incorrect structure anchored too long. Conversely, a correct but diffuse prediction may receive a low reliability estimate and be released prematurely, weakening observation anchoring during subsequent generative refinement. Posterior miscalibration or distribution shift can amplify both effects.

H Real-World Acquisition and Cross-Camera Alignment

This section provides the acquisition and calibration protocol for the real-world evaluation in Sec. 4.3. The benchmark contains 208 reflective RGB inputs from 10 indoor/outdoor scenes and 14 reflector configurations. Reflective inputs are captured with a Canon EOS 77D, while the corresponding panorama references are captured with an Insta360 X5. No real image is used for training, validation, hyperparameter selection, checkpoint selection, or model adaptation.

H.1 Acquisition Protocol

For each scene, the primary tripod remains fixed throughout acquisition. We first mount the Insta360 X5 and capture the reference panorama. The X5 is then replaced by the Canon EOS 77D using a quick-release adapter to minimize changes in the nominal camera position and orientation.

A reflective object is mounted on a second tripod and positioned near the center of the Canon viewfinder. After the Canon position, adapter height, and orientation are fixed, images for the available reflector configurations are captured without changing the Canon pose. Consequently, all Canon images acquired within the same scene share a common nominal camera pose.

The evaluation includes four indoor scenes and six outdoor scenes, covering laboratory and corridor interiors, a lounge, campus walkways, a sculpture plaza, a tree-lined road, a sports-field environment, and a shaded EV-charging area. The reflector set contains seven spherical configurations and seven non-spherical metalware objects, providing different degrees of curvature, directional coverage, self-occlusion, and geometric complexity.

H.2 Geometric Alignment

Because the reflective RGB input and the panorama reference are captured by different cameras, we explicitly estimate their relative orientation before computing paired reconstruction metrics.

For each Canon capture, we construct its projection model from the EXIF focal length together with the known sensor size and crop geometry. We mask the reflector and its supporting structure, and establish correspondences between the directly visible Canon background and the Insta360 panorama. These corresponding pixels are converted into spherical viewing directions, and the relative camera-to-panorama rotation

$$R_{\mathrm{C}\to\mathrm{P}} \in SO(3) \tag{112}$$

is robustly estimated using spherical RANSAC.

In implementation, the Canon view is kept fixed while the Insta360 panorama is inverse-warped and resampled into the Canon image plane for calibration and validation. The estimated rotation is subsequently frozen and shared by all reconstruction methods evaluated on the corresponding capture pair.

H.3 Photometric Alignment

Cross-device differences in exposure, white balance, camera response, and color rendering can bias pixel-level metrics even when the geometry is well aligned. We therefore perform a model-independent photometric calibration after geometric alignment.

The calibration uses only mutually observed background regions. The reflector, its support, foreground objects, invalid pixels, and strongly under- or over-exposed regions are excluded from the fitting mask. To reduce sensitivity to fine-texture displacement, sensor noise, and small residual geometric misalignment, the aligned image pair is Gaussian-filtered before fitting.

For each capture pair, we robustly estimate a spatially global RGB affine transformation

$$\mathbf{c} = A\mathbf{p} + \mathbf{b}, \tag{113}$$

where $\mathbf{p} \in \mathbb{R}^3$ denotes the projected Insta360 RGB value and $\mathbf{c} \in \mathbb{R}^3$ denotes the corresponding Canon RGB value. The matrix $A \in \mathbb{R}^{3\times3}$ accounts for global exposure, white-balance, and cross-channel response differences, while $\mathbf{b} \in \mathbb{R}^3$ models global color offsets.

We use a spatially global transformation rather than a spatially varying correction to avoid artificially modifying local scene content. The transformation is fitted on spatial training blocks and evaluated on disjoint held-out blocks, with a guard band separating the two sets to prevent the Gaussian filtering support from crossing the split.

H.4 Residual Alignment Error

We independently quantify the remaining geometric and photometric discrepancies after calibration.

Table 8: **Residual cross-camera alignment error.** Geometric error is measured using pairwise median structural-edge residual normalized by Canon crop width. Photometric error is the held-out median absolute RGB residual after low-frequency filtering.

Statistic	Median	Mean	P90	Max	SD
Geometry (% width)	1.64	2.10	4.78	8.39	1.96
Photometric	0.0450	0.0595	0.1301	0.1830	0.0444

For geometric alignment, the calibrated panorama is rendered back into the Canon image plane. The rendered panorama and Canon image are converted to grayscale and processed with contrast normalization, Gaussian smoothing, and Sobel edge extraction. After excluding the reflector, its support, and invalid regions, we measure for each valid structural edge pixel in the Canon image the distance to the nearest edge in the aligned panorama.

Across capture pairs, the median of the pairwise median edge residual is 16.79 pixels, corresponding to 1.64% of the Canon crop width. The across-pair mean, P90, maximum, and standard deviation are 2.10%, 4.78%, 8.39%, and 1.96% of the crop width, respectively.

For photometric alignment, we evaluate the median absolute RGB residual on held-out spatial blocks after low-frequency filtering. The across-pair median residual is 0.0450, with a mean of 0.0595, P90 of 0.1301, maximum of 0.1830, and standard deviation of 0.0444.

The remaining geometric discrepancy mainly arises from local parallax around finite-depth objects and occlusion boundaries, imperfect coincidence of the two optical centers, lens-model inaccuracies, weak or repeated structural edges, and temporal scene changes. These residuals therefore prevent the real-world references from being interpreted as perfectly pixel-aligned ground truth.

H.5 Sensitivity of Reconstruction Metrics to Alignment Error

To examine the sensitivity of reconstruction metrics to the measured cross-camera discrepancies, we additionally perturb ReFlex outputs on 3,700 held-out OmniObject3D samples using the measured geometric and photometric error distributions. We evaluate the same metric pipeline under unperturbed, geometry-only, photometric-only, and joint perturbations, and average the perturbed results over three random seeds.

The unperturbed model obtains 14.8096 PSNR, 0.5992 SSIM, and 0.4671 LPIPS. Geometry-only perturbations have little influence on PSNR and SSIM, while photometric perturbations are the larger source of metric sensitivity. Applying geometric and photometric perturbations jointly decreases PSNR by approximately 0.91 dB and SSIM by 0.0149, while increasing LPIPS by 0.0454.

These results indicate that residual cross-camera discrepancies can modestly reduce the absolute paired reconstruction scores, but are unlikely to account for the full real-world degradation observed in Sec. 4.3. Importantly, geometric alignment, photometric calibration, and the resulting reference panoramas are estimated independently of the reconstruction method and are frozen and shared across all evaluated methods.

I Extended Real-World Evaluation

This section extends the real-world evaluation in Sec. 4.3 with aggregate, reflector-wise, and scene-wise results. The evaluation includes all 208 inputs from 10 scenes and 14 reflector configurations. All methods use the same calibrated input–reference pairs under a model-level zero-shot protocol: no real image is used for training, validation, hyperparameter tuning, or checkpoint selection. Acquisition, cross-camera calibration, and residual-alignment analyses are provided in Appendix H.

I.1 Overall Performance and Score Distribution

Tab. 9 reports the aggregate reconstruction results. ReFlex outperforms LuxDiT on all three paired metrics and also improves over OCE alone. These comparisons provide a common-protocol assessment of zero-shot transfer; the analyses below further examine how reconstruction quality varies with the available reflection evidence and scene conditions.

Table 9: **Aggregate reconstruction results on the real-world evaluation set.** All methods are evaluated on the same 208 calibrated pairs without real-data adaptation. PSNR is reported in dB.

Method	PSNR↑	SSIM↑	LPIPS↓
LuxDiT	10.8513	0.2952	0.7970
OCE only	12.6295	0.3522	0.6895
ReFlex	**12.7506**	**0.3619**	**0.6777**

Table 10: **Distribution of ReFlex reconstruction scores across all 208 real inputs.** SD denotes the sample standard deviation across inputs, not training-seed variability. PSNR is reported in dB.

Metric	Minimum	Median	Mean	SD	Maximum
PSNR↑	7.471	12.773	12.751	1.582	20.027
SSIM↑	0.120	0.363	0.362	0.105	0.716
LPIPS↓	0.397	0.683	0.678	0.093	0.856

To contextualize the aggregate scores, Tab. 10 summarizes their distribution over all 208 inputs. The means are close to the medians for all three metrics, suggesting that the reported averages describe the central performance rather than a few unusually favorable examples. The broad ranges also show substantial variation across inputs. This distributional summary characterizes the results; by itself, it does not establish reconstruction faithfulness.

I.2 Reconstruction across Reflector Configurations

To examine the relationship between reconstruction quality and the available reflection evidence, we report results for all 14 reflector configurations in Tab. 11. Each configuration is evaluated across the available scenes containing that reflector, and no reflector configuration is excluded from the overall evaluation.

The seven spherical configurations average 13.624 dB PSNR, 0.3844 SSIM, and 0.6111 LPIPS, compared with 12.287 dB, 0.3524, and 0.7125 for non-spherical metalware. The sphere-level PSNR values range from 13.086 to 14.057 dB, indicating that the stronger spherical-reflector results are not attributable to a single unusually favorable configuration.

The reflector-wise trend is consistent with the continuity and directional coverage of the observations. Smooth convex spheres provide continuous curvature, broadly distributed reflection evidence, and limited self-occlusion. In contrast, handles, concavities, ornamentation, and thin boundaries fragment or attenuate the reflected observations. For example, the Silver Ornate Cup obtains 11.240 dB PSNR and 0.7590 LPIPS, whereas the spherical configurations average 13.624 dB and 0.6111, respectively. The Gold Goblet and Silver Ewer provide intermediate reconstruction quality, with PSNR values of 12.309 and 12.515 dB.

This improvement with more continuous and broadly distributed reflection evidence provides evidence consistent with observation-grounded reconstruction: recovery becomes more accurate when the input supplies more informative constraints. The same qualitative trend appears in the synthetic benchmark, where ReFlex achieves stronger reconstruction on RandBalls than on the more geometrically complex OmniObject3D regime (Tab. 1).

I.3 Reconstruction across Real-World Scenes

Tab. 12 reports reconstruction quality across all 10 captured scenes. The evaluation covers four indoor scenes: a laboratory, two corridors, and a lounge, and six outdoor scenes. Each scene-level entry reports the mean and sample standard deviation across its available input samples.

The results do not support a universal ordering between indoor and outdoor environments. Outdoor inputs achieve higher pooled PSNR (12.905 versus 12.513 dB), whereas indoor inputs achieve better SSIM and LPIPS ($0.4110/0.6570$ versus $0.3300/0.6912$). The relatively stable illumination and regular structure of the evaluated indoor scenes may partly explain their stronger SSIM and LPIPS, while broad sky and ground regions may contribute to the higher outdoor PSNR.

Table 11: **ReFlex reconstruction quality across real reflector configurations.** Results are aggregated across the available scenes for each configuration. The group-summary rows average the seven listed configuration-level results within each group. PSNR is reported in dB.

Reflector configuration	PSNR↑	SSIM↑	LPIPS↓
Spherical configurations			
Black Sphere	13.461	0.3862	0.5939
Cyan Sphere	13.518	0.3772	0.6121
Green Sphere	13.567	0.3665	0.6018
Gold Sphere	13.806	0.4088	0.5994
Red Sphere	14.057	0.3890	0.6588
Purple Sphere	13.873	0.3882	0.6283
Silver Sphere	13.086	0.3752	0.5835
Mean across seven configurations	13.624	0.3844	0.6111
Non-spherical metalware			
Silver Ewer	12.515	0.3713	0.6866
Silver Ornate Cup	11.240	0.3481	0.7590
Silver Creamer	12.348	0.3565	0.7054
Silver Round Jar	12.586	0.3509	0.6971
Gold Coffee Pot	13.080	0.3411	0.6986
Gold Goblet	12.309	0.3539	0.7205
Gold Tray	11.934	0.3450	0.7203
Mean across seven configurations	12.287	0.3524	0.7125

Table 12: **Real-world reconstruction quality across scenes.** Results are reported as mean ± sample standard deviation over available reflector inputs. Indoor and outdoor totals pool all inputs within each subset. PSNR is in dB.

Scene	PSNR↑	SSIM↑	LPIPS↓
Indoor scenes			
Laboratory Interior	13.316 ± 0.789	0.3915 ± 0.0378	0.7053 ± 0.0501
Research Building Third-Floor Corridor	12.784 ± 1.428	0.3570 ± 0.0544	0.5810 ± 0.0815
Research Building Second-Floor Lounge	12.572 ± 0.483	0.5264 ± 0.0329	0.6059 ± 0.0366
Basement Entrance Corridor	11.407 ± 1.395	0.3638 ± 0.0305	0.7284 ± 0.0572
Indoor Total	12.513 ± 1.284	0.4110 ± 0.0795	0.6570 ± 0.0848
Outdoor scenes			
Teaching-Building Sculpture Plaza	14.950 ± 2.089	0.4730 ± 0.1008	0.5483 ± 0.0756
Cafeteria–Basketball Court Path	12.998 ± 0.939	0.4255 ± 0.0367	0.6783 ± 0.0679
Bridge-Side Tree-Lined Road	12.731 ± 1.547	0.2717 ± 0.0314	0.6991 ± 0.0626
Mini Sports Field	12.697 ± 1.192	0.3389 ± 0.0431	0.7142 ± 0.0745
Shaded EV Charging Area	12.438 ± 0.913	0.2220 ± 0.0301	0.7563 ± 0.0536
Research Building–Cafeteria Walkway	11.671 ± 1.622	0.2683 ± 0.0734	0.7435 ± 0.0750
Outdoor Total	12.905 ± 1.737	0.3300 ± 0.1072	0.6912 ± 0.0964

More informative differences appear at the individual-scene level. The Teaching-Building Sculpture Plaza achieves the highest scene-level PSNR and lowest LPIPS, with 14.950 dB PSNR, 0.4730 SSIM, and 0.5483 LPIPS. It contains stable daylight, regular facades, a clear horizon, and broad ground regions, which are relatively close to the outdoor environments represented during synthetic training. In contrast, the Shaded EV Charging Area has the lowest SSIM and highest LPIPS, with 12.438 dB PSNR, 0.2220 SSIM, and 0.7563 LPIPS. Its uneven tree shade, mixed illumination, signboards, and charging equipment provide more irregular observations and include content less represented in the synthetic training distribution.

These scene-wise differences complement the reflector-wise analysis. Reflector geometry affects the continuity of the observed evidence, while scene structure and illumination affect how reliably

Table 13: **Stage-wise training cost on a single RTX 4090.** The total combines one OCE training and one diffusion-prior LoRA adaptation. The reported peak training memory is approximately 14 GB.

Training stage	GPU-hours
OCE training	108
Diffusion-prior LoRA adaptation	12
Two-stage total	120

Table 14: **End-to-end inference profile on a single RTX 4090.** ReFlex includes both OCE and RGR. Parameter counts refer to trainable parameters, and TFLOPs denotes floating-point operations per panorama.

Method	Latency (ms/image)↓	FPS↑	Peak inference memory (MB)↓	Trainable parameters (M)	TFLOPs per panorama
DeLightNet [17]	487.73	2.050	197	7.763	0.048
Georgoulis et al. [5]	493.20	2.028	286	17.306	0.066
LDMs Outpainting [8]	1247.03	0.802	1978	294.974	6.438
DRMNet [6]	4821.88	0.207	6345	418.792	29.766
LuxDiT [11]	17106.22	0.058	12549	148.636	365.072
ReFlex	1200.63	0.833	4424	281.122	48.203

that evidence supports reconstruction. The observed variation is therefore consistent with stronger recovery when both the reflected observations and the surrounding scene provide stable, informative constraints.

J Computational Complexity and Resource Usage

This section reports the hardware configuration, stage-wise training cost, end-to-end inference profile, and component-wise runtime of ReFlex.

J.1 Hardware and Training Cost

The experiments were conducted on a local Windows workstation with one NVIDIA RTX 4090 GPU with 24 GB memory, an Intel Core i9-14900KF CPU, and 32 GB system RAM. OCE and RGR are optimized in two separate stages. OCE training uses a batch size of 2 for 80 epochs, with the DINO encoder and VQ encoder/decoder kept frozen. For RGR, only LoRA adapters in the U-Net attention projections are optimized; the Stable Diffusion 2.1 Base backbone, VAE, and text encoder remain frozen. Detailed training objectives are provided in Appendix B.

OCE training and diffusion-prior LoRA adaptation require approximately 108 and 12 GPU-hours, respectively, on the single RTX 4090. The reported peak GPU memory during training is approximately 14 GB. Tab. 13 summarizes the stage-wise training costs. The 120 GPU-hour total covers one OCE training and one LoRA adaptation, rather than all ablations, baseline training, or additional training-seed runs.

J.2 End-to-End Inference Profile

We profile complete inference pipelines on the same RTX 4090. For ReFlex, the measurements include both OCE reconstruction and RGR refinement. The default inference configuration uses eight MaskGIT iterations per hypothesis/detail stage and 50 DDIM refinement steps, with maximum anchoring duration $r_{\mathrm{ub}} = 50$. Baseline adaptation and checkpoint selection follow the protocol in Appendix C.

Tab. 14 reports latency, throughput, peak inference memory, trainable parameter count, and floating-point operations per panorama. Trainable parameter counts exclude frozen pretrained parameters; TFLOPs per panorama measures computational work rather than hardware throughput.

Table 15: **Component-wise runtime of RGR on a single RTX 4090.** Measurements use the default inference configuration.

RGR component	Latency (ms)
U-Net denoising	665.38
VAE encoding	13.68
VAE decoding	27.67
Reliability computation and conditioning	2.73

The full ReFlex pipeline requires 1200.63 ms per panorama and 4424 MB peak inference memory. Its latency is comparable to LDMs Outpainting, while the measured latencies correspond to approximately $4.02\times$ and $14.25\times$ speedups over DRMNet and LuxDiT, respectively. Compared with lightweight DeLightNet, ReFlex incurs additional computation but achieves stronger paired reconstruction fidelity (Tab. 1). The profile therefore reflects an efficiency–quality trade-off, rather than uniformly lower computational cost across all methods.

J.3 Runtime Breakdown of Reliability-Guided Refinement

We further profile RGR components separately on the same RTX 4090 under the default inference configuration. Tab. 15 distinguishes U-Net denoising, VAE encoding and decoding, and reliability computation together with conditioning.

U-Net denoising accounts for the dominant refinement cost, whereas reliability computation and conditioning require only 2.73 ms. The additional runtime of RGR is therefore primarily attributable to diffusion sampling rather than the reliability formulation itself.

K Dataset Documentation and Responsible Release

This section documents the synthetic ReFlex-Bench, its source assets, and its release conditions. The separately collected real-world evaluation set is described in Appendix H. Computational resources are reported in Appendix J.

K.1 Source Assets and Licenses

Tab. 16 documents the source assets, access dates, license terms, and source-specific release conditions.

K.2 ReFlex-Bench Documentation

Tab. 17 summarizes the benchmark composition, generation pipeline, data format, release plan, and known limitations.

K.3 Safeguards and Responsible Release

ReFlex reconstructs out-of-frustum scene content from reflective observations, which may raise privacy concerns if applied to real-world images without consent. To mitigate this risk, our released benchmark is based on rendered synthetic data and does not include private real-world photographs, human-subject data, or personally identifying information. We will release the dataset, code, and model checkpoints with an intended-use statement restricting their use to research, benchmarking, and scene-understanding studies. The release will explicitly prohibit privacy-invasive applications, including surveillance, covert reconstruction of private spaces, or inference from reflective surfaces without consent from the scene owner or affected individuals. We will also provide a model card and dataset documentation describing the data source, intended use, limitations, and misuse risks. The released code is designed for offline research evaluation and does not provide a real-time surveillance or deployment interface.

Table 16: **Existing source assets used to construct ReFlex-Bench.** All assets were accessed on 2026/01/12. We list the source URL, number of HDR environments or object sources used where applicable, license or terms, and release notes.

Asset	Source URL	Count	Access date	License / terms	Use and release notes
CGEES HDR environments	https://cgees.com/	34 HDRs	2026/01/12	CC0 for downloadable HDRI content.	Used as HDR source environments. CGEES-derived rendered pairs are included in the redistributable subset of ReFlex-Bench.
HDRMAPS / HDRI Maps	https://hdrmaps.com/	200 HDRs	2026/01/12	Custom royalty-free, non-transferable license; redistribution or repackaging of source images is restricted by the provider's terms.	Used as HDR source environments under the provider's terms. We will not redistribute original HDRMAPS files or HDRMAPS-derived rendered pairs unless redistribution permission is obtained. We will release manifests and rendering scripts for reproducibility.
HDRI-Skies	https://hdri-skies.com/	58 HDRs	2026/01/12	Custom royalty-free, non-transferable license; redistribution of source images as standalone assets is restricted by the provider's terms.	Used as HDR source environments under the provider's terms. We will not redistribute original HDRI-Skies files or HDRI-Skies-derived rendered pairs unless redistribution permission is obtained. We will release manifests and rendering scripts for reproducibility.
Laval Indoor HDR Database	http://hdrdb.com/indoor/	2,362 HDRs	2026/01/12	Written research permission obtained; redistribution permission has not yet been obtained.	Used as indoor HDR source environments. We will not redistribute original Laval panoramas or Laval-derived rendered pairs unless explicit redistribution permission is granted. We will release manifests and rendering scripts for users who obtain the Laval assets from the official source.
Laval Outdoor HDR Database	http://hdrdb.com/outdoor/	205 HDRs	2026/01/12	Written research permission obtained; redistribution permission has not yet been obtained.	Used as outdoor HDR source environments. We will not redistribute original Laval panoramas or Laval-derived rendered pairs unless explicit redistribution permission is granted. We will release manifests and rendering scripts for users who obtain the Laval assets from the official source.
PolyHaven HDRIs	https://polyhaven.com/hdris/	833 HDRs	2026/01/12	CC0.	Used as HDR source environments. PolyHaven-derived rendered pairs are included in the redistributable subset of ReFlex-Bench.
Digital Gene	https://github.com/sii-research/DigitalGene	object source	2026/01/12	License not publicly specified in the repository at the time of writing.	Used as reflective object geometry for the Digital Gene regime. We cite the original work and will redistribute Digital-Gene-derived rendered pairs only if the asset license or author permission permits redistribution. Otherwise, we release manifests and rendering scripts.
OmniObject3D	https://github.com/omniobject3d/OmniObject3D	object source	2026/01/12	CC BY 4.0.	Used as scanned real-world object geometry for the OmniObject3D regime. We preserve attribution metadata and include OmniObject3D-derived rendered pairs when all upstream HDR sources in the corresponding samples permit redistribution.

Table 17: **Documentation for the new asset introduced by this work.** ReFlex-Bench is a paired benchmark of reflective RGB inputs and pose-aligned target panoramas. We document its composition, generation pipeline, file format, release plan, and known limitations.

Documentation item	**Description**
Asset name	ReFlex-Bench.
Asset type	A paired synthetic benchmark containing reflective RGB inputs and pose-aligned equirectangular target panoramas for faithful panorama reconstruction from a single reflective image.
Dataset scale	ReFlex-Bench contains 110,760 rendered image–panorama pairs. The benchmark is evenly divided into three observation regimes: RandBalls, Digital Gene, and OmniObject3D, with 36,920 pairs per regime.
Source environments	The benchmark uses 3,692 HDR source environments: 34 from CGEES, 200 from HDRMAPS / HDRI Maps, 58 from HDRI-Skies, 2,362 from Laval Indoor, 205 from Laval Outdoor, and 833 from PolyHaven. All source assets were accessed on 2026/01/12.
Observation regimes	RandBalls provides dense and controlled reflection cues; Digital Gene provides complex reflective geometry and self-occlusion; OmniObject3D provides scanned real-world object geometry.
Data split	We first split source HDR environments at the scene level into training and test environments with a 9:1 ratio to prevent environment leakage. A validation subset is then carved out from the training environments, resulting in an effective 8:1:1 train/validation/test split.
File format	Each reflective RGB input is stored as an 8-bit PNG image at 1024×1024 resolution. Each pose-aligned equirectangular target panorama is stored as an 8-bit PNG image at 512×256 resolution.
Rendering pipeline	All samples are rendered in Blender 4.5 with random seed 42. For each HDR environment, we render a reflective RGB observation paired with the corresponding pose-aligned target panorama.
Release plan	We will release code, checkpoints, rendering/preprocessing pipelines, scene-level splits, evaluation code, and all benchmark pairs cleared for redistribution. For restricted-source subsets, we will provide source manifests, fixed random seeds, and deterministic rendering scripts, enabling users who independently obtain the upstream assets from the official providers to regenerate the corresponding samples.
License of ReFlex-Bench	The released subset will use a license compatible with all upstream assets included in that subset. Source-specific attribution, license notices, and redistribution restrictions will be documented in the released dataset card.
Restricted-source subsets	HDRMAPS, HDRI-Skies, and Laval-derived samples are treated as restricted-source subsets unless redistribution permission is obtained. Digital-Gene-derived samples are also treated conservatively until the license or author permission is confirmed. These samples are included in our experiments but will not be directly redistributed without the required permission.
Privacy and filtering	We do not collect new personal data or conduct human-subject studies. Since some third-party HDR environments may depict real-world indoor or outdoor scenes, released documentation will include a takedown and correction procedure for privacy or rights-related concerns.
Maintenance	The released benchmark will include versioned manifests, split files, rendering scripts, and evaluation scripts. We will document updates, corrections, and any removed samples in a version history. We will provide versioned updates as additional subsets become eligible for redistribution.
Known limitations	ReFlex-Bench is based on physically based rendering and may not fully capture real-world reflection formation, including sensor noise, mixed materials, imperfect calibration, and inter-reflections.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the main contributions: ReFlex, the Observation-Consistent Estimator, the Reliability-Guided Refiner, and ReFlex-Bench. These claims are supported by the experiments in Section 4, including image-level reconstruction, distributional realism, downstream perception, and ablation results.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses limitations in Section 5 and Appendix K, including the use of physically based simulation and the potential gap to real-world reflection formation involving sensor noise, mixed materials, imperfect calibration, and inter-reflections.

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Appendix A.3.1 states the assumptions of the oracle categorical model and provides proofs for Claims 1–3. The inference-time reliability construction is presented as a surrogate rather than a guarantee of prediction correctness.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper describes the benchmark construction, data splits, evaluation metrics, baselines, training setup, inference schedule, and detailed model components in Sections 3 and 4. Additional training objectives, baseline adaptation protocols, downstream metrics, statistical uncertainty analysis, dataset documentation, and compute resources are provided in the appendix.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: We provide a release plan for code, rendering scripts, split files, evaluation scripts, source manifests, and redistributable benchmark subsets. Some rendered pairs also cannot be redistributed unless upstream licenses or written permissions allow it.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 4.1 specifies the benchmark construction, scene-level split, input and target resolutions, metrics, baselines, optimizer, learning rate, batch size, number of epochs, mask-ratio schedule, and inference schedule. Further implementation, objective, baseline, and evaluation details are provided in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: Appendix F reports the mean and standard deviation over three independently trained Stage-2 RGR runs with a fixed OCE checkpoint.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix J reports the workstation configuration, GPU type and memory, CPU, system RAM, ReFlex training time, peak GPU memory, and ReFlex inference memory and runtime, including both OCE and RGR. Appendix C summarizes how baseline training or adaptation was handled under the shared benchmark protocol.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work uses rendered data and existing research assets, does not involve human-subject experiments or collection of new private personal data, and includes license, release, privacy, and responsible-use documentation in Appendix K. We are not aware of any aspect of the research that violates the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.

- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Sec. 5 discusses potential benefits for robotic navigation, embodied perception, and scene understanding. Appendix K.3 addresses privacy and misuse risks from reflection-based out-of-view reconstruction and describes the corresponding safeguards. Appendix K.2 documents the dataset release plan and privacy-related procedures.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: Appendix K.3 describes safeguards for responsible release, including synthetic-data-based release, intended-use restrictions, prohibition of privacy-invasive applications, model and dataset documentation, and avoiding a real-time surveillance or deployment interface. Appendix K.2 also describes privacy filtering, takedown/correction procedures, and restricted-source subsets.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Appendix K.1 lists the source assets used to construct ReFlex-Bench, including source URLs, counts, access dates, license or terms, and release notes. The paper also states that restricted-source samples will not be redistributed unless permission is obtained and that manifests and rendering scripts will be released for reproducibility where redistribution is restricted.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The paper introduces ReFlex-Bench and documents it in Appendix K.2, including asset type, scale, source environments, observation regimes, data split, file format, rendering pipeline, release plan, license plan, privacy and filtering, maintenance, and known limitations.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

Justification: The paper does not involve crowdsourcing or research with human subjects. All benchmark samples are rendered from existing environment and object assets.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not involve human subjects, user studies, crowdsourcing, or collection of new personal data. Therefore, IRB approval or equivalent human-subjects review is not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs are not used as an important, original, or non-standard component of the proposed method, benchmark construction, experiments, data analysis, or evaluation. Any LLM assistance, if used, was limited to writing, editing, formatting, or checklist preparation and did not affect the core methodology, scientific claims, experimental results, or originality of the research.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.