SCENE RECONSTRUCTION USING UBIQUITOUS WIRELESS SIGNALS VIA DIFFUSION POSTERIOR INFERENCE OVER PHYSICS-GROUNDED REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Knowing the layout of the physical world matters for many applications, yet the sensors we rely on for it (e.g., cameras and LiDAR) can fail in darkness, in adverse weather, or behind walls. Wireless signals do not face these limitations: every link between a base station and a user is shaped by the buildings it passes, reflects from, or is blocked by, leaving information about the surrounding scene embedded in the wireless channel measured at the endpoints. However, mapping channel measurements to the environment is a severely ill-posed problem: a measurement records only the structures that interact with the transmitted electromagnetic waves along its propagation paths, and even these observations can often be explained by multiple plausible scenes. We therefore formulate wireless scene reconstruction as posterior inference and develop a hierarchical framework with two complementary stages to approach this problem. A wireless foundation model, pretrained to respect propagation physics, transforms each channel measurement into a representation of the scene structure that shaped the observation. A diffusion model then jointly integrates representations from multiple measurements to model and sample the posterior over building occupancy, with consensus over samples producing the final reconstruction. We demonstrate that the framework can accurately recover building footprints from wireless channel measurements alone and show that its performance depends critically on both design components: physics-grounded representations expose scene-relevant information, while joint posterior inference coherently integrates complementary observations.

1 INTRODUCTION

Every wireless link is already an unintentional measurement of its surroundings. A signal that leaves a base station and arrives at a receiver has encountered several components in an environment: blocked by one building, reflected off another, diffracted around a corner. The environment is, in effect, written into the wireless channel measured at the receiver. If that writing could be read back, the radio infrastructure already deployed at scale could be repurposed as a sensor network, mapping the world through occluded regions, in darkness, and in severe weather, where cameras and lidar degrade or fail outright (Zhao et al., 2018; Dodds et al., 2026). It would also keep maps current when they change fastest: after an earthquake or a flood, the layout of a neighborhood can change in minutes, and the network is often the only sensor still running. Site-specific digital twins, which today require expensive surveys or hand-built geometry, could also be constructed and kept current from the traffic a network already carries (Hoydis et al., 2023; Li et al., 2024; Testolina et al., 2024). The knowledge of the propagation environment can also be used to optimize communication network throughput (Zhao et al., 2023; Demirhan & Alkhateeb, 2023), and could advance integrated communication and sensing, one of the main pillars of next-generation networks (Akyildiz et al., 2020; Wang et al., 2024).

Recovering geometry from wave measurements is well studied in acoustics, where room shape is inferred from resolvable echoes and geometric propagation models (Dokmanić et al., 2013; El Baba et al., 2018); recent learning-based approaches relax some of these assumptions and predict floorplans directly (Yeon & Choi, 2024; Yeon et al., 2024), but still rely on echo structures that can be attributed to individual surfaces. In the RF domain, scene reconstruction has primarily relied on dedicated radar sensing systems (Casmin & Oliveira, 2025; Borts et al., 2026), recovering occupancy grids or floor plans from millimeter-wave and ultra-wideband reflections using geometric inference, motion-synthesized aperture, structural priors, or learned generative models (Lu et al., 2020; Sun et al., 2022; Zhou et al., 2025; Lai et al., 2026). Recent approaches include diffusion-based indoor

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

scene completion from sparse radar heatmaps (Zhou et al., 2025) and reconstruction around corners from multi-bounce returns collected by a mobile radar (Lai et al., 2026). These approaches require dedicated sensing hardware and often movement through the environment, incurring additional hardware and deployment costs.

In this paper, we aim to use the communication signals from already deployed infrastructure for scene reconstruction. The challenge is that extracting environmental components out of a raw wireless channel measurement without any assumptions is a severely ill-posed inverse problem, and the ill-posedness is structural rather than incidental. A measurement is shaped only by the structures its propagation paths interact with, so the likelihood is exactly invariant to the rest of the scene. Even within the observed region, the geometry-to-channel mapping is many-to-one: finite bandwidth and aperture limit the resolution of path delays and angles, so scatterers displaced within those limits yield indistinguishable measurements. Compounding this, a structure impacts the channel only through the *role* it plays for a given link — reflecting, blocking, or leaving the signal untouched — and that role varies with where the structure sits relative to the transmitter and receiver. Different occupancy patterns whose structures play the same roles for a given link would therefore be indistinguishable. Hence, a single measurement admits not one scene but a family of them.

In this work, we formulate scene reconstruction as posterior inference over occupancy: the solution is a distribution over scenes, and the observations determine which parts of it are pinned down and which remain free. To our knowledge, this is the first treatment of wireless scene reconstruction that assumes nothing about the number, shape, convexity, or location of scatterers.

Learning this posterior directly from raw channel measurements is impractical. A single decoder would have to solve two difficult problems at once: work out from finite data how propagation physics maps geometry into a channel, and at the same time learn to generate plausible scenes from a combinatorially large space. We therefore adopt a *hierarchical* design that separates these burdens. First, a multi-modal wireless foundation model, pretrained to capture electromagnetic propagation physics and align channel observations with propagation-relevant scene structure, maps each measurement to a representation of the scene properties that could have produced it. Second, representations from multiple observations of the same environment are fused to condition a diffusion decoder, which performs posterior inference over occupancy and generates scenes consistent with the observed channels. Finally, we aggregate the sampled scenes through per-cell consensus to obtain the reconstructed occupancy map.

We evaluate reconstruction of arbitrary scene geometries against natural alternatives to each level of the hierarchy on $\sim$700 environments held out from all stages of training, spanning diverse spatial layouts and structural configurations, from dense urban blocks to open areas with isolated structures. Specifically, we show through controlled comparisons that both stages of the hierarchy are essential: independently reconstructing and aggregating per-observation scene estimates markedly underperforms, while conditioning the decoder directly on raw channel measurements collapses. By varying the input representation and posterior inference mechanism independently, we isolate the contribution of each stage. Together, these experiments demonstrate that physics-grounded representation learning followed by generative posterior inference provides the complementary capabilities needed to make reconstruction from sparse wireless observations tractable.

Contributions.

- **Formulation and theory.** We formulate wireless scene reconstruction as probabilistic posterior inference over occupancy with no parametric prior; characterize its ill-posedness and the resulting recoverable set; and derive finite-sample bounds showing that a consensus estimator over posterior samples is Bayes-optimal per cell and converges to the true scene on the recoverable set as the number of measurements grows.
- **Physics-grounded representations.** We show that pretraining a wireless foundation model and aligning its representation with propagation-relevant scene structures yields per-observation embeddings from which the scene occupancy distribution can be recovered. Each embedding captures the portion of the environment that shaped the corresponding measurement while expressing calibrated uncertainty elsewhere.
- **Hierarchical generative reconstruction.** We design a diffusion decoder that conditions on the set of per-observation representations and samples the occupancy posterior, and propose a consensus readout that fuses many viewpoints into a single reconstruction without assuming anything about scene structure.

2 PRELIMINARIES AND PROBLEM FORMULATION

2.1 WIRELESS PROPAGATION

A signal transmitted from a base station reaches a receiver along a set of propagation paths determined by the surrounding environment. Each path is a sequence of interactions with the physical structures of the scene: it may travel directly, reflect specularly off a building, diffract at an edge, or be blocked. For a receiver at position $\mathbf{p}$, propagation through a scene $\mathbf{s}$ yields a path set $\mathcal{P}(\mathbf{s}, \mathbf{p}) = \{(\alpha_\ell,\ \tau_\ell,\ \theta_\ell,\ \phi_\ell)\}_{\ell=1}^{L}$, where α_ℓ is the complex gain, τ_ℓ the delay, and θ_ℓ, ϕ_ℓ the angles of arrival and departure. Each of these parameters depends on scene geometry: delays are set by the lengths of the path segments, angles by the orientation of the surfaces the wave meets, and gains by free-space loss and the reflection coefficients of the structures encountered.

The receiver does not observe the paths individually. For a wideband multiple-input multiple-output (MIMO) link with N_{tx} transmit antennas, N_{rx} receive antennas, and N_{sc} subcarriers, it observes their superposition in the measured channel $H \in \mathbb{C}^{N_{tx} \times N_{rx} \times N_{sc}}$, whose entry for transmit/receive antenna t/r and subcarrier frequency f is

$$H[t, r, f] \;=\; \sum_{\ell \in \mathcal{P}(\mathbf{s},\mathbf{p})} \alpha_\ell \, a_r(\theta_\ell) \, a_t^*(\phi_\ell) \, e^{-j2\pi f \tau_\ell}, \tag{1}$$

with $a_t(\cdot)$ and $a_r(\cdot)$ the array responses (Appendix A). This composition defines a forward operator $H = \mathcal{F}(\mathbf{s}, \mathbf{p})$ mapping scene and link geometry to the observed channel: the scene is written into the channel through the path parameters, in a form the receiver observes only in superposition. Importantly, not all parts of the environment participate in the measurement. Only structures a path interacts with, allowing free propagation, blocking the direct ray, or serving as reflection or diffraction points, enter the path set and thus the channel. We call this region the *propagation-relevant set* of an observation; structures outside it leave the observation exactly unchanged.

2.2 THE INVERSE PROBLEM

Objective. We observe a scene through the wireless links a deployed network already maintains. A base station serves K users, where user k at position $\mathbf{p}_k$ contributes a channel measurement H_k, giving the observation set $\mathcal{O} = \{(H_k, \mathbf{p}_k)\}_{k=1}^{K}$. Denoting a_k for the propagation-relevant set of observation k, the *observable set* $\mathcal{A} = \bigcup_k a_k$ is the region of the scene influencing at least one measurement. We represent the environment as a two-dimensional occupancy grid over a square region with spatial resolution ΔR, $\mathbf{s} \in \{0, 1\}^{N \times N}$, where $s_i = 1$ indicates that cell i contains a structure, and seek to recover $\mathbf{s}$ from $\mathcal{O}$, inverting $\mathcal{F}$, without parametric assumptions on the number, shape, convexity, or location of the structures.

Why the problem is ill-posed. Under the standard Hadamard framework (Bakushinsky & Goncharsky, 2012), the problem is ill-posed because the observations do not uniquely determine the scene, despite a feasible solution existing. Non-uniqueness arises from three sources (Appendix B). First, structures outside $\mathcal{A}$ leave every measurement unchanged, so the likelihood is exactly invariant to them: flipping any combination of those cells yields a scene explaining the observations equally well, and the number of such scenes is exponential in $|\mathcal{A}^c|$. Second, even inside $\mathcal{A}$, the forward map is many-to-one, since finite bandwidth and finite aperture resolve delays and angles only to a limit, so scatterers displaced within those limits yield indistinguishable measurements. Third, a structure enters the channel only through the *role* it plays for a given link — reflecting, blocking, or leaving the signal untouched — and that role varies with where it sits relative to transmitter and receiver: the same building reflects for one user, blocks another, and is invisible to a third. Scenes whose structures play the same roles for every link are indistinguishable at any resolution, so unlike the second source, this degeneracy is not removed by better hardware. Multiple observations narrow the ambiguity by intersecting the families of scenes each admits, and additional users and reflected paths enlarge $\mathcal{A}$, but the intersection does not collapse to a single scene: the invariance outside $\mathcal{A}$ persists, and the resolution and role ambiguities remain within it.

2.3 A PROBABILISTIC FORMULATION

Because the data do not determine a unique scene, we do not seek one. We take the solution object to be the posterior over scenes $p(\mathbf{s}|\mathcal{O}) \propto p(\mathcal{O}|\mathbf{s})p(\mathbf{s})$, whose structure is dictated by the physics of wave propagation. Since structures outside $\mathcal{A}$ leave every measurement unchanged, the likelihood is invariant to them, and hence factorizing the prior gives

$$p(\mathbf{s} \mid \mathcal{O}) \;=\; \underbrace{p(\mathbf{s}_{\mathcal{A}} \mid \mathcal{O})}_{\text{extracted from the data}} \;\cdot\; \underbrace{p(\mathbf{s}_{\mathcal{A}^c} \mid \mathbf{s}_{\mathcal{A}})}_{\text{supplied by the prior}} . \tag{2}$$

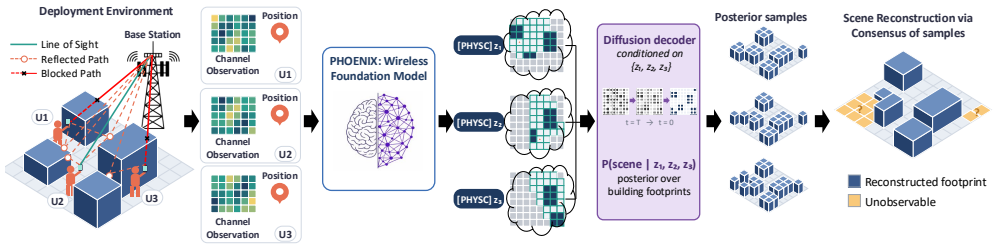

Figure 1: Overview of the proposed framework for wireless scene reconstruction using diffusion posterior inference over wireless foundation model representations.

The observation set does not appear in the second factor: it acts on the observable part of the scene and leaves the rest to the prior, conditioned on what the data did determine. The per-cell marginals $\pi_i = p(s_i = 1|\mathcal{O})$ make this explicit. For $i \in \mathcal{A}$ the measurements can drive π_i toward 0 or 1 as evidence accumulates; for $i \notin \mathcal{A}$ the marginal reduces to $\pi_i = p(s_i = 1 \mid \mathbf{s}_{\mathcal{A}})$, which the measurements cannot move, and its variance converges to a prior-conditional floor rather than to zero. A deterministic regressor, which must commit a value to every cell, cannot represent this distinction and produces confident predictions on cells the measurements never constrained; a posterior can, which is what makes it the appropriate object here. Since the likelihood $p(\mathcal{O}|\mathbf{s})$ is defined only implicitly through ray tracing and is unavailable in closed form, we learn the posterior directly from paired physically consistent data $(\mathbf{s}, \mathcal{O}) \sim \mathcal{D}$, as described next.

3 Design and Methodology

Our objective is to learn the posterior over scenes given the measurements a network already collects. The difficulty is not only that the posterior is high-dimensional, but that a model trained directly on raw channels would have to solve two problems at once. It would have to discover, from finite data, how geometry writes itself into a channel — which structures produce which delays, which angle of arrivals, which signal blockage — and simultaneously learn to generate a coherent scene consistent with that mapping. The first of these is the physics of propagation and is shared across every link; the second is the inference problem specific to a given set of observations. Entangling them enforces supervision on rediscovering the underlying physics, and leaves the generative model to place mass over a combinatorially large space of grids with no guidance about where the plausible scenes lie. We therefore propose a hierarchical framework, illustrated in Fig. 1, that separates these two burdens. The first stage learns the shared physical characteristics once: a multi-modal foundation model is pretrained to aggregate channel and positional information into a representation that is explicitly aligned, during pretraining, with the scene structure that shaped the observation. Each measurement is thereby encoded into a summary of the scene properties that could have produced it. The second stage performs the inference: the encoded representations of many observations of the same environment are fused as the conditioning set of a diffusion model that learns the posterior over occupancy, and a consensus over its samples yields the reconstructed scene. Because conditioning is expressed in a representation space that is aligned with geometrical properties rather than raw channel information, the generative model just resolves which scenes are consistent with the measurements rather than first inverting propagation.

3.1 Physics-grounded representation learning

In this section, we discuss the design and implementation of PHOENIX (a Physics-grounded Heterogeneous Observation Embedding for Neural Inference across wireless conteXts), a multi-modal foundation model that learns a unified representation of the physical information governing wireless propagation. The overall pipeline of PHOENIX is illustrated in Fig. 2.

At its core, PHOENIX employs a single-stream multi-modal transformer-based feature embedding framework to learn global, robust representations of the wireless channel by capturing cross-modality associations among the input data covering the wireless medium, the transmitter-receiver positioning, and the spatial-temporal characteristics of the received signal. Specifically, the channel tensor in the input is converted into amplitude and phase and tokenized over antenna pairs and neighboring subcarriers; the height map is partitioned into non-overlapping patches; and the receiver location is projected to a single token. Each token additionally encodes its physical origin through antenna/subcarrier indices or normalized spatial coordinates, together with a modality embedding. A dedicated [PHYSC] token aggregates information across modalities under an asymmetric attention mask: it attends to all modality tokens but is excluded as a key from the remaining tokens.

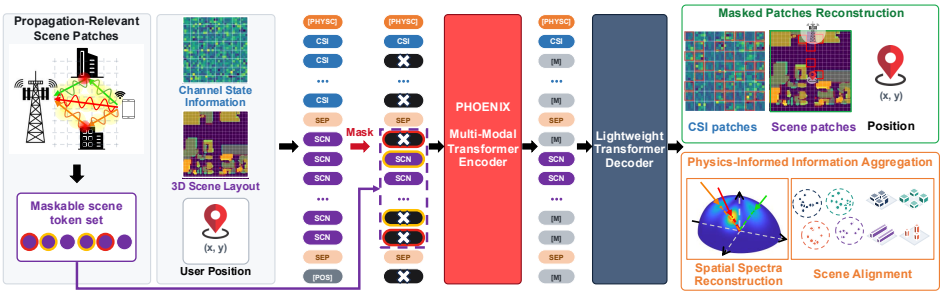

Figure 2: PHOENIX self-supervised pretraining: enabling physics-grounded representation learning over physical-layer wireless systems.

The [PHYSC] token is designed to aggregate information across modalities into a shared, physically grounded representation that reflects the underlying wireless propagation structure. The key question is therefore how to ensure that this representation remains consistent with the observed wireless propagation and underlying physical correspondence across modalities. Architectural and tokenization details are provided in Appendix C.

Physics-guided self-supervised pretraining. What makes the resulting representation physically meaningful is the pretraining objective. Conventional contrastive image–text style objectives (Radford et al., 2021) are a poor fit here, as different scenes and user positions can produce similar channels through different combinations of propagation paths, so matching and non-matching channel–scene pairs are not well defined. We instead pretrain with two complementary mechanisms.

The first mechanism is *bidirectional reconstruction*, which learns cross-modal physical associations by reconstructing masked information in one modality from the others: masked complex channel values are recovered from the remaining channel tokens together with scene and location information, while masked regions of the building map are recovered from the observed channel. Crucially, scene masking is not applied uniformly. Because a wireless observation constrains only the environment reached by its propagation paths, asking the model to infer regions that left no trace in the channel would train it to hallucinate unsupported geometry. We therefore identify the scene patches intersected by the collective propagation paths of a given link and mask *only within* this propagation-relevant set. This encourages the model to associate channel structure with the scene geometry that actually shaped the observation. To capture dominant propagation effects, including blockage, reflection, and diffraction, without overfitting fine-grained geometry, the reconstruction target is limited to coarse occupancy at each masked location.

The second is *physics-guided cross-modal distillation*, which shapes the [PHYSC] token. We train it to reconstruct a physical quantity jointly determined by the environment, the link geometry, and the observed channel: the *spatial spectrum* which describes how signal energy is distributed over propagation directions and reveals the dominant angles of departure and arrival. Since those directions are determined by the propagation medium and the link geometry, anchoring [PHYSC] to them ties its representation to physically interpretable propagation characteristics that remain meaningful across environments. We further align [PHYSC] with the part of the scene that produced it: the encoded tokens of the propagation-relevant scene patches, the structures that block, reflect, or diffract this particular link, are pooled into a scene representation, and [PHYSC] is encouraged to match it, while representations of unrelated links are pushed apart, with the separation relaxed for spatially nearby links that plausibly interact with the same structures. This alignment is the mechanism on which the rest of the paper depends: it is what makes $\mathbf{z}_k$ a summary of the scene structure responsible for the observation. Further pretraining details are provided in Appendix D.

After pretraining, we freeze f_ϕ and use the resulting [PHYSC] embedding $\mathbf{z}_k = f\phi(H_k, \mathbf{p}_k)$ as the compact representation of observation k, collecting $\mathcal{Z} = \{\mathbf{z}_k\}_{k=1}^{K}$ for each scene. These propagation-grounded representations serve as the basis for reconstructing the target environment.

3.2 Diffusion Posterior over Scenes

The physics-grounded representations provide a compact description of what each wireless observation reveals about the environment. The remaining challenge is to combine these partial and potentially ambiguous views into a posterior over complete scene configurations. We therefore model $p_\theta(\mathbf{s}|\mathcal{Z})$ directly, with the second stage learning the posterior by minimizing $\mathbb{E}(\mathbf{s}, \mathcal{O}) \sim \mathcal{D}\big[-\log p\theta(\mathbf{s} \mid \mathcal{Z})\big]$.

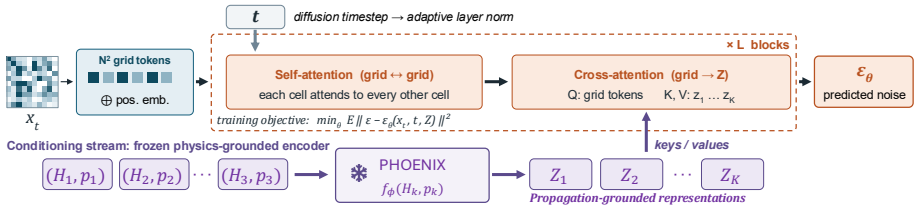

Figure 3: Posterior diffusion training over physics-grounded representations from wireless foundation model allowing conditional denoising of individual cells based on observed measurements.

Parameterization. Although the target scene layout is expressed in a binary format, we use a continuous relaxation to enable diffusion over the occupancy grid. Specifically, we represent the scene using analog bits, $\mathbf{x}_0 = 2\mathbf{s} - \mathbf{1} \in \{-1, +1\}^{N^2}$, and recover the binary occupancy through thresholding, $\hat{\mathbf{s}} = \mathbf{1}[\mathbf{x}_0 > 0]$. We then define a forward diffusion process that progressively perturbs $\mathbf{x}_0$ with Gaussian noise following the standard denoising diffusion formulation Ho et al. (2020), using the cosine noise schedule $\bar{\alpha}_t$ (Nichol & Dhariwal, 2021):

$$q(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}\big(\mathbf{x}_t;\ \sqrt{\bar{\alpha}_t}\mathbf{x}_0,\ (1 - \bar{\alpha}_t)\mathbf{I}\big)\,, \qquad \mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t}\boldsymbol{\epsilon}, \tag{3}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. This closed-form parameterization allows us to obtain a noisy scene $\mathbf{x}_t$ directly from the clean scene at any diffusion timestep.

To reverse this corruption, the denoiser $\boldsymbol{\epsilon}_\theta$ predicts the noise added to the scene while conditioning on the propagation-grounded representations $\mathcal{Z}$. It is implemented as a Transformer over N^2 grid tokens, one per cell, each combining the noisy occupancy value at that cell with a positional embedding of its location relative to the base station. Every block applies self-attention among the grid tokens, with the diffusion timestep injected through adaptive layer normalization with zero-initialized gates, followed by cross-attention in which the grid tokens are queries and $\mathcal{Z}$ supplies keys and values. Cross-attention suits this conditioning structure for two reasons: it is permutation invariant and accepts a variable number of observations, so one model serves any K without retraining; and since each $\mathbf{z}_k$ encodes the scene structure that shaped observation k, cross-attention allows each grid cell to draw selectively on the observations carrying evidence about that part of the scene. Fig. 3 illustrates the overall framework, including the conditioning and denoising process.

Training objective. We optimize the denoiser using the standard noise-prediction objective for diffusion models. For each training example, we sample a diffusion timestep t and Gaussian noise $\boldsymbol{\epsilon}$, construct the corresponding noisy scene $\mathbf{x}_t$, and train the model to recover the injected noise:

$$\theta^\star = \arg\min_\theta\ \mathbb{E}_{(\mathbf{s},\mathcal{O})\sim\mathcal{D}}\,\mathbb{E}_{t\sim\mathcal{U}\{1,\dots,T\},\,\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\left\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta\left(\mathbf{x}_t,\ t,\ \mathcal{Z}\right)\right\|_2^2\right], \tag{4}$$

where $\mathbf{x}_t$ is obtained from $\mathbf{x}_0 = 2\mathbf{s} - \mathbf{1}$ and $\boldsymbol{\epsilon}$ through equation 3, and $\mathcal{Z} = f_\phi(\mathcal{O})$ is computed by the frozen PHOENIX encoder. We update only the diffusion parameters θ. Architecture details and the noise schedule are provided in Appendix E.

Readout by consensus. At inference, we draw M independent scenes from the learned posterior by ancestral sampling, with each trajectory initialized from Gaussian noise and driven by an independent noise sequence. Each sample is then binarized to obtain an occupancy grid $\hat{\mathbf{s}}^{(m)} \sim p_\theta(\cdot \mid \mathcal{Z})$. We estimate the posterior marginal for each cell by averaging across samples and obtain the final reconstruction by thresholding this marginal:

$$\hat{\pi}_i = \frac{1}{M}\sum_m \hat{s}_i^{(m)}, \qquad \hat{s}_i^{\mathrm{cons}} = \mathbf{1}\left[\hat{\pi}_i > \tfrac{1}{2}\right]. \tag{5}$$

This readout follows directly from the posterior formulation. Proposition 3 in Appendix F shows that $\hat{\pi}_i$ concentrates around the learned posterior marginal $\pi_i^\theta = p_\theta(s_i = 1 \mid \mathcal{Z})$ at a Hoeffding rate. Consequently, $\hat{s}_i^{\mathrm{cons}}$ converges to the per-cell Bayes-optimal decision under p_θ, where the number of samples required for reliable recovery at a cell scales inversely with the squared posterior margin $|\pi_i^\theta - \frac{1}{2}|^2$. Proposition 4 shows that, when the posterior is sufficiently close to the underlying distribution, this decision recovers the true scene on the observable set $\mathcal{A}$, with additional measurements improving prediction on cells they constrain, while cells outside $\mathcal{A}$ remain governed by the conditional prior. Thus, consensus converts the learned posterior into a deterministic scene estimate.

An instructive analogy to our framework is unCLIP (Ramesh et al., 2022), which also inverts a frozen representation with a diffusion decoder, conditioning on a CLIP image embedding to generate an image. Our framework follows the same high-level paradigm, but the role of sampling

variability differs fundamentally. In unCLIP, the embedding specifies semantic content, while diffusion stochasticity supplies visual detail the embedding leaves undetermined, making diversity desirable. In our setting, a single physical scene underlies the observations, and variation across posterior samples reflects genuine ambiguity that the measurements leave unresolved. We therefore seek the structure consistently supported by the observations rather than diverse scenes, which is what the consensus readout in equation 5 provides: it retains footprints supported across samples and suppresses structure arising from unresolved uncertainty.

4 Experiments

We pretrain the foundation model with the proposed physics-guided self-supervised objective on a digital-twin corpus of 11,000 outdoor environments and 30 million ray-traced channel observations. The pretrained model then provides physics-grounded representations that condition the posterior diffusion model in our hierarchical pipeline, a configuration we refer to as **Diffusion + FM**. To provide an interpretable alternative to the diffusion formulation and isolate the contribution of joint posterior modeling, **Marginal Prob. + FM** replaces the diffusion posterior with a deterministic per-observation decoder: each observation yields a cell-wise occupancy belief and a confidence score marking the propagation-relevant region over which that belief is meaningful. The predicted beliefs are then fused into the final reconstruction. To our knowledge, prior work does not directly address scene reconstruction from *communication-infrastructure data* and without geometric priors, precluding a direct head-to-head comparison; we therefore use controlled baselines to isolate each component of our design. **Diffusion + Raw Meas.** conditions the identical diffusion model on raw observations and user locations, testing whether the scene posterior can be learned without the intermediate scene-aligned representation; this is closest in spirit to (Hehn et al., 2025), which reconstructs indoor geometry from a single link's received power with the assumption of knowing floor and ceiling geometry. **Marginal Prob. + Raw Meas.** predicts occupancy over the full grid from each measurement under a binary cross-entropy objective and aggregates the per-measurement predictions, without modeling the observation-specific propagation-relevant region. Training objectives and implementation details for all models are given in Appendices G–I. All models are evaluated on $\sim$700 environments held out from every stage of training, spanning layouts from dense urban blocks to open areas with isolated structures. Each observation consists of the full wireless channel collected by a single base station together with the user's position. We report balanced accuracy, F1, and intersection-over-union (IoU) of the reconstructed occupancy grid against ground truth. Appendix G details the array and waveform configuration, Appendix J defines the metrics, and Appendices L–M study the effect of incomplete/noisy position information.

4.1 Main Results

Figure 4 presents scene reconstruction produced by our framework on previously unseen representative environments. Given 500 wireless channel measurements from a single base-station node, our framework recovers the surrounding building footprints, their location, extent, and arrangement, from the wireless channel observations and user position, without any prior over scene information. The reconstructions closely follow ground-truth layouts across environments with substantially different spatial density and structure. *To our knowledge, this is the first demonstration that communication-infrastructure channel measurements can support reconstruction of arbitrary outdoor scene geometry without assumptions on the number, shape, or placement of structures.* By performing posterior inference over physics-grounded representations, our framework makes this ill-posed inverse problem tractable and allows existing radio links to act as environmental sensors. The scope of what is recovered is set by the measurements: with a ULA at the base station, the channel resolves paths only in azimuth, so building height is severely underspecified; hence we reconstruct the scene as a two-dimensional occupancy grid of building footprints. A three-dimensional extension using a uniform rectangular array that can resolve paths in elevation is left for future work.

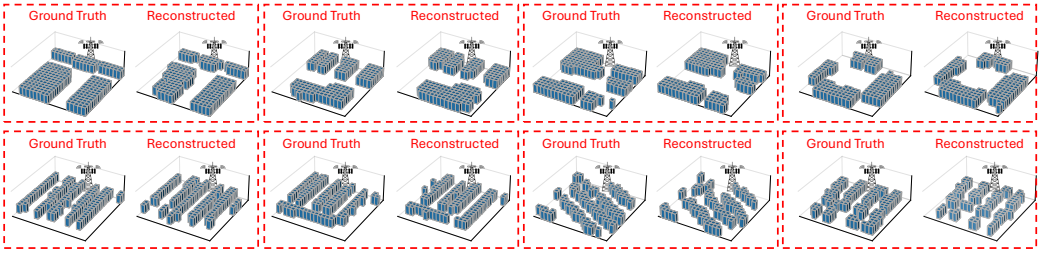

Figure 4: Scene reconstruction from wireless measurements across spatially distinct environments.

Table 1: Scene reconstruction performance across different models.

Model	Balanced Accuracy ↑			F1 ↑			IoU ↑		
	Mean	Median	P90	Mean	Median	P90	Mean	Median	P90
Diffusion + Raw Meas.	0.505	0.500	0.519	0.054	0.016	0.162	0.030	0.008	0.088
Marginal Prob. + Raw Meas.	0.621	0.618	0.668	0.397	0.398	0.503	0.251	0.248	0.336
Marginal Prob. + FM (Ours)	0.738	0.739	0.803	0.633	0.642	0.737	0.470	0.473	0.584
Diffusion + FM (Ours)	**0.758**	**0.746**	**0.868**	**0.663**	**0.655**	**0.810**	**0.506**	**0.487**	**0.680**

Table 1 summarizes the reconstruction performance. Across all metrics, our framework achieves the strongest performance, substantially outperforming both raw-measurement baselines and improving on the marginal-probability variant. To the best of our knowledge, these are the strongest scene reconstruction results reported from communication-infrastructure measurements, and substantially improve on prior attempts at this problem. Several conclusions follow from the results.

The comparison with *Diffusion + Raw Meas.* isolates the benefit of the physics-grounded representation. The decoder, noise schedule, and sampler are otherwise identical, yet conditioning on raw channels collapses to near-chance performance. This indicates a failure of learnability rather than information: raw channels contain at least as much scene information as their learned representations, but the decoder must simultaneously discover the propagation-to-geometry mapping and perform posterior generation from a high-dimensional input. In contrast, the foundation model provides scene-aligned representations that substantially simplify posterior inference.

Marginal Prob. + Raw Meas. fails for a different reason. Its full-grid cross-entropy requires every observation to predict every scene cell, even though a measurement is informative only about propagation-relevant regions. The resulting overconfident predictions in uninformed regions accumulate during fusion, where observations that never sensed a cell still contribute votes to it. *Marginal Prob. + FM* uses the same fusion but forms each per-observation belief on the physics-grounded representation and around the propagation-relevant cells of that observation, abstaining elsewhere; each vote is therefore informative where it is cast, and the aggregation better recovers the scene. Our proposed design improves over this baseline by **22.1%** in balanced accuracy and **67.0%** in F1 score.

Lastly, the results demonstrate that the joint posterior outperforms fused marginals in this setting. The main advantage is concentrated in the upper tail of the per-scene distribution: the median IoU of the two models differs by three percent relative, while at the 90th percentile the gap widens to sixteen percent. The marginal variant treats observations independently and combines them through a fixed rule: confidence gating discards evidence that is weak in each observation but informative in combination, and nothing in the fusion enforces that the result be a plausible scene. The diffusion decoder instead learns the joint posterior over the entire grid given the full set of observations. It fuses evidence in representation space, and it completes the scene through the learned prior where the measurements are silent, the second factor in equation 2. Additional illustrative scene reconstruction examples comparing all four methods across diverse environments are provided in Appendix N.

4.2 Physical Interpretability of the Learned Beliefs

We examine the physical interpretability of the per-observation marginal predictions using *Marginal Prob. + FM*, which exposes a belief for each measurement individually. Specifically, we ask whether each belief captures the environment regions implicated by its propagation paths while remaining uncertain elsewhere. Figure 5 visualizes the marginal occupancy probability for two measurements in the same environment, together with the ground-truth layout and corresponding propagation paths. The belief recovers the free-space corridors along the direct and reflected paths, the reflecting structures, and the obstructing component when the direct path is blocked. Elsewhere, predictions remain near one-half, indicating uncertainty in regions not informed by the measurement. What the beliefs recover is therefore not an arbitrary partial reconstruction but the specific geometry that is physically responsible for shaping the observation, leaving unrelated regions unknown. This behavior, induced by the physics-guided pretraining, makes the marginal predictions physically interpretable: each belief can be read back as a statement about which structures the link encountered.

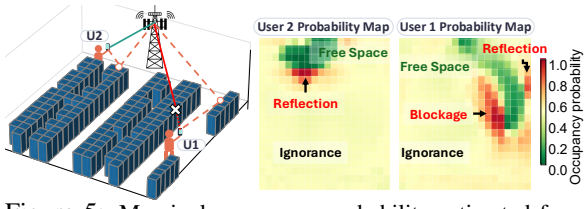

Figure 5: Marginal occupancy probability estimated from wireless measurements.

4.3 Performance Across Scene Density

Scene density governs the complexity of the propagation conditions and with it the difficulty of the inverse problem: in sparse environments most links propagate through free space, and each measurement admits a narrow family of consistent scenes, whereas blockage, reflection, and higher-order interaction widen that and force the posterior to reconcile many such families at once. We therefore partition the test environments by occupied fraction, spanning 1–70% across test set, into three equal-width bands: sparse, moderate, and dense, and report performance within each in Table 2.

Table 2: Mean reconstruction performance by scene density.

	Sparse (88 scenes)			Moderate (565 scenes)			Dense (18 scenes)		
Model	Bal. Acc. ↑	F1 ↑	IoU ↑	Bal. Acc. ↑	F1 ↑	IoU ↑	Bal. Acc. ↑	F1 ↑	IoU ↑
Diffusion + Raw Meas.	0.518	0.066	0.039	0.503	0.048	0.027	0.508	0.179	0.105
Marginal Prob. + Raw Meas.	0.647	0.434	0.283	0.617	0.393	0.248	0.589	0.316	0.190
Marginal Prob. + FM (Ours)	0.764	0.631	0.468	0.735	0.635	0.471	0.712	0.607	0.457
Diffusion + FM (Ours)	**0.786**	**0.665**	**0.510**	**0.752**	**0.659**	**0.501**	**0.806**	**0.775**	**0.652**

Diffusion + FM is best in every regime, and its advantage is most pronounced for dense scenes. Both marginal estimators degrade monotonically in balanced accuracy as density increases, whereas *Diffusion + FM* remains robust across the three bands; relative to *Marginal Prob. + FM*, its balanced-accuracy advantage grows from 2.9% to 13.2% and its IoU advantage from 9.0% to 42.7%. This is consistent with how the two combine evidence: denser environments leave each observation informative about a smaller portion of the scene, and while the marginal formulation must aggregate independently estimated beliefs under this incomplete evidence, the diffusion model infers the scene jointly and completes regions where direct evidence is limited through the learned prior.

4.4 Effect of the Number of Wireless Measurements

Figure 6 reports reconstruction quality for 50 to 1,000 measurements per scene, showing how additional observations provide complementary information. Since both diffusion decoders are trained with $K \in [100, 500]$, the extreme sweep points also test generalization in the number of observations.

Diffusion + FM and the marginal-probability variants improve with more measurements before saturating. Additional observations help in two ways: new links, particularly new reflected paths, reach cells no earlier observation touched and enlarge the observable set, while additional coverage of observed cells sharpens the posterior there, as described by Proposition 4. Saturation follows from the same analysis, since once the users of a single base station have illuminated most of the scene its links can reach, further observations revisit covered cells and the remaining errors concentrate on structures no link reaches. *Diffusion + Raw Meas.* instead stays near chance throughout: without a representation exposing the scene-relevant structure of the observations, more raw data does not become usable evidence.

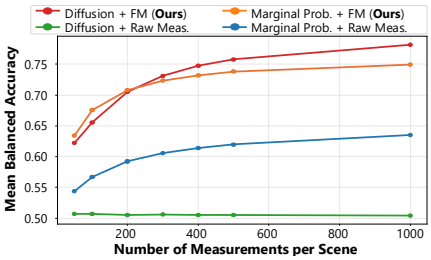

Figure 6: Reconstruction accuracy as a function of collected measurements per scene.

The two foundation-model approaches scale differently. *Marginal Prob. + FM* is comparable to *Diffusion + FM* under few or moderately many observations, as both rely on the same representations under similarly incomplete evidence, but is increasingly surpassed as measurements accumulate. The value of modeling the joint posterior thus emerges once enough complementary observations are available to integrate, reflecting the complementary roles of our two components.

5 Conclusion

We presented a hierarchical framework for reconstructing environmental geometry from wireless channel measurements by combining physics-grounded representation learning with joint generative posterior inference. A wireless foundation model first extracts scene-relevant information from individual observations, while a diffusion model jointly integrates these representations to infer the posterior over scene occupancy, from which a per-cell consensus over sampled scenes recovers the building footprints. Results demonstrate our framework can recover previously unseen scene structures from communication signals, and controlled comparisons show the importance of our hierarchical design for effectively solving the underlying ill-posed inverse problem. The results highlight the potential of existing wireless infrastructure to serve for environmental perception.

AI USE STATEMENT

We used generative AI tools as a general-purpose writing and coding assistant. Specifically, they were used to polish the phrasing of text drafted by the authors, and to assist with the implementation of experiment and figure-generation scripts. We did not use generative AI tools for research ideation, for deriving the theoretical results, or for retrieving or summarizing related work. All AI-assisted output was reviewed by the authors: text was checked for accuracy against the underlying results, all AI-assisted code was read, executed, and verified against the reported numbers, and the theoretical claims, their assumptions, and their proofs were derived and checked by the authors. We take responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We have aimed to make every component of this work reproducible. The problem formulation, the assumptions underlying each theoretical claim, and complete proofs of propositions are given in Appendices B and F. Dataset construction, including the simulation configuration, the derivation of occupancy grids, and the extraction of propagation-relevant sets, is described in Appendices A and G. Architecture and tokenization details for the foundation model are given in Appendix C, and the complete pretraining objective in Appendix D. The diffusion decoder, its noise schedule, and its sampler are specified in Appendix E, and full training configurations — optimizer, learning rate, schedule, batch size, and hardware — for the proposed pipeline, the ablated variant, and both baselines are given in Appendices G, H, and I. Evaluation metrics are defined in Appendix J, and the evaluation protocol, including how observation subsets are drawn across the measurement sweep, is described in Appendix G. Source code for training and evaluation, together with the pretrained dataset and checkpoints, will be provided upon publication.

REFERENCES

Ian F Akyildiz, Ahan Kak, and Shuai Nie. 6G and beyond: The future of wireless communications systems. *IEEE Access*, 8:133995–134030, 2020.

Anatoly Bakushinsky and A Goncharsky. *Ill-posed problems: theory and applications*. Springer Science & Business Media, 2012.

David Borts, Julian Ost, Shamik Basu, Tim Broedermann, Andrea Ramazzina, Christos Sakaridis, Mario Bijelic, and Felix Heide. Carafe: Camera-radar radiance fields for scene reconstruction. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pp. 1–12, 2026.

Eugene Casmin and Rodolfo Oliveira. Survey on context-aware radio frequency-based sensing. *Sensors*, 25(3):602, 2025.

Umut Demirhan and Ahmed Alkhateeb. Integrated sensing and communication for 6G: Ten key machine learning roles. *IEEE Communications Magazine*, 61(5):113–119, 2023.

Laura Dodds, Maisy Lam, Waleed Akbar, Yibo Cheng, and Fadel Adib. Wave-former: Through-occlusion 3d reconstruction via wireless shape completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21713–21724, 2026.

Ivan Dokmanić, Reza Parhizkar, Andreas Walther, Yue M Lu, and Martin Vetterli. Acoustic echoes reveal room shape. *Proceedings of the National Academy of Sciences*, 110(30):12186–12191, 2013.

Yara El Baba, Andreas Walther, and Emanuël A P Habets. 3D room geometry inference based on room impulse response stacks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(5):857–872, 2018.

Thomas Hehn, Markus Peschl, Tribhuvanesh Orekondy, Arash Behboodi, and Johann Brehmer. Differentiable and learnable wireless simulation with geometric transformers. In *International Conference on Learning Representations (ICLR)*, 2025.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Jakob Hoydis, Fayçal Aït Aoudia, Sebastian Cammerer, Merlin Nimier-David, Nikolaus Binder, Guillermo Marcus, and Alexander Keller. Sionna RT: Differentiable ray tracing for radio propagation modeling. In *IEEE Globecom Workshops (GC Wkshps)*, pp. 317–321, 2023.

Haowen Lai, Zitong Lan, and Mingmin Zhao. Non-line-of-sight 3d reconstruction with radar. *Advances in Neural Information Processing Systems*, 38:72844–72869, 2026.

Yiming Li, Zeyu Li, Zhihui Gao, and Tingjun Chen. Geo2SigMap: High-fidelity RF signal mapping using geographic databases. In *IEEE International Symposium on Dynamic Spectrum Access Networks (DySPAN)*, pp. 277–285, 2024.

Chris Xiaoxuan Lu, Stefano Rosa, Peijun Zhao, Bing Wang, Changhao Chen, John A Stankovic, Niki Trigoni, and Andrew Markham. See through smoke: Robust indoor mapping with low-cost mmWave radar. In *Proceedings of the International Conference on Mobile Systems, Applications, and Services (MobiSys)*, pp. 14–27, 2020.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, pp. 2579–2605, 2008.

Krista Merry and Pete Bettinger. Smartphone gps accuracy study in an urban environment. *PloS one*, 14(7):e0219890, 2019.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning (ICML)*, pp. 8162–8171, 2021.

OpenStreetMap contributors. OpenStreetMap. `https://www.openstreetmap.org`, 2023.

Joshua Roy Palathinkal, Hardani Ismu Nabil, Muhammad Iqbal Rochman, Hossein Nasiri, Francis A Gatsi, and Monisha Ghosh. Evaluating smartphone gnss accuracy for geofenced 6 ghz operations. In *2026 IEEE International Symposium on Spectrum Innovation (DySPAN)*, pp. 1–6. IEEE, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint arXiv:2204.06125*, 2022.

Yifei Sun, Jie Li, Tong Zhang, Rui Wang, Xiaohui Peng, Xiao Han, and Haisheng Tan. An indoor environment sensing and localization system via mmwave phased array. *Journal of communications and information networks*, 7(4):383–393, 2022.

Paolo Testolina, Michele Polese, Pedram Johari, and Tommaso Melodia. Boston Twin: The Boston digital twin for ray-tracing in 6G networks. In *Proceedings of the ACM Multimedia Systems Conference (MMSys)*, pp. 441–447, 2024.

Jiacheng Wang, Hongyang Du, Dusit Niyato, Jiawen Kang, Shuguang Cui, Xuemin Shen, and Ping Zhang. Generative AI for integrated sensing and communication: Insights from the physical layer perspective. *IEEE Wireless Communications*, 31(5):246–255, 2024.

Inmo Yeon and Jung-Woo Choi. RGI-Net: 3D room geometry inference from room impulse responses with hidden first-order reflections. In *IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 439–443, 2024.

Inmo Yeon, Iljoo Jeong, Seungchul Lee, and Jung-Woo Choi. EchoScan: Scanning complex room geometries via acoustic echoes. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

Mingmin Zhao, Tianhong Li, Mohammad Abu Alsheikh, Yonglong Tian, Hang Zhao, Antonio Torralba, and Dina Katabi. Through-wall human pose estimation using radio signals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7356–7365, 2018.

Xiaopeng Zhao, Zhenlin An, Qingrui Pan, and Lei Yang. NeRF^2: Neural radio-frequency radiance fields. In *Proceedings of the Annual International Conference on Mobile Computing and Networking (MobiCom)*, 2023.

Kaichen Zhou, Laura Dodds, Sayed Saad Afzal, and Fadel Adib. RISE: Single static radar-based indoor scene understanding. *arXiv preprint arXiv:2511.14019*, 2025.

A Forward model details

Array responses. For a uniform linear array of N elements with spacing d at wavelength λ, the array response at angle ϑ is

$$a(\vartheta) \;=\; \Big[1,\; e^{-j\frac{2\pi d}{\lambda}\sin\vartheta},\; \ldots,\; e^{-j\frac{2\pi d(N-1)}{\lambda}\sin\vartheta}\Big]^{\top}. \tag{6}$$

Substituting equation 6 into the path sum of the main text gives the channel entry for each transmit–receive pair and subcarrier. The angular resolution used below follows from the Rayleigh criterion for this aperture, $\Delta\theta \approx 2/N_{tx}$ radians near broadside, and the delay resolution is $\Delta\tau = 1/B$ for occupied bandwidth B.

Path generation. The path set $\mathcal{P}(\mathbf{s}, \mathbf{p})$ is produced by ray tracing the scene geometry for the transmitter–receiver pair. Each path is characterized by its interaction sequence with the structures of the scene: the direct path when it exists, first- and higher-order specular reflections off building facades, and diffractions at building edges. The gain α_ℓ aggregates free-space path loss with the reflection and diffraction coefficients along the interaction sequence.

Propagation-relevant set. For observation k we define

$$a_k \;=\; \big\{\, i \;:\; \mathrm{cell}_i \cap \ell \neq \emptyset \;\text{ for some } \ell \in \mathcal{P}(\mathbf{s}, \mathbf{p}_k) \,\big\}, \qquad A \;=\; \bigcup_{k=1}^{K} a_k, \tag{7}$$

recording, for every path, the grid cells traversed by its segments together with the cells at which it interacts with a structure. We additionally record the first-order neighborhood ∂a_k of these cells, used to define the boundary zone in the analysis.

B Ill-posedness of the inverse problem

This appendix formalizes the three sources of non-uniqueness described in the main text.

B.1 Exact invariance outside the propagation-relevant set

By the definition equation 7, cells outside a_k do not participate in any path of observation k, so the likelihood of that observation does not depend on them; aggregating over the observation set gives the invariance stated in the main text. Its consequence for the number of scenes consistent with a measurement is the following.

Proposition 1 (Size of the consistent family)**.** *Fix an observation k and let $\mathbf{s}^\star$ be the true scene. Let $\mathcal{S}_k(H_k) = \{\mathbf{s} : \mathcal{F}(\mathbf{s}, \mathbf{p}_k) = H_k\}$ be the set of scenes reproducing the noiseless measurement exactly. Then $|\mathcal{S}_k(H_k)| \geq 2^{|a_k^c|}$.*

Proof. By the definition of a_k, no cell in a_k^c participates in any path of $\mathcal{P}(\mathbf{s}^\star, \mathbf{p}_k)$; the path parameters $(\alpha_\ell, \tau_\ell, \theta_\ell, \phi_\ell)$ therefore do not depend on $\mathbf{s}_{a_k^c}$, and neither does the path sum. Hence for any subset $\mathcal{J} \subseteq a_k^c$ the scene $\mathbf{s}^\star \oplus \mathbf{e}_\mathcal{J}$, obtained by flipping the cells in $\mathcal{J}$, satisfies $\mathcal{F}(\mathbf{s}^\star \oplus \mathbf{e}_\mathcal{J}, \mathbf{p}_k) = H_k$ and lies in $\mathcal{S}_k(H_k)$. Distinct subsets give distinct scenes, and there are $2^{|a_k^c|}$ of them. □

The inequality is strict in general, since the count enumerates only scenes differing from $\mathbf{s}^\star$ outside a_k; the ambiguities of Sections B.2 and B.3 contribute further members that differ *inside* a_k. Aggregating over the observation set replaces a_k^c by A^c, giving $|\mathcal{S}(\mathcal{O})| \geq 2^{|A^c|}$: the residual non-uniqueness is exponential in the number of cells no path touches.

B.2 Finite resolution

Within A the map from geometry to channel remains many-to-one. A measurement of bandwidth B resolves path delays only to $\Delta\tau = 1/B$, and an aperture of N_{tx} elements resolves angles only to $\Delta\theta \approx 2/N_{tx}$. If two scenes induce path parameter sets that agree within these limits,

$$\big\{(\tau_\ell, \theta_\ell, \phi_\ell, \alpha_\ell)\big\}_{\mathbf{s}} \;\equiv\; \big\{(\tau_\ell, \theta_\ell, \phi_\ell, \alpha_\ell)\big\}_{\mathbf{s}'} \quad \text{within } (\Delta\tau, \Delta\theta), \tag{8}$$

then $\mathcal{F}(\mathbf{s}, \mathbf{p}_k) \approx \mathcal{F}(\mathbf{s}', \mathbf{p}_k)$ and the two scenes cannot be told apart from observation k. This source of ambiguity shrinks as bandwidth and aperture grow.

B.3 Propagation-Equivalent Scenes

By the path-sum forward model, the channel of link k depends on the scene only through the path set it induces, $\mathcal{P}(\mathbf{s}, \mathbf{p}_k)$, and its parameters. Two scenes that induce the same path parameters for every link are therefore indistinguishable from $\mathcal{O}$, regardless of how they differ as occupancy grids. We call such scenes *propagation-equivalent*.

Proposition 2 (Propagation-equivalent scenes are indistinguishable)**.** *If two scenes $\mathbf{s}, \mathbf{s}'$ satisfy $\mathcal{P}(\mathbf{s}, \mathbf{p}_k) = \mathcal{P}(\mathbf{s}', \mathbf{p}_k)$ for a given observation, in the sense that the two path sets carry identical parameters $(\alpha_\ell, \tau_\ell, \theta_\ell, \phi_\ell)$, then $\mathcal{F}(\mathbf{s}, \mathbf{p}_k) = \mathcal{F}(\mathbf{s}', \mathbf{p}_k)$, no estimator can distinguish them from $\mathcal{O}$, irrespective of bandwidth and aperture.*

The channel entry of link k is a sum over $\ell \in \mathcal{P}(\mathbf{s}, \mathbf{p}_k)$ of terms depending only on $(\alpha_\ell, \tau_\ell, \theta_\ell, \phi_\ell)$. Equality of the parameterized path sets therefore implies equality of the path sums and of the likelihoods $p(H_k \mid \cdot)$ for that observation.

The content of this statement lies in how large the set of propagation-equivalent scenes is. The map from occupancy to path parameters is many-to-one, and it is so for reasons that are geometric rather than instrumental. A cell influences a link only through the *role* it plays for that link,

$$\rho_k(i) = \rho\big(i, \mathbf{p}_k, \mathbf{s}\big) \in \{\texttt{block}, \texttt{reflect}, \texttt{free}, \texttt{none}\}, \tag{9}$$

where `block` denotes an occupied cell that removes a path that would otherwise reach the receiver, `reflect` an occupied cell that supports a specular or diffracted interaction contributing a path, `free` an unoccupied cell traversed by a contributing path, and `none` a cell that neither carries nor obstructs any contributing path. Additionally, $\rho_k(i)$ is a function of the link geometry, not of s_i alone: different occupancy value can provide the same role for a given link.

Aliasing of interaction sequences. Distinct interaction sequences can produce the same path parameters. A single specular reflection from one surface and consecutive reflections from two others can yield the same total delay and the same departure and arrival angles, in which case the two scenes contribute identical terms to the path sum. Here the roles themselves differ — different cells reflect — yet the scenes remain propagation-equivalent. This mechanism is exact when the geometry makes the parameters coincide, as symmetric layouts often do, and approximate otherwise, in which case it compounds with the finite-resolution ambiguity of Appendix B.2.

Occupancy does not imply blockage. A cell may be occupied and still fail to obstruct a link whose line-of-sight segment passes over it. The base station and the user sit at different heights, so the direct ray clears any structure shorter than its elevation at that point, and a two-dimensional occupancy grid records only that a building is present, not whether it is tall enough to intercept the ray. A cell can therefore belong to the propagation-relevant set of an observation — it lies along a path the link uses — while the measurement still carries no evidence against its being occupied: the unobstructed channel is equally consistent with free space and with a structure the ray passes above.

B.4 COLLECTIVE OBSERVATIONS

Under conditional independence of the measurements given the scene,

$$p(\mathbf{s} \mid \mathcal{O}) \;\propto\; p(\mathbf{s}) \prod_{k=1}^{K} p\big(H_k \mid \mathbf{s}\big), \tag{10}$$

so additional observations intersect the families $\mathcal{S}_k$ and sharpen the posterior. The intersection does not collapse to a point: by Proposition 1 the residual family has size at least $2^{|A^c|}$, and within A the ambiguities of Sections B.2 and B.3 persist. Substituting the invariance into Bayes' rule gives the factorization stated in the main text, whose second factor is independent of $\mathcal{O}$; consequently a cell $i \notin A$ has $\pi_i = p(s_i = 1 \mid \mathbf{s}_A)$, and as the data sharpen $p(\mathbf{s}_A \mid \mathcal{O})$ its variance converges to the prior-conditional variance rather than to zero.

C FOUNDATION MODEL ARCHITECTURE AND TOKENIZATION

Channel tokenization. The CSI tensor is converted from real and imaginary components to amplitude and phase $(a_{t,r,s}, \phi_{t,r,s})$. For each transmit antenna t, receive antenna r, and group g of n_{sc} neighboring subcarriers, one channel token is formed by concatenating the corresponding amplitude and phase values, $\mathbf{x}^{ch}_{t,r,g} = [\, a_{t,r,g,1:n_{sc}} \,\|\, \phi_{t,r,g,1:n_{sc}} \,] \in \mathbb{R}^{2n_{sc}}$, giving $N_{ch} = N_{tx}N_{rx}N_{sc}/n_{sc}$ tokens. Each is mapped to the shared embedding dimension d by a learnable linear projection followed by layer normalization. This preserves full transmit- and receive-antenna resolution, since each antenna pair carries distinct spatial information, while grouping subcarriers to represent local frequency structure compactly.

Scene and location tokenization. The height map $\mathbf{S} \in \mathbb{R}^{N \times N}$ is partitioned into non-overlapping $p \times p$ patches, flattened, and projected analogously, yielding $(N/p)^2$ scene tokens. The receiver location is projected by a single linear layer into one location token.

Physically grounded positional and modality embeddings. Tokenization captures content but not physical origin. For channel tokens we encode position separately along the transmit, receive, and subcarrier axes using sinusoidal encodings and concatenate them, $\mathbf{e}^{ch}_{t,r,g} = [\,\mathrm{PE}_{tx}(t) \,\|\, \mathrm{PE}_{rx}(r) \,\|\, \mathrm{PE}_{sc}(g)\,]$, preserving both the spatial structure of the arrays and the frequency structure of the channel. For scene patches we apply a two-layer MLP to sinusoidal features of the patch's normalized distance and bearing relative to the base station, so each token carries its geometric relationship to the reference position. A learnable modality embedding $\mathbf{m}^{(i)} \in \{\mathbf{m}^{ch}, \mathbf{m}^{scn}, \mathbf{m}^{loc}\}$ identifies the token type. The final input token is the sum of the three: what the token contains, where it comes from physically, and which modality it belongs to.

Encoder. Because channel tokens vastly outnumber scene and location tokens, lightweight pre-encoder blocks first operate on channel tokens alone, using rotary position embeddings across the (t, r, g) axes to capture local correlations across nearby antennas and frequencies before cross-modal fusion; absolute encodings are retained for the subsequent stage. All tokens are then concatenated into a single sequence prefixed by `[PHYSC]` and processed by a single-stream transformer with merged-token self-attention, so all modalities interact from the first layer. A binary attention mask enforces the asymmetry described in the main text: `[PHYSC]` is an unrestricted query over all tokens but is excluded as a key everywhere, so it can read from the sequence without any token reading from it.

D PRETRAINING OBJECTIVE IN FULL

Masked channel reconstruction. With $\mathcal{M}_{ch}$ the masked channel tokens,

$$\mathcal{L}_{ch} \;=\; \sqrt{\frac{1}{|\mathcal{M}_{ch}|} \sum_{i \in \mathcal{M}_{ch}} \big\|\hat{\mathbf{h}}_i - \mathbf{h}_i\big\|_2^2}. \tag{11}$$

Masked scene reconstruction. Masking is restricted to the propagation-relevant patches of the link. With $\mathcal{M}_{scn} \subseteq a_k$ the masked propagation-relevant patches and o_i the occupancy label,

$$\mathcal{L}_{scn} \;=\; -\frac{1}{|\mathcal{M}_{scn}|} \sum_{i \in \mathcal{M}_{scn}} \big[\, o_i \log \hat{o}_i + (1 - o_i) \log(1 - \hat{o}_i) \,\big]. \tag{12}$$

Restricting the mask to a_k is what prevents the model from being trained to infer regions that left no trace in the channel.

Masked location reconstruction. $\mathcal{L}_{loc}$ is the root-mean-square error between predicted and true receiver location.

Spatial-spectrum distillation. Let $H_n^{(\Psi)}$ denote slices of the CSI tensor along the transmit or receive antenna axis at a reference antenna and subcarrier. The angle-of-departure and angle-of-arrival spectra are

$$s_\Psi[b] \;=\; \Big| \sum_{n=0}^{N_\Psi - 1} H_n^{(\Psi)}\, e^{-j 2\pi n b / N_\Psi} \Big|^2, \qquad \Psi \in \{tx, rx\}, \tag{13}$$

with b indexing angular bins. The `[PHYSC]` token is trained to reconstruct these spectra under

$$\mathcal{L}_\Psi \;=\; D_{\mathrm{KL}}\big(\mathrm{softmax}(s_\Psi) \,\|\, \mathrm{softmax}(\hat{s}_\Psi)\big) \;+\; \lambda\, \mathrm{MSE}(\hat{s}_\Psi, s_\Psi), \tag{14}$$

where the KL term emphasizes alignment of spectral peak locations and the MSE term preserves their relative strengths.

PHYSC–scene contrastive alignment. The encoded tokens of the propagation-relevant patches are pooled into a scene representation z^{scn}, and

$$\mathcal{L}_{con} \;=\; -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp\big(z_i^{physc} \cdot z_i^{scn} / \tau\big)}{\sum_{j=1}^{N} w_{ij} \exp\big(z_i^{physc} \cdot z_j^{scn} / \tau\big)}, \tag{15}$$

with temperature τ and weights $w_{ij} \in (0, 1]$ that down-weight negative pairs drawn from spatially nearby links, preventing the objective from penalizing genuinely similar propagation geometries as hard negatives.

Pretraining objective. The five terms are combined as

$$\mathcal{L}_{\mathrm{pre}} \;=\; \lambda_{ch}\, \mathcal{L}_{ch} \;+\; \lambda_{scn}\, \mathcal{L}_{scn} \;+\; \lambda_{loc}\, \mathcal{L}_{loc} \;+\; \lambda_{spec}\big(\mathcal{L}_\Psi^{tx} + \mathcal{L}_\Psi^{rx}\big) \;+\; \lambda_{con}\, \mathcal{L}_{con}, \tag{16}$$

where $\mathcal{L}_{ch}$ and $\mathcal{L}_{loc}$ are reconstruction errors on the masked channel entries and the masked location, $\mathcal{L}_{scn}$ is a binary cross-entropy on the masked propagation-relevant scene patches, $\mathcal{L}_\Psi$ combines a Kullback–Leibler term that aligns the peaks of the predicted and true spatial spectra with a mean-squared term that preserves their relative strengths, and $\mathcal{L}_{con}$ is the distance-weighted contrastive alignment between `[PHYSC]` and the propagation-relevant scene representation.

Decoder and configuration. Following masked autoencoding, a lightweight transformer decoder receives the encoded visible tokens together with learnable tokens standing for the missing inputs and predicts the corresponding channel values, scene occupancy, and location. It is used only during pretraining and discarded afterwards.

E DIFFUSION DECODER DETAILS

Noise schedule. We use the cosine schedule, $\bar{\alpha}_t = f(t)/f(0)$ with $f(t) = \cos^2\!\big(\frac{t/T + \varsigma}{1+\varsigma} \cdot \frac{\pi}{2}\big)$ and offset $\varsigma = 0.008$, and $\beta_t = 1 - \bar{\alpha}_t / \bar{\alpha}_{t-1}$ clipped at 0.999 for numerical stability.

Denoiser. Each block comprises timestep-modulated self-attention over the grid tokens, plain cross-attention into the conditioning set, and a timestep-modulated feed-forward network. Modulation follows the adaptive layer normalization scheme in which the timestep embedding produces shift, scale, and gate parameters per block, with gates initialized to zero so that each block begins as an identity map. Cross-attention is left unmodulated, since a variable-size conditioning set does not fold naturally into a single global modulation vector the way a scalar timestep does. A linear adapter maps the encoder embedding dimension to the decoder width, and the output head is zero-initialized.

Conditioning dropout. During training the entire conditioning set is replaced by a learned null embedding with probability $p_{\text{drop}} = 0.15$. We use this as a regularizer against over-reliance on any particular conditioning configuration; at sampling time the conditional prediction is used directly, without guidance.

Sampler. We use ancestral sampling from $\mathbf{x}_T \sim \mathcal{N}(0, \mathbf{I})$. At each step the clean-scene estimate $\hat{\mathbf{x}}_0 = (\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, \mathcal{Z}))/\sqrt{\bar{\alpha}_t}$ is clamped to $[-1, 1]$ before forming the posterior mean, which stabilizes the trajectory for a target that is binary at $t = 0$. Each of the M trajectories is driven by independent noise and binarized by sign at the final step, which is the condition under which the concentration results of Appendix F apply.

F Proofs of the consensus propositions

In the following, we establish the propositions that characterize the theoretical properties of our consensus estimator. We first show that consensus under the learned posterior yields the Bayes-optimal per-cell decision. We then show that consensus from a near-optimal posterior recovers the true scene values on the observable set, with the reconstruction error decreasing as the number of observations increases.

Proposition 3 (Consensus is the plug-in Bayes decision)**.** *Let $\hat{\mathbf{s}}^{(1)}, \ldots, \hat{\mathbf{s}}^{(M)}$ be i.i.d. draws from $p_\theta(\cdot \mid \mathcal{Z})$ and let $\pi_i^\theta = p_\theta(s_i = 1 \mid \mathcal{Z})$. Then for every cell i: (a) $\hat{\pi}_i \to \pi_i^\theta$, with $\Pr[\,|\hat{\pi}_i - \pi_i^\theta| \geq \varepsilon\,] \leq 2e^{-2M\varepsilon^2}$; (b) if $\pi_i^\theta \neq \frac{1}{2}$, then $\hat{s}_i^{\text{cons}}$ equals the per-cell Bayes-optimal decision under p_θ with probability at least $1 - 2e^{-2M(\pi_i^\theta - 1/2)^2}$; and (c) $\hat{\pi}_i(1-\hat{\pi}_i) \to \text{Var}_\theta[s_i \mid \mathcal{Z}]$.*

Proposition 4 (Limit under the true posterior)**.** *Assume the model is well specified with $\mathcal{Z}$ sufficient and the observations conditionally independent given $\mathbf{s}$. For $i \notin \mathcal{A}$ the marginal equals the conditional prior $p(s_i = 1 \mid \mathbf{s}_\mathcal{A})$ and the data cannot reduce its variance. For $i \in \mathcal{A}$, let n_i count the observations separating $\mathbf{s}^\star$ from the alternatives differing at cell i; the consensus recovers $s_i^\star$ with error decaying in n_i.*

F.1 Proof of Proposition 3

Fix a cell i. Each sample $\hat{s}_i^{(m)}$ is a Bernoulli variable with parameter $\pi_i^\theta = p_\theta(s_i = 1 \mid \mathcal{Z})$, and the samples are independent because each reverse trajectory is driven by independent noise.

(a) Concentration. The variables are bounded in $[0, 1]$, so Hoeffding's inequality gives, for any $\varepsilon > 0$,

$$\Pr\Big[\,\big|\hat{\pi}_i - \pi_i^\theta\big| \geq \varepsilon\,\Big] \;\leq\; 2\exp\big(-2M\varepsilon^2\big), \tag{17}$$

and $\hat{\pi}_i \to \pi_i^\theta$ as $M \to \infty$ by the strong law of large numbers.

(b) Decision. The consensus $\hat{s}_i^{\text{cons}} = \mathbf{1}[\hat{\pi}_i > 1/2]$ differs from $\mathbf{1}[\pi_i^\theta > 1/2]$ only if $\hat{\pi}_i$ falls on the opposite side of $1/2$, which requires $|\hat{\pi}_i - \pi_i^\theta| \geq |\pi_i^\theta - 1/2|$. Applying equation 17 with $\varepsilon = |\pi_i^\theta - 1/2|$,

$$\Pr\Big[\hat{s}_i^{\text{cons}} \neq \mathbf{1}[\pi_i^\theta > \tfrac{1}{2}]\Big] \;\leq\; 2\exp\Big(-2M\big(\pi_i^\theta - \tfrac{1}{2}\big)^2\Big). \tag{18}$$

The limit $\mathbf{1}[\pi_i^\theta > 1/2]$ is the Bayes decision for cell i under p_θ with 0–1 loss, i.e. the minimizer of the expected per-cell Hamming error, since for a Bernoulli posterior the optimal decision is the more probable value.

(c) Variance. The map $x \mapsto x(1-x)$ is continuous, so $\hat{\pi}_i(1-\hat{\pi}_i) \to \pi_i^\theta(1-\pi_i^\theta) = \text{Var}_\theta[s_i \mid \mathcal{Z}]$ by the continuous mapping theorem. □

Two remarks. First, equation 18 shows that the number of samples required scales as $1/(\pi_i^\theta - 1/2)^2$: cells whose marginal is close to $1/2$ are exactly those that need many draws to stabilize, and they are also the cells on which the samples disagree most. The instability of the estimate and the ambiguity it reflects are the same phenomenon. Second, the proposition concerns the *learned* posterior; it says that sampling extracts a decision already encoded in p_θ and adds no information of its own.

F.2 Proof of Proposition 4

Rather than assuming a uniform information gap at every covered cell, we define the relevant count directly in terms of which observations can distinguish the competing hypotheses.

Definition 1 (Separating observation)**.** Observation k *separates* two scenes $\mathbf{s}, \mathbf{s}'$ if $\mathcal{F}(\mathbf{s}, \mathbf{p}_k) \neq \mathcal{F}(\mathbf{s}', \mathbf{p}_k)$, i.e. the two scenes induce different path parameters for link k. For a cell i, we write n_i for the number of observations that separate $\mathbf{s}^\star$ from the alternatives differing at cell i.

Two cases give $X_k = 0$ and are excluded from n_i by construction: observations with $i \notin a_k$, which are invariant to cell i; and observations that are role-degenerate for the pair, which by Proposition 2 induce identical path parameters under both hypotheses despite covering the cell. Neither requires a separate assumption, and $n_i \leq |\{k : i \in a_k\}|$ in general.

We assume:

- (A1) *Well specification.* $p_\theta(\cdot \mid \mathcal{Z}) = p(\cdot \mid \mathcal{O})$, with $\mathcal{Z}$ sufficient for $\mathbf{s}$.
- (A2) *Forward invariance.* $p(\mathcal{O} \mid \mathbf{s})$ depends on $\mathbf{s}$ only through $\mathbf{s}_\mathcal{A}$.
- (A3) *Conditional independence.* The observations are independent given $\mathbf{s}$.
- (A4) *Regularity.* The log-likelihood ratios X_k are sub-exponential, so their moment generating functions exist near the origin.

(i) Unobservable cells. Let $i \notin \mathcal{A}$. Factorizing the prior as $p(\mathbf{s}) = p(\mathbf{s}_\mathcal{A})\, p(\mathbf{s}_{\mathcal{A}^c} \mid \mathbf{s}_\mathcal{A})$ and applying (A2) to Bayes' rule gives the factorization of equation 2, in which $\mathcal{O}$ does not appear in the second factor. Marginalizing, $\pi_i = p(s_i = 1 \mid \mathbf{s}_\mathcal{A})$, so the belief at cell i is supplied entirely by the prior conditioned on the observable part of the scene. As the data sharpen $p(\mathbf{s}_\mathcal{A} \mid \mathcal{O})$, the variance at cell i converges to the prior-conditional variance $\mathrm{Var}_{\mathrm{prior}}[s_i \mid \mathbf{s}^\star_\mathcal{A}]$, a floor the measurements cannot lower. This holds in both regimes below.

(ii) Observable cells. The forward map is deterministic in the *physical* scene — the mesh geometry, building heights, and material properties from which the ray tracer computes propagation — but the occupancy grid $\mathbf{s}$ is a coarse projection of that description. Writing $\mathbf{g}$ for the physical scene, many $\mathbf{g}$ are consistent with the same grid, and they produce different channels; the likelihood entering the posterior is therefore the marginal

$$p(H_k \mid \mathbf{s}) \;=\; \int p\big(H_k \mid \mathbf{g}\big)\, p(\mathbf{g} \mid \mathbf{s})\, d\mathbf{g}, \tag{19}$$

which is a density over channel realizations rather than a point mass, even when the simulator itself is noiseless. The Kullback–Leibler divergences below are accordingly well defined.

Let $\delta_i > 0$ denote the smallest divergence $\mathrm{KL}\big(p(H_k \mid \mathbf{s}^\star) \,\|\, p(H_k \mid \mathbf{s}')\big)$ over the separating observations counted by n_i and the alternatives $\mathbf{s}'$ in contention at cell i; separation guarantees $\delta_i > 0$. Using (A3), the log posterior odds between $\mathbf{s}^\star$ and such an alternative decompose as

$$L \;=\; \log \frac{p(\mathbf{s}^\star \mid \mathcal{O})}{p(\mathbf{s}' \mid \mathcal{O})} \;=\; \underbrace{\log \frac{p(\mathbf{s}^\star)}{p(\mathbf{s}')}}_{=:L_0} \;+\; \sum_{k=1}^{K} X_k, \qquad X_k \;=\; \log \frac{p(H_k \mid \mathbf{s}^\star)}{p(H_k \mid \mathbf{s}')}, \tag{20}$$

where $X_k \equiv 0$ for every non-separating observation, leaving n_i nonzero terms. Taking expectations under the truth, $\mathbb{E}[X_k] \geq \delta_i$, so the drift $\mathbb{E}[\sum_k X_k] \geq n_i \delta_i$ grows linearly in n_i while the standard deviation of the sum grows as $O(\sqrt{n_i})$ by independence. An error at cell i requires $L \leq 0$, so the sum must fall below its mean by at least $n_i\delta_i + L_0$. A Chernoff bound gives, for suitable $\lambda > 0$,

$$\Pr\big[\textstyle\sum_k X_k + L_0 \leq 0\big] \;\leq\; e^{\lambda L_0} \prod_{k \text{ separating}} \mathbb{E}\big[e^{-\lambda X_k}\big] \;\leq\; e^{-c\, n_i \delta_i}, \tag{21}$$

for a constant $c > 0$ depending on the sub-exponential parameters, with L_0 absorbed for n_i large enough. Combining with (A1) and Proposition 3, the consensus recovers $s_i^\star$ with error decaying in $n_i\delta_i$. □

Remarks. The stochasticity in equation 19 is intrinsic to the problem rather than an artifact of measurement noise: the occupancy grid does not determine building heights, facade materials, or sub-cell geometry, all of which shape the channel, so a cell's occupancy constrains the observed

channel only in distribution. This is the same representational gap discussed in Appendix B.3, where a link may clear an occupied cell entirely. It also bounds what any amount of data can achieve, since δ_i is limited by how much of the physical scene the grid leaves unspecified, and is therefore one reason accuracy saturates below unity in Section 4.4 even as observations accumulate. The quantity governing recovery at a cell is n_i, the number of observations that actually separate the competing hypotheses there, which excludes both uncovered and role-degenerate observations. Role degeneracy therefore does not halt recovery; it reduces the effective number of informative observations, and only if every link in the deployment were non-separating for a pair differing at a cell would that cell remain unresolved. Regarding (A1), diffusion models are consistent in the limit of data and capacity, not at finite scale, so the proposition describes the ideal posterior and the empirical question is how closely p_θ approaches it; the comparison against a decoder conditioned on raw channels probes the sufficiency half of this assumption. Regarding (A2), real propagation leaks weak energy off the dominant paths, so in practice the boundary of $\mathcal{A}$ is soft and concentration on $\mathcal{A}^c$ is slow rather than exactly absent.

G Dataset and training details

Dataset. Channels are simulated with Sionna RT (Hoydis et al., 2023) over more than 11,000 outdoor urban environments spanning 200 cities, each covering a $200\,\mathrm{m} \times 200\,\mathrm{m}$ area, with three-dimensional building geometry extracted from OpenStreetMap (OpenStreetMap contributors, 2023) using the Geo2SigMap pipeline (Li et al., 2024). Simulations use a carrier frequency of 5.8 GHz with 100 MHz bandwidth and 128 subcarriers. The base station is equipped with a uniform linear array of $N_{tx} = 32$ antennas at a height of 15 m and serves more than 2,000 randomly placed users per environment, each equipped with an $N_{rx} = 8$-element uniform linear array at 1.5 m. The dataset comprises over 30 million samples, each consisting of a complex channel tensor, the user location, and the building height map. Occupancy grids are obtained on a 20×20 grid, i.e. $\Delta R = 10$ m, whose cells coincide with the scene patches of the foundation model. Propagation-relevant sets are recorded from the ray tracer as described in Appendix A.

Foundation model pretraining. The encoder consists of 12 transformer layers with 512-dimensional embeddings and a 2,450-token context window, totaling approximately 60M trainable parameters. Pretraining uses the objective of equation 16 with progressive masking combinatorial ratios. It was performed on 32 NVIDIA L40 GPUs for 9,600 GPU-hours over 48 epochs, processing approximately 882 billion tokens.

Diffusion decoder training. The denoiser has depth 6, width 192, 6 attention heads, and MLP ratio 4, with a linear adapter mapping the 512-dimensional encoder embedding to the decoder width. Because the encoder is frozen, its `[PHYSC]` embeddings are extracted once for up to 2,000 observations per scene and cached, so decoder training involves no forward or backward pass through the encoder. We use $T = 1000$ steps with the cosine schedule of Appendix E and conditioning dropout $p_{\mathrm{drop}} = 0.15$. To serve any number of observations with one model, the number of observations per training step is redrawn uniformly from $[K_{\min}, K_{\max}] = [100, 500]$ and shared across the scenes of the batch. Training uses AdamW with learning rate 10^{-4} and weight decay 10^{-4}, a linear warmup over the first epoch from $0.1\times$ the base rate followed by cosine decay, and gradient-norm clipping at 1.0, with 32 scenes per GPU on 8 GPUs for 100 epochs of 500 steps each. Samples are drawn with the conditional prediction, without guidance, and the consensus is formed over $M = 32$ samples.

Evaluation. All models are evaluated on 671 held-out test environments. For each scene, observations are loaded once at the largest sweep value, $K = 1000$, and every smaller K uses a prefix of the same sequence, so observation sets are nested across sweep points and differences across K reflect added observations rather than resampling. The per-observation encodings of every model are computed once per scene and reused across sweep points; only the fusion step, or the reverse-diffusion sampling for the two diffusion models, is repeated for each K.

H Marginal Prob. + FM

Per-observation decoder. This variant reuses the pretrained model in full, including the lightweight decoder of the pretraining stage. Each observation is encoded with its height-map input entirely masked, and the decoder predicts the occupancy of every scene patch from the channel and location alone,

$$q_{k,i} \;=\; \sigma\big(g_\psi(f_\phi(H_k, \mathbf{p}_k))_i\big), \tag{22}$$

which is the occupancy belief supported by observation k alone. Unlike Diffusion + FM, the encoder is not frozen: ϕ and ψ are fine-tuned jointly under the objective below.

Confidence-shaping objective. Let a_k be the propagation-relevant set of observation k, ∂a_k the cells within one step of a_k (8-connected) that are not in a_k, and $u_k = (a_k \cup \partial a_k)^c$ the remaining cells. Write $\mathrm{BCE}_\beta(q, s) = -[\,\beta\, s \log q + (1-s)\log(1-q)\,]$ for class-balanced cross-entropy with $\beta = (1-\rho)/\rho$, where ρ is the fraction of occupied cells in the scene, clipped to $[0.01, 0.99]$. The model is trained with

$$\begin{aligned}\mathcal{L}_{\mathrm{marg}} = \frac{1}{|a_k|}\sum_{i\in a_k}\mathrm{BCE}_\beta(q_{k,i}, s_i) - \lambda_c\,\frac{1}{|a_k|}\sum_{i\in a_k}\big|2q_{k,i}-1\big| \\ + \lambda_n\,\frac{1}{|\partial a_k|}\sum_{i\in\partial a_k}\mathrm{BCE}(q_{k,i}, s_i) + \lambda_u\,\frac{1}{|u_k|}\sum_{i\in u_k}\Big(q_{k,i}-\tfrac{1}{2}\Big)^2.\end{aligned} \tag{23}$$

The first term fits the occupancy of the cells the observation actually constrains, and the second rewards confident predictions there. The third supervises the boundary of the relevant set with a reduced weight, since cells adjacent to a path are partially constrained. The fourth drives predictions on cells the observation does not reach toward $1/2$. This last term is what gives the belief its meaning: under cross-entropy alone, every cell would be pushed toward a confident label and the belief could no longer indicate where the observation is uninformative. The propagation-relevant sets are used only to form this loss during training and are not required at inference. We use $\lambda_c = 0.1$, $\lambda_n = 0.3$, and $\lambda_u = 0.5$, with unit weight on the first term. Training uses AdamW with learning rate 10^{-5} and weight decay 10^{-4}, 32 observations per GPU, a linear warmup over the first epoch followed by cosine decay, and gradient-norm clipping at 1.0, for 20 epochs, each covering one quarter of the training corpus.

Confidence-gated geometric fusion. Per-observation beliefs are combined by a confidence-gated, geometrically weighted average: an observation contributes to a cell only if its belief there is confident, and in proportion to how close the cell lies to its link,

$$\begin{aligned} w_{k,i} &= \gamma_{k,i}\,\mathbf{1}\big[\,|2q_{k,i}-1| > \tau_c\,\big], \qquad \gamma_{k,i} = \exp\!\Big(-\frac{d_{k,i}^2}{2\sigma^2}\Big), \\ P_i &= \frac{\sum_k w_{k,i}\,q_{k,i}}{\sum_k w_{k,i}}, \qquad \hat{s}_i = \mathbf{1}\Big[\,P_i > \tfrac{1}{2} \;\wedge\; \sum_k w_{k,i} > \kappa\,\Big]. \end{aligned} \tag{24}$$

where $d_{k,i}$ is the distance from the center of cell i to the segment joining the base station and user k. The geometric weight depends only on the base-station and user positions, which are known at inference; the propagation-relevant sets are not used. Cells that receive no sufficiently weighted vote are assigned the free label, so every cell receives a decision and all metrics are computed over the full grid. We use $\tau_c = 0.75$, $\sigma = 50$ m, and $\kappa = 0.05$. The confidence gate is what allows the fusion to discard observations that are uninformative about a cell, and it is effective only because equation 23 drives such beliefs toward $1/2$.

I BASELINES

Diffusion + Raw Meas. The denoiser architecture, noise schedule, sampler, and number of consensus samples are identical to those of Diffusion + FM. The conditioning set is formed from the raw observations by a learnable tokenizer τ_ω, trained jointly with the decoder,

$$\mathcal{Z}_{\mathrm{raw}} \;=\; \big\{\tau_\omega(H_k, \mathbf{p}_k)\big\}_{k=1}^{K},\; (\theta^\star, \omega^\star) = \arg\min_{\theta,\omega}\; \mathbb{E}_{(\mathbf{s},\mathcal{O})\sim\mathcal{D}}\,\mathbb{E}_{t,\boldsymbol{\epsilon}}\Big[\,\big\|\,\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta\big(\mathbf{x}_t,\, t,\, \mathcal{Z}_{\mathrm{raw}}\big)\big\|_2^2\Big]. \tag{25}$$

The tokenizer τ_ω is a shallow transformer encoder with 3 layers, width 256, and 8 heads. Each raw observation is tokenized as in the channel front end of the foundation model — one token per transmit–receive antenna pair and group of 16 subcarriers, giving 2,048 channel tokens with factored sinusoidal positional encodings along the transmit, receive, and subcarrier axes — together with a location token, and the encoded sequence is mean-pooled into a single token per observation. The

conditioning set $\mathcal{Z}_{\text{raw}}$ therefore has the same cardinality K as the set of [PHYSC] embeddings it replaces and enters the identical denoiser through a linear adapter to the decoder width. τ_ω and θ are trained jointly from scratch with learning rate 10^{-4}, conditioning dropout 0.15, and K drawn from $[100, 500]$, for 100 epochs over the training scenes on 32 GPUs. Because the decoder is held fixed, any difference from Diffusion + FM is attributable to the conditioning.

Marginal Prob. + Raw Meas. An encoder h_ω with 12 layers, width 512, and 16 heads — capacity-matched to the foundation-model encoder and using the same channel and location tokenization — is trained from scratch on individual raw observations, with N^2 learned positional query tokens, one per cell, reading out the occupancy of the full grid:

$$\tilde{q}_{k,i} = \sigma\big(h_\omega(H_k, \mathbf{p}_k)_i\big),\ \mathcal{L}_{\text{naive}} \ =\ \frac{1}{N^2}\sum_{i=1}^{N^2} w_{k,i}\,\mathrm{BCE}_\beta\big(\tilde{q}_{k,i}, s_i\big),\ w_{k,i} = \begin{cases} 3, & i \in a_k, \\ 1, & \text{otherwise.} \end{cases} \tag{26}$$

Training uses AdamW with learning rate 5×10^{-5} and weight decay 10^{-4}, 32 observations per GPU on 32 GPUs, a linear warmup over the first epoch followed by cosine decay, and gradient-norm clipping at 1.0, for 32 epochs, each covering one quarter of the training corpus. The per-observation estimates are fused with the same rule as Marginal Prob. + FM, equation 24, with identical parameters, so the two models differ in representation and training objective but not in fusion. Here the confidence gate is ineffective. The loss equation 26 supervises every cell with a hard label whether or not the observation constrains it: relevant cells are up-weighted, but cells outside a_k are still driven toward their occupancy label rather than toward $1/2$. Beliefs about cells an observation never sensed are therefore confident, pass the gate, and enter the average. The geometric weight is then the only remaining filter, and it reflects proximity to the line-of-sight segment rather than the propagation-relevant set: reflected paths reach cells far from that segment, while cells near it need not interact with the link at all.

J Evaluation metrics

For a reconstruction $\hat{\mathbf{s}}$ and ground truth $\mathbf{s}$, let TP, FP, TN, and FN count the cells in each outcome, with occupied cells as the positive class. We report

$$\mathrm{BA} = \frac{1}{2}\left(\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} + \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}\right),\ \mathrm{F1} = \frac{2\,\mathrm{TP}}{2\,\mathrm{TP}+\mathrm{FP}+\mathrm{FN}},\ \mathrm{IoU} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}+\mathrm{FN}}. \tag{27}$$

Balanced accuracy weights the two classes equally, so a constant or random predictor scores 50% irrespective of the fraction of occupied cells; this makes it the appropriate indicator of chance-level behavior. F1 and IoU depend only on the occupied class and are insensitive to correctly predicted free space, which dominates most scenes, and are therefore stricter measures of footprint recovery. For a single scene the two are related by $\mathrm{IoU} = \mathrm{F1}/(2-\mathrm{F1})$; they differ after averaging over scenes and are both reported for comparability with prior work. All metrics are computed per scene over the full $N \times N$ grid and averaged over the 671 test scenes.

K Physical Interpretation of Pretrained Representations

The hierarchical design of Section 3 rests on the claim that the [PHYSC] token summarizes the physical structure that produced an observation, rather than the channel as a signal. This appendix examines that claim directly, independently of the reconstruction task, by inspecting the geometry of the representation space itself.

Setup. We encode a set of held-out observations with the frozen foundation model and project the resulting [PHYSC] embeddings to two dimensions with t-SNE (Maaten & Hinton, 2008), which preserves local neighborhood structure. As a control, we apply the identical projection to the raw channel measurements from which those embeddings were computed. Any structure visible in the embeddings but absent from the raw channels is structure the pretraining objective has introduced, since no information is created by the encoder (Section 2.3). We color the same two projections by three properties, each relevant to a different aspect of the propagation problem: the spatial cluster of the user position, the dominant propagation direction of the link, and the identity of the environment in which the measurement was taken.

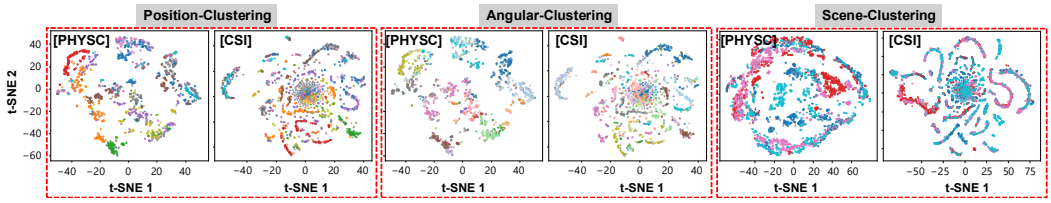

Figure 7: t-SNE of the encoded [PHYSC] token and raw CSI samples, colored by spatial cluster ID, dominant propagation path, and scene identity, showing foundation model representation forms clearer clustering across physically-relevant wireless link properties.

Observations. Figure 7 shows the projections. Under all three colorings, the `[PHYSC]` embeddings form markedly clearer structure than the raw channels, and each of the three carries a distinct implication for this work.

Separation by *user position* indicates that the representation resolves where a link terminates, which is a prerequisite for placing the evidence an observation carries at the correct location in the scene. Separation by *dominant propagation direction* indicates that the representation disentangles the angular components of the channel — the departure and arrival directions that determine which structures a link interacts with — rather than retaining the raw superposition in which those components are mixed. Separation by *scene identity* is the most directly relevant here: observations taken in the same environment are encoded to nearby points, even though the users occupy different positions and therefore observe different channels. The representation thus carries information about the environment that produced a measurement and not merely about the measurement itself, which is what makes a set of embeddings from one scene an informative conditioning set for the posterior over that scene.

Relation to the conditioning argument. The raw channels contain at least as much information as the embeddings computed from them, so the contrast in Figure 7 is not one of information content but of accessibility. In the raw space, observations that share an environment are not neighbors, and a decoder conditioned there must learn the mapping from channel structure to scene structure before it can exploit the conditioning at all. In the representation space, that structure is already exposed. This is the geometric counterpart of the learnability argument of Section 3, and it is consistent with the behavior of the two conditioning schemes in Section 4: the diffusion decoder conditioned on raw measurements does not improve as observations accumulate, while the same decoder conditioned on `[PHYSC]` embeddings does.

L Availability of Position Information

Each wireless observation is represented as a channel measurement and the corresponding user position, $(H_k, \mathbf{p}_k)$. To evaluate the effect of incomplete position information, we vary the fraction of measurements with available position labels from 10% to 100%, while keeping the total number of measurements per scene fixed $K = 500$. Figure 8 reports the resulting balanced accuracy.

All models except *Diffusion + Raw Meas.* improve as position availability increases and eventually saturate. *Diffusion + FM* improves from 0.56 at 10% availability to 0.76 at 100%, while *Marginal Prob. + FM* increases from 0.62 to 0.74 and *Marginal Prob. + Raw Meas.* from 0.54 to 0.62. In contrast, *Diffusion + Raw Meas.* remains near 0.50 throughout.

Position information is essential because a channel measurement reveals propagation-induced geometric information but does not by itself anchor that information to an absolute location in the scene. The user position maps the propagation evidence to specific scene cells, enabling the model to spatially localize the structures associated with each observation. As a result, reducing position availability directly reduces the

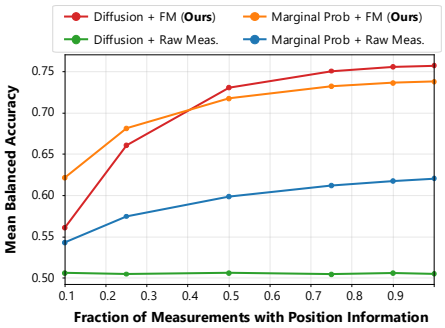

Figure 8: Mean scene reconstruction accuracy as a function of the fraction of measurements with position information.

amount of spatially anchored evidence available for reconstruction.

The performance curves largely saturate once roughly half of the measurements have position information, indicating that additional labeled observations provide diminishing benefits after sufficient spatial coverage has been established. At low availability, *Marginal Prob. + FM* initially outperforms *Diffusion + FM*, whereas the ordering reverses as more position-labeled observations become available. This behavior reflects the different inference mechanisms: the marginal formulation applies an explicit geometric fusion rule that remains effective with sparse anchored evidence, while the diffusion model benefits more from a sufficiently rich set of observations that it can jointly integrate through the learned posterior.

Finally, the near-constant performance of *Diffusion + Raw Meas.* across the entire sweep further indicates that its poor reconstruction quality is not primarily caused by insufficient position information, but by its inability to effectively exploit the raw channel measurements as conditioning information.

M Robustness to Noisy Position Information

Appendix L studies the case where position labels are missing for a fraction of observations. In deployment, another non-ideality can be that every observation carries a position, but the position is inaccurate. We therefore evaluate all models with perturbed user positions $\tilde{\mathbf{p}}_k = \mathbf{p}_k + \mathbf{e}_k$, $\mathbf{e}_k \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_2)$, with $\sigma = 5\,\mathrm{m}$ per horizontal axis, drawn independently for each observation. All models are trained on exact positions; noise is applied only at test time, with $K = 500$ observations per scene and all other settings as in Section 4. Under this model the radial error $\|\mathbf{e}_k\|$ is Rayleigh distributed, with median $\sigma\sqrt{2\ln 2} \approx 5.9\,\mathrm{m}$, mean $\sigma\sqrt{\pi/2} \approx 6.3\,\mathrm{m}$, and 95th percentile $\approx 12.2\,\mathrm{m}$. This range is representative of consumer GNSS in outdoor urban settings (Merry & Bettinger, 2019; Palathinkal et al., 2026).

Table 3: Reconstruction under noisy user positions ($\sigma = 5\,\mathrm{m}$ per axis, $K = 500$). Values in parentheses give the relative change of the mean with respect to exact positions (Table 1).

	Balanced Accuracy ↑			F1 ↑			IoU ↑		
Model	Mean	Median	P90	Mean	Median	P90	Mean	Median	P90
Diffusion + Raw Meas.	0.505 (±0.0%)	0.500	0.517	0.053 (-1.5%)	0.018	0.150	0.030 (±0.0%)	0.009	0.081
Marginal Prob. + Raw Meas.	0.609 (-1.9%)	0.604	0.655	0.374 (-5.9%)	0.371	0.480	0.233 (-7.1%)	0.228	0.315
Marginal Prob. + FM (Ours)	**0.728** (-1.4%)	**0.725**	0.797	**0.617** (-2.5%)	**0.619**	0.714	**0.453** (-3.7%)	**0.449**	0.556
Diffusion + FM (Ours)	0.696 (-8.2%)	0.676	**0.829**	0.553 (-16.5%)	0.538	**0.763**	0.397 (-21.6%)	0.368	**0.617**

Results. Table 3 shows that both of our designed pipelines remain far above the baselines under position information non-idealities. It is worth noting that the diffusion decoder, treats each embedding as precise evidence about a specific location and has never seen mislocalized evidence during training. Observations whose evidence is shifted in different directions then conflict. Posterior samples disagree on the affected cells, and their marginals move toward $1/2$, and the consensus readout suppresses footprints that are otherwise recovered. This mechanism mainly removes true positives, which is consistent with the larger drop in F1 and IoU than in balanced accuracy.

The sensitivity of Diffusion + FM is a train–test mismatch rather than a limitation of the posterior formulation, and it can be addressed directly. The position error/incompleteness can be treated as part of the observation model by training the decoder on embeddings computed from perturbed/missing positions, while keeping the clean occupancy $\mathbf{s}$ as the target. The decoder then learns $p_\theta(\mathbf{s} \mid \tilde{\mathcal{Z}})$ under such conditions. The architecture and training objective will be unchanged. We leave this implementation for future work.

N VISUAL DEMONSTRATION OF SCENE RECONSTRUCTION ALL MODELS

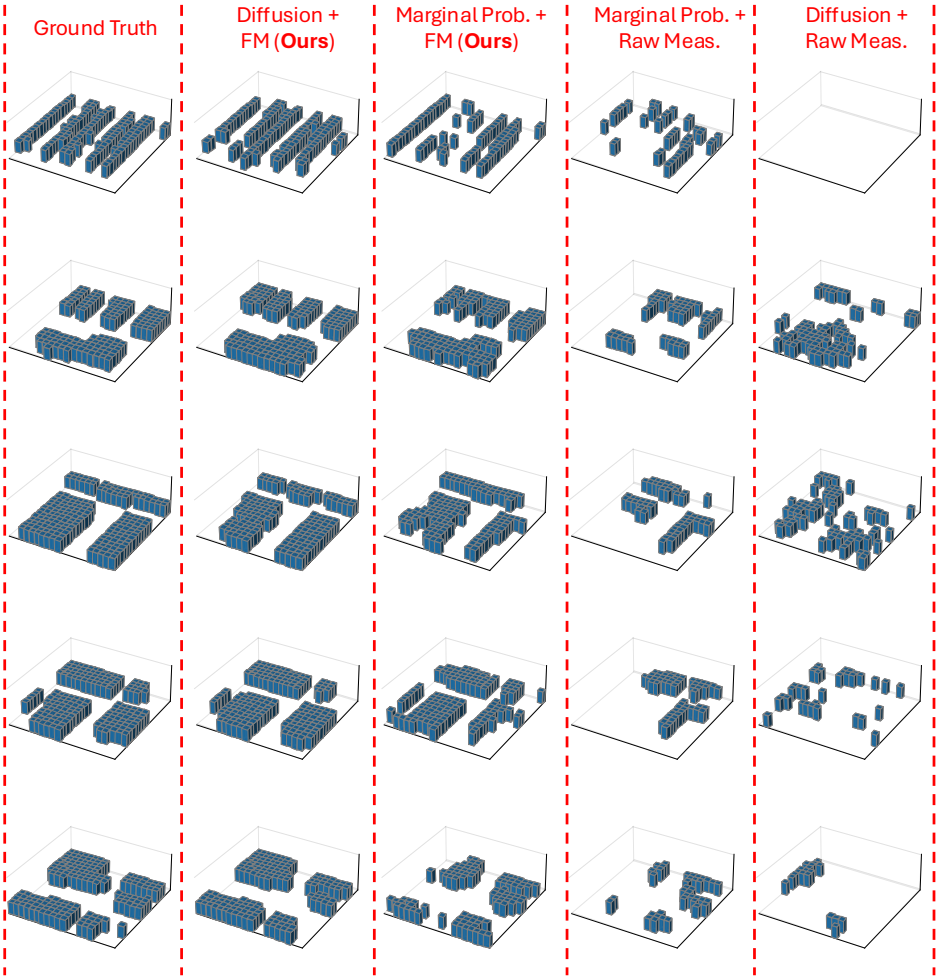

Figure 9: Visual comparison of scene reconstructions produced by all models against the ground-truth scenes.

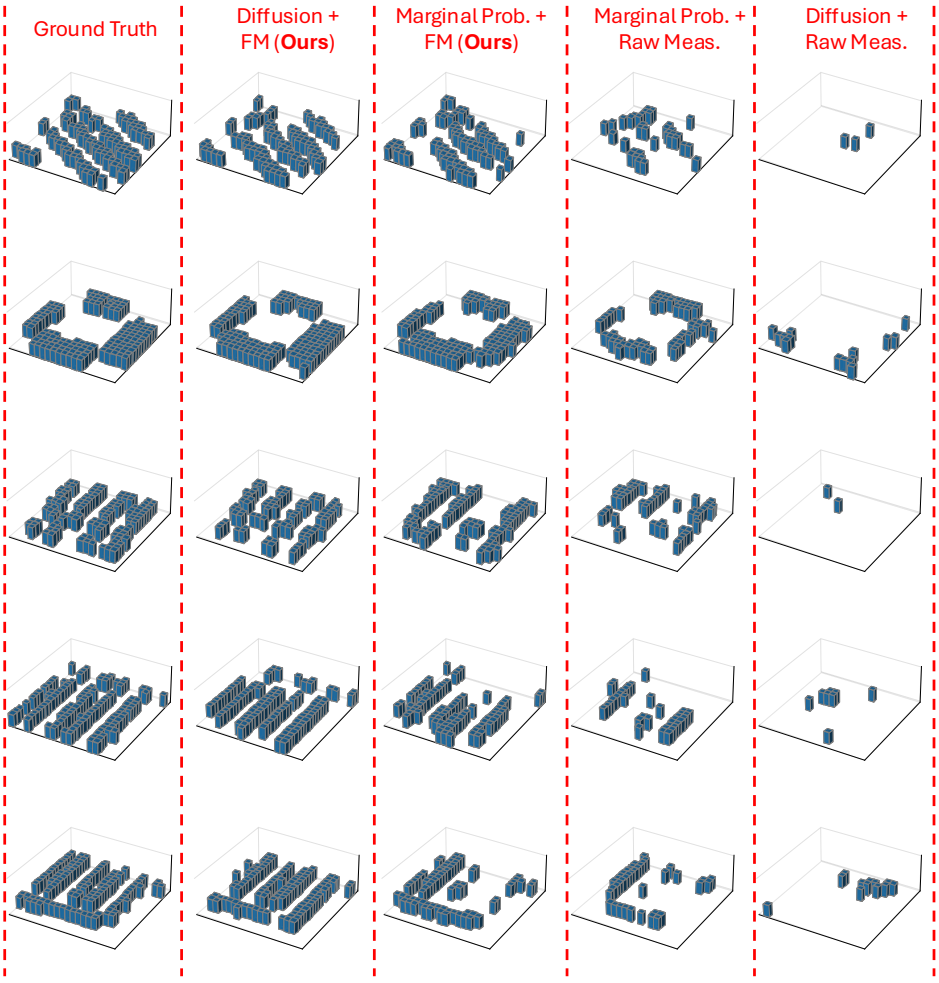

Figure 10: Visual comparison of scene reconstructions produced by all models against the ground-truth scenes.