LEARNING TO TRUST SPARSE METRIC EVIDENCE FOR MONOCULAR DEPTH FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Monocular depth foundation models provide strong dense geometric priors, while their predictions can still retain errors in metric scale and local depth. We study how little external metric evidence is needed to ground such priors, focusing on the extreme few-point regime where only one to a handful of sparse metric observations are available. We find that a few reliable observations can already yield substantial metric gains, but that this regime is inherently fragile: when evidence is scarce, each observation carries disproportionate influence, so unreliable measurements can erase the benefit of sparse grounding. We introduce Reliable Few-Point Metric Grounding (RFMG), a lightweight, backbone-agnostic module that learns both which observations to trust and how much correction the available evidence should be allowed to induce. RFMG estimates observation-level reliability and uses the amount of reliable evidence to control correction authority: scarce reliable evidence primarily supports global metric calibration, while spatially varying local refinement is progressively enabled as reliable evidence accumulates. The monocular backbone remains frozen throughout. Experiments across three frozen monocular depth backbones, four datasets, observation budgets, and corruption conditions show effective few-point grounding and reduced sensitivity to unreliable observations. On NYUv2 with UniDepthV2, only eight clean observations reduce AbsRel from 0.0766 to 0.0540, a 29.5% relative reduction, while RFMG still achieves 0.0608 when 80% of the sparse observations are corrupted. These results support viewing sparse depth not as incomplete scene geometry that must be densified from scratch, but as compact metric evidence for grounding the dense geometric prior already encoded by a monocular foundation model.

1 INTRODUCTION

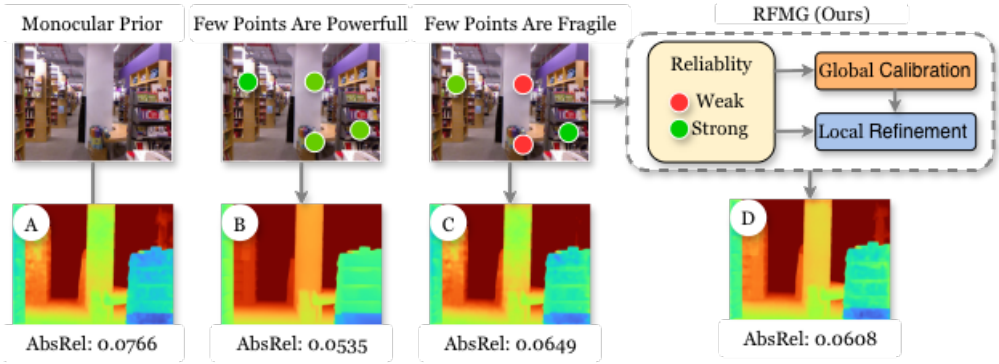


Figure 1: **Few metric observations are powerful, but fragile.** **(A)** RGB-only prior. **(B)** Direct local grounding with $N = 8$ clean observations. **(C)** The same grounding under 80% corruption. **(D)** RFMG under the same corrupted setting. On NYUv2 with UniDepthV2, AbsRel↓ changes from 0.0766 to 0.0535, 0.0649, and 0.0608, respectively. RFMG improves robustness through observation reliability, global metric calibration, and evidence-aware local refinement. Marker colors indicate observation quality for visualization only.

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.

Monocular depth foundation models increasingly provide strong dense geometric priors from a single RGB image (Hu et al., 2024; Piccinelli et al., 2026; Lin et al., 2026). In practical systems, high-quality dense metric depth is often costly or difficult to acquire, while lightweight or lower-resolution depth sensing provides sparse metric measurements with imperfect reliability. This motivates a basic question: *how little metric evidence is sufficient to ground a strong monocular prior and produce accurate dense metric depth?*

Depth-completion methods traditionally fuse RGB appearance with sparse metric measurements to recover dense depth (Zhang et al., 2023; Zuo & Deng, 2024). More recent approaches leverage pretrained monocular depth models as strong geometric priors, using sparse measurements to guide or adapt the pretrained prediction (Viola et al., 2025; Wang et al., 2026; Yu et al., 2026). While these methods effectively exploit sparse metric guidance, they are not primarily designed around the reliability problem that arises when only a handful of observations are available. Under such extreme scarcity, each measurement carries unusually high leverage over the resulting metric correction, so corrupted or otherwise unreliable evidence can disproportionately affect the prediction. This exposes two coupled challenges: *which observations should be trusted*, and *how much correction should the available evidence be allowed to induce*?

We study these challenges across $N \in \{1, \ldots, 64\}$ observations, with particular emphasis on the *extreme few-point regime* of $N = 1$–16. Our experiments reveal a characteristic *power–fragility trade-off*. When a strong monocular prior already captures much of the scene structure, only a few metric observations can provide substantial gains, while the same scarcity makes unreliable measurements particularly consequential (Fig. 1). Rather than treating sparse measurements primarily as incomplete geometry to be densified, we view them as compact metric evidence for grounding an already strong dense geometric prior.

We address these challenges with *Reliable Few-Point Metric Grounding (RFMG)*, a lightweight, backbone-agnostic module for frozen monocular depth foundation models. RFMG explicitly separates *observation reliability* from *correction authority*. It first estimates the reliability of each sparse observation and uses reliability-weighted evidence for global metric calibration. Spatially varying correction is then treated more conservatively. As reliable evidence accumulates, RFMG progressively enables local refinement; when reliable evidence remains scarce, correction remains predominantly global. Thus, reliability determines which observations influence the prediction, while the amount of reliable evidence determines what those observations are allowed to change.

We evaluate RFMG across three frozen monocular depth backbones, four datasets, multiple observation budgets, and clean and corrupted measurement conditions. Across these settings, few observations already provide substantial metric gains, while unreliable evidence can markedly reduce those gains under extreme scarcity. RFMG consistently reduces this sensitivity by controlling both observation influence and correction authority.

Our contributions are threefold:

- We formulate and systematically study *extreme few-point metric grounding* for monocular depth foundation models under scarce and potentially unreliable metric observations.
- We identify a power–fragility trade-off under extreme scarcity and introduce RFMG, which explicitly separates *observation reliability* from *correction authority* through reliability-weighted global metric calibration and evidence-dependent local refinement.
- We validate the proposed grounding principle across three frozen monocular depth backbones, four datasets, multiple observation budgets, and corruption conditions, showing effective few-point metric grounding with reduced sensitivity to unreliable observations.

2 Related Work

Monocular depth foundation models. Large-scale pretraining has substantially improved the transferability and geometric quality of monocular depth estimation. Depth Anything (Yang et al., 2024) learns transferable depth representations from large-scale unlabeled imagery, while metric-depth models such as Metric3D v2 (Hu et al., 2024) and UniDepthV2 (Piccinelli et al., 2026) target zero-shot metric geometry from a single image. More recent models such as Depth Anything 3 (Lin et al., 2026) further strengthen monocular geometric prediction across diverse scenes and view-

points. These developments provide increasingly strong dense geometric priors before any external metric measurement is introduced. Our work builds directly on this setting by keeping the monocular backbone frozen and studying how sparse external metric evidence can ground its dense prediction.

Sparse depth completion. Depth completion traditionally combines RGB images with sparse metric measurements to recover dense depth. Existing methods propagate or fuse sparse depth through convolutional, spatial-propagation, optimization, and transformer-based architectures. CompletionFormer (Zhang et al., 2023), for example, combines convolutional and transformer representations for RGB-guided depth completion, while OGNI-DC (Zuo & Deng, 2024) develops iterative geometric refinement for sparse-to-dense prediction. Input sparsity has also been studied explicitly. SpAgNet (Conti et al., 2023) is designed to remain effective across varying sparse-depth densities, including highly sparse inputs. These methods establish strong mechanisms for recovering dense geometry from incomplete metric observations. In this formulation, sparse depth remains a primary geometric signal to be propagated or densified into the final prediction.

Prior-guided depth completion. More recent methods instead exploit pretrained monocular depth models as strong dense priors and use sparse metric measurements to guide, adapt, or refine the pretrained prediction. Prompt Depth Anything (Lin et al., 2025) uses sparse depth as prompts for a depth foundation model. Marigold-DC (Viola et al., 2025) injects sparse metric guidance into a pretrained diffusion-based monocular depth model at test time, demonstrating that strong monocular priors can remain effective even under highly sparse guidance. Depth Anything with Any Prior (PriorDA) (Wang et al., 2026) combines incomplete metric priors with dense monocular geometry, while LDCM (Yu et al., 2026) studies generalizable depth completion from sparse observations using strong monocular priors. Together, these methods shift the role of sparse depth from geometry that must be densified toward metric evidence that guides or corrects an already informative dense prediction.

RFMG focuses on the reliability problem that becomes especially pronounced when metric evidence is extremely scarce. With only a handful of observations, each measurement carries unusually high leverage over the resulting correction, making corrupted or otherwise unreliable evidence particularly consequential. Existing prior-guided methods primarily study how sparse metric information can improve a strong monocular prior, while the joint problem of estimating *observation reliability* and controlling the *authority of the resulting correction* under extreme scarcity remains less explored. RFMG addresses this problem by explicitly separating these two factors, using reliability to determine which observations should influence the prediction and the amount of reliable evidence to determine whether the correction should remain predominantly global or extend to spatially varying local refinement.

3 METHOD

3.1 PROBLEM FORMULATION

Given an RGB image I and a sparse set of metric depth observations

$$\mathcal{P} = \{(\mathbf{u}_i, d_i)\}_{i=1}^{N}, \tag{1}$$

our goal is to estimate dense metric depth $\widehat{Z}$. Here, $\mathbf{u}_i$ denotes the image location of observation i and d_i its observed metric depth, which may be unreliable or corrupted. We focus on the *extreme few-point regime*, where only a handful of observations are available and each can therefore exert disproportionate influence on the prediction.

Rather than reconstructing dense geometry from $\mathcal{P}$, we use the sparse observations as metric evidence for grounding a frozen monocular depth prior,

$$(F, Z_0, C) = \Phi_{\text{frozen}}(I), \tag{2}$$

where F denotes dense features, Z_0 the initial metric depth, and C a backbone-specific uncertainty-related cue associated with the frozen monocular prior.

RFMG follows a simple principle: observation reliability determines *which observations to trust*, while the amount of reliable evidence determines *how much correction authority* that evidence re-

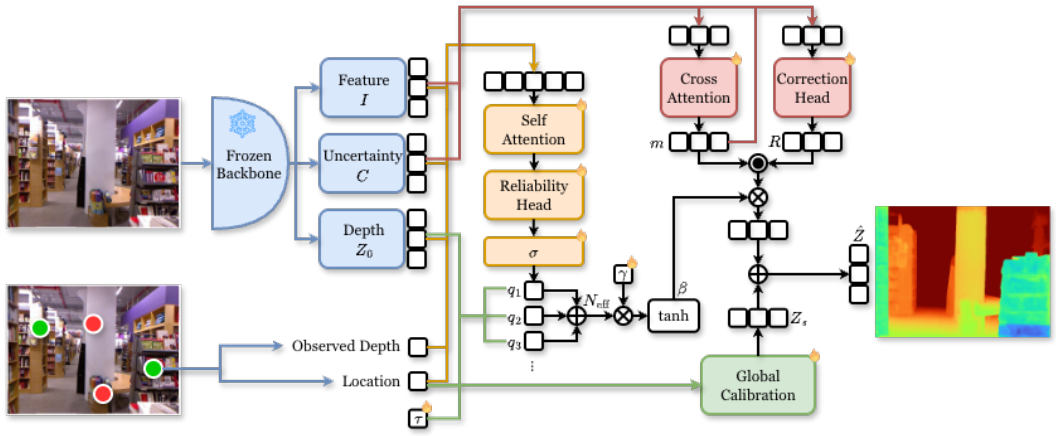


Figure 2: **Overview of Reliable Few-Point Metric Grounding (RFMG).** Given an RGB image and sparse metric depth observations, a frozen monocular depth backbone provides dense features F, an initial metric depth Z_0, and an uncertainty map C. RFMG contextualizes the sparse observations with self-attention and predicts observation-level reliability q_i using a lightweight reliability head. Reliability-weighted metric evidence is used for global calibration to obtain Z_s, while reliability-aware cross-attention produces a spatial support map m and a correction head predicts the residual R. The effective amount of reliable evidence N_{eff} controls the local correction authority β, producing the final metric depth $\widehat{Z}$.

ceives. As illustrated in Fig. 2, RFMG first estimates observation-level reliability, uses reliability-weighted metric evidence for global calibration, and progressively enables spatially varying local refinement as reliable evidence accumulates.

3.2 Learning to Trust Metric Evidence

Because only a few metric observations are available, an unreliable measurement can exert disproportionate influence on the refined prediction. We therefore estimate the reliability of each observation before using it to modify the monocular prior.

For each observation $(\mathbf{u}_i, d_i)$, we sample its local backbone feature, prior depth, and uncertainty value

$$f_i = F(\mathbf{u}_i), \qquad z_i = Z_0(\mathbf{u}_i), \qquad c_i = C(\mathbf{u}_i). \tag{3}$$

Here, $F(\mathbf{u}_i)$ denotes the feature vector sampled from the dense feature map F at location $\mathbf{u}_i$, while $Z_0(\mathbf{u}_i)$ and $C(\mathbf{u}_i)$ denote the corresponding sampled scalar values. We additionally measure the agreement between the observed depth and the monocular prior in log-depth space,

$$r_i = \log d_i - \log z_i, \tag{4}$$

and encode the observation as

$$x_i = [f_i, \mathrm{PE}(\mathbf{u}_i), \log d_i, \log z_i, r_i, c_i]. \tag{5}$$

Here, $\mathrm{PE}(\mathbf{u}_i)$ denotes a positional encoding of the observation location. A lightweight self-attention encoder contextualizes the sparse observations and produces representations h_i. A lightweight reliability head $g(\cdot)$ maps each representation to a scalar reliability logit, from which we obtain

$$q_i = \sigma(g(h_i)), \qquad q_i \in [0, 1],$$

where $\sigma(\cdot)$ is the sigmoid function and larger q_i indicates greater predicted reliability.

The uncertainty cue c_i is used only as an auxiliary signal, whereas q_i is explicitly learned to represent whether observation i should be trusted. These reliability scores subsequently control the contribution of each observation to both global metric calibration and evidence-aware local refinement.

3.3 RELIABILITY-WEIGHTED GLOBAL CALIBRATION

Reliable observations can constrain global metric scale even when they are too sparse to support spatially varying geometric correction. We therefore first use reliability-weighted metric evidence to calibrate the monocular prediction globally.

Each observation provides a scale cue

$$t_i = \frac{d_i}{z_i}. \tag{6}$$

We aggregate these cues using the predicted reliability and the backbone uncertainty signal,

$$w_i = q_i \widetilde{c}_i, \qquad W = \sum_{i=1}^{N} w_i, \qquad \bar{t} = \frac{\sum_{i=1}^{N} w_i t_i}{W + \epsilon}, \tag{7}$$

where $\widetilde{c}_i$ denotes a bounded transformation of the uncertainty cue c_i, and $\epsilon > 0$ is a small numerical constant. Here, W summarizes the total weighted evidence available for global calibration, while $\bar{t}$ aggregates the reliability-weighted scale cues.

To prevent scarce evidence from inducing an overly aggressive global correction, we form a shrunk scale estimate,

$$\widetilde{s} = 1 + \frac{W}{\tau + W + \epsilon}(\bar{t} - 1), \tag{8}$$

where $\tau > 0$ is learned. A bounded transformation of $\widetilde{s}$ produces the final global scale s, and the calibrated prior is

$$Z_s = sZ_0. \tag{9}$$

This low-dimensional correction allows sparse reliable evidence to adjust scene-level metric scale, while the evidence-dependent shrinkage limits the magnitude of the global correction when weighted support is scarce. The exact bounded scale transformation, clipping ranges, and numerical safeguards are provided in Appendix A.2.

3.4 EVIDENCE-AWARE LOCAL REFINEMENT

Global metric calibration corrects scene-level scale but cannot resolve spatially varying depth errors. Such local correction requires stronger evidence because very few observations provide limited support for deciding how geometry should change across the image.

After global calibration, we re-encode the sparse observations relative to the calibrated prior Z_s and propagate their metric evidence to dense image features using reliability-aware cross-attention. The predicted reliabilities q_i modulate the contribution of individual observations, while a null attention option allows the model to abstain when the available sparse evidence provides insufficient local support. The resulting assigned attention mass defines a dense support map m, which indicates where the sparse evidence supports local correction, while a lightweight correction head predicts the corresponding dense residual R.

We summarize the amount of reliable metric evidence as

$$N_{\text{eff}} = \sum_{i=1}^{N} q_i, \tag{10}$$

which acts as a soft effective count of the available reliable evidence rather than a literal integer observation count. We use it to control the correction authority granted to local refinement,

$$\beta = \tanh(\gamma \, \max(N_{\text{eff}} - 1, 0)), \tag{11}$$

where $\gamma > 0$ is learned, yielding $\beta \in [0, 1)$.

The final prediction is

$$\widehat{Z} = Z_s + \beta \, (m \odot R). \tag{12}$$

This decomposition assigns complementary roles to the available evidence: q_i determines *which observations to trust*, m determines *where local correction is supported*, R predicts *what local*

residual correction to apply, and β determines *how much local correction authority* the available reliable evidence receives.

In particular, when $N = 1$, we have $N_{\text{eff}} = q_1 \leq 1$ and therefore $\beta = 0$, reducing RFMG exactly to global metric calibration. As reliable evidence accumulates, spatially varying local refinement is progressively enabled. The exact cross-attention, abstention, and support-map formulations are provided in Appendix A.3.

3.5 Training

We train only RFMG while keeping the monocular depth backbone frozen. For each training image, we sample $N \sim \mathcal{U}\{1, \ldots, 16\}$ valid locations and let d_i^* denote the clean metric depth at location $\mathbf{u}_i$. To expose the model to unreliable evidence, we additionally sample a corruption prevalence $\rho \sim \mathcal{U}(0, 0.8)$ and independently perturb each observation with probability ρ, producing the observed input depth d_i. We associate each observation with a binary reliability target $y_i^{\text{rel}} \in \{0, 1\}$, where $y_i^{\text{rel}} = 1$ denotes an uncorrupted observation and $y_i^{\text{rel}} = 0$ a corrupted one. These targets provide direct supervision for the predicted reliabilities q_i. The exact corruption process is described in Appendix B.1.

We jointly supervise the final prediction, global metric calibration, and observation reliability. Let Z^* denote the dense ground-truth metric depth. The training objective is

$$\mathcal{L} = \mathcal{D}(\widehat{Z}, Z^*) + \lambda_s \, \mathcal{D}(Z_s, Z^*) + \lambda_r \, \mathcal{L}_{\text{rel}}, \tag{13}$$

where $\mathcal{L}_{\text{rel}}$ is binary cross-entropy between q_i and the synthetic reliability targets y_i^{rel}, and $\mathcal{D}$ is a depth reconstruction loss combining a log-depth term with an absolute metric-depth term. We use $\lambda_s = 0.5$ and $\lambda_r = 0.1$. The first term supervises the final dense prediction, the second encourages the global calibration branch to remain useful independently of local refinement, and the third trains observation-level reliability. The exact depth loss and optimization details are provided in Appendix B.2.

4 Experiments

4.1 Experimental Setup

Datasets. We use NYUv2 (Silberman et al., 2012) as our primary benchmark, following a split of 700 training, 95 validation, and 654 test images. We additionally evaluate on ETH3D (Schöps et al., 2017), KITTI (Geiger et al., 2012), and DIODE (Vasiljevic et al., 2019) to examine whether few-point metric grounding exhibits consistent behavior across diverse indoor and outdoor datasets. For these experiments, a separate RFMG refiner is trained on each dataset while the pretrained monocular depth backbone remains frozen. Dataset-specific evaluation ranges and preprocessing details are provided in Appendices D.1 and C.

Sparse metric evidence. Sparse metric observations are sampled uniformly without replacement from valid ground-truth depth pixels. During training, the observation budget is sampled as $N \sim \mathcal{U}\{1, \ldots, 16\}$. The main benchmark evaluates $N \in \{1, 8, 16\}$, while the fine-grained scaling analysis considers $N \in \{1, 2, 4, 8, 16, 32, 64\}$. The native monocular prediction at $N = 0$ serves as the no-observation reference.

To study unreliable metric evidence, we synthetically corrupt sparse observations using the depth-dependent perturbation process described in Appendix B.1. During corruption-aware training, the corruption prevalence is sampled as $\rho \sim \mathcal{U}(0, 0.8)$, while evaluation considers Clean, 40%, and 80% corruption. This protocol probes sensitivity to imperfect metric evidence rather than attempting to reproduce the complete noise characteristics of a particular physical sensor.

Backbones and training. We instantiate RFMG with Metric3D v2 ViT-L, UniDepthV2 ViT-L/14, and Depth Anything 3 (DA3) Metric-Large, keeping all backbone parameters frozen. Unless otherwise specified, controlled analyses use UniDepthV2. RFMG is optimized with AdamW, and checkpoints are selected using validation AbsRel. Dataset-specific training schedules and implementation details are provided in the appendix.

Table 1: **Extreme few-point metric grounding under unreliable observations on NYUv2.** AbsRel↓ for the RGB-only prediction ($N = 0$) and $N \in \{1, 8, 16\}$ sparse observations. For $N > 0$, entries denote **Clean** | **40%** | **80%** corruption. All RFMG models use corruption-aware training with $\rho_{\max} = 0.8$. **Bold** marks the best measured result per setting; dashes denote non-applicable settings.

Method	$N = 0$ **RGB**	$N = 1$ Clean | 40% | 80%	$N = 8$ Clean | 40% | 80%	$N = 16$ Clean | 40% | 80%
RGB-only				
Metric3D v2 (Hu et al., 2024)	.1257	–	–	–
UniDepthV2 (Piccinelli et al., 2026)	.0766	–	–	–
DA3 (Lin et al., 2026)	.0912	–	–	–
Sparse-guided				
CompletionFormer (Zhang et al., 2023)	–	.3246 | .3243 | .3249	.2795 | .2806 | .2809	.2472 | .2493 | .2512
OGNI-DC (Zuo & Deng, 2024)	–	.3045 | .3086 | .3123	.1259 | .1337 | .1388	.0782 | .0861 | .0945
PromptDA (Lin et al., 2025)	–	.4345 | .4402 | .4407	.1640 | .1717 | .1794	.1100 | .1200 | .1291
Marigold-DC (Viola et al., 2025)	–	.4368 | .9252 | 1.5001	.0992 | .7246 | 1.1772	.0503 | .0701 | .0852
Marigold-SSD (Gregorek et al., 2026)	–	.9021 | .9022 | .9024	.1218 | .1258 | .1297	.0975 | .1075 | .1081
PriorDA (Wang et al., 2026)	–	.4239 | .4247 | .4288	.0608 | .0712 | .0818	.0381 | .0491 | .0598
LDCM (Yu et al., 2026)	–	.1598 | .1693 | .1823	**.0488** | .0603 | .0740	**.0338** | .0431 | .0569
Ours				
Metric3D v2 + RFMG	–	.1289 | .1343 | .1359	.0677 | .0705 | .0732	.0482 | .0510 | .0538
UniDepthV2 + RFMG	–	**.0689** | **.0730** | **.0767**	.0540 | **.0563** | .0608	.0401 | **.0426** | **.0442**
DA3 + RFMG	–	.0831 | .0875 | .0901	.0545 | .0569 | **.0597**	.0448 | .0472 | .0505

Evaluation. We use Absolute Relative Error (AbsRel) as the primary metric. Metrics are computed over valid ground-truth pixels, including sparse observation locations. Unless otherwise specified, compared methods receive the same sparse observation coordinates under each evaluation condition. Additional metrics and evaluation details are provided in the appendix.

4.2 Main Results

Few points are powerful, but fragile. We first examine how much metric evidence is needed to improve a strong monocular depth prior. Table 1 compares RFMG with the native RGB-only predictions and representative sparse-guidance methods on NYUv2 under $N \in \{1, 8, 16\}$ sparse metric observations.

A small number of clean observations already provides substantial metric grounding. With UniDepthV2, AbsRel decreases from 0.0766 without metric evidence to 0.0689, 0.0540, and 0.0401 using 1, 8, and 16 clean observations, respectively. DA3 similarly improves from 0.0912 to 0.0831, 0.0545, and 0.0448. Metric3D v2 does not benefit from a single observation, but improves from 0.1257 to 0.0677 at $N = 8$ and 0.0482 at $N = 16$. Thus, the single-point regime is not uniformly sufficient across backbones, but substantial gains consistently emerge once only a few metric observations are available.

The same scarcity also makes grounding fragile to unreliable evidence. For UniDepthV2 at $N = 8$, AbsRel increases from 0.0540 with clean observations to 0.0563 and 0.0608 at 40% and 80% corruption, respectively. DA3 exhibits the same trend, increasing from 0.0545 to 0.0597 between clean and 80% corruption. Nevertheless, at $N = 8$ and $N = 16$, all three RFMG instantiations retain clear improvements over their corresponding frozen RGB-only predictions even under 80% corruption. The sparse-guidance baselines further illustrate that highly sparse metric guidance can become difficult to exploit once individual observations are unreliable. Together, these results expose the central power–fragility trade-off: scarce metric evidence can be highly informative, but its influence must be controlled when it cannot be trusted uniformly.

From global calibration to local refinement. Figure 3 evaluates a fixed UniDepthV2+RFMG model over $N \in \{1, 2, 4, 8, 16, 32, 64\}$ observations, with the native monocular prediction at $N = 0$

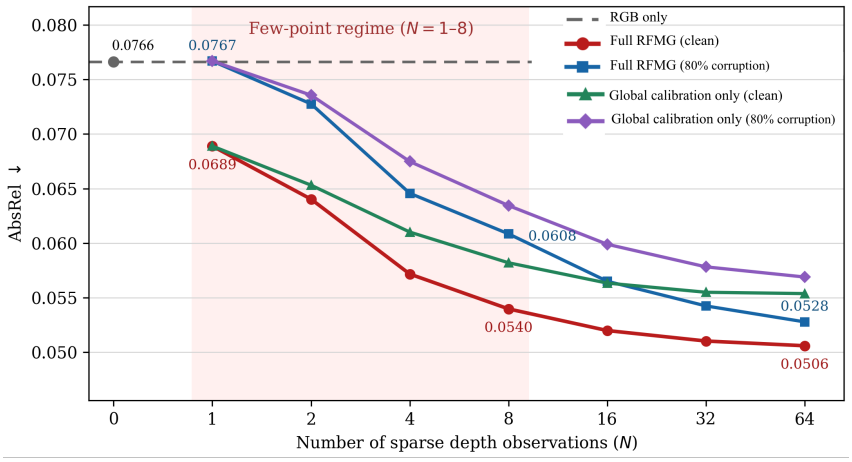


Figure 3: **Few-point scaling and global-to-local grounding on NYUv2.** AbsRel↓ as the number of sparse metric observations increases from $N = 1$ to 64, using UniDepthV2 as the frozen monocular prior. We compare full RFMG with global calibration only under clean and 80% corrupted observations. At $N = 1$, the two coincide because local refinement is disabled by design; as more reliable evidence becomes available, full RFMG increasingly improves over global calibration alone. The shaded region highlights the most data-scarce range ($N = 1$–8).

Table 2: **In-domain generality across datasets and observation budgets.** AbsRel ↓ using UniDepthV2 as the frozen monocular prior. A separate RFMG refiner is trained on each dataset while the backbone remains frozen. We evaluate $N \in \{1, 8, 16\}$ observations under Clean, 40%, and 80% corruption.

	RGB-only	**RFMG** ($N = 1$)			**RFMG** ($N = 8$)			**RFMG** ($N = 16$)		
Dataset	Z_0	**Clean**	**40%**	**80%**	**Clean**	**40%**	**80%**	**Clean**	**40%**	**80%**
NYUv2	.0766	.0689	.0730	.0767	.0540	.0563	.0608	.0401	.0426	.0442
ETH3D	.2166	.0982	.1045	.1118	.0509	.0562	.0615	.0418	.0452	.0496
KITTI	.1068	.0882	.0925	.0981	.0562	.0607	.0654	.0485	.0512	.0550
DIODE	.6887	.5420	.5680	.5990	.4269	.4617	.4820	.3650	.3980	.4320

as reference. Most of the gain already emerges within the most data-scarce range $N = 1$–8, after which performance improves more gradually.

At $N = 1$, full RFMG coincides with reliability-weighted global calibration by design because $N_{\text{eff}} \leq 1$ implies $\beta = 0$. As the observation budget grows, the full model increasingly improves over the global-only prediction under both clean and corrupted conditions. This is consistent with the intended correction hierarchy: very scarce reliable evidence first calibrates global metric scale, while additional evidence enables increasingly useful spatially varying refinement.

Generality across datasets. Table 2 evaluates UniDepthV2+RFMG on NYUv2, ETH3D, KITTI, and DIODE using separately trained in-domain refiners while keeping the monocular backbone frozen. Increasing the observation budget consistently reduces AbsRel across all four datasets and all three corruption conditions.

Under clean observations, all four datasets improve substantially as the observation budget increases. The single-point regime remains sensitive under heavy corruption: on NYUv2, $N = 1$ with 80% corruption yields 0.0767, essentially returning to the RGB-only error of 0.0766. By $N = 8$, however, RFMG remains better than the frozen RGB-only prior under 80% corruption on all four datasets. Together with Table 1, these results show that the power–fragility behavior is not specific to a single backbone or dataset.

Table 3: **Component analysis of RFMG on NYUv2.** Results use UniDepthV2 with $N = 8$ sparse metric observations. AbsRel ↓ is reported under clean, 40%, and 80% corruption. Direct local grounding attains slightly lower clean error but is more sensitive to unreliable observations, while full RFMG combines reliability, global calibration, and evidence-aware local refinement.

Variant	Reliability	Global Calib.	Local Refine.	Clean	40% AbsRel ↓	80%
Direct local grounding	×	×	✓	**0.0535**	0.0594	0.0649
w/o Reliability	×	✓	✓	0.0588	0.0621	0.0653
Global calibration only	✓	✓	×	0.0582	0.0601	0.0634
Full RFMG (Ours)	✓	✓	✓	0.0540	**0.0563**	**0.0608**

4.3 Analysis and Ablations

Which evidence to trust and how much to change. Table 3 isolates the roles of observation reliability, global metric calibration, and local refinement using UniDepthV2 on NYUv2 with $N = 8$ observations.

Direct local grounding achieves the lowest clean error, 0.0535, showing that aggressive local use of sparse evidence can be effective when the observations are reliable. However, its error rises to 0.0649 under 80% corruption. Full RFMG incurs only a 0.0005 increase in clean AbsRel (0.0540 versus 0.0535), while reducing the 80% corruption error to 0.0608, exposing the same power–fragility trade-off at the component level.

Explicit reliability estimation is consistently beneficial. Removing reliability increases AbsRel from 0.0540 to 0.0588 under clean observations, from 0.0563 to 0.0621 at 40% corruption, and from 0.0608 to 0.0653 at 80% corruption. Full RFMG also improves over global calibration alone in all three conditions, showing that spatially varying local refinement remains useful when its influence is controlled by reliable evidence. Together, these results support the central principle of RFMG: reliability determines *which evidence to trust*, while the amount of reliable evidence determines *how much spatially varying correction that evidence should be allowed to induce*.

Qualitative analysis. Additional error maps illustrating the progression from the frozen prior to global calibration and evidence-aware local refinement are provided in Appendix E.

5 Conclusion

We studied how much sparse metric evidence is needed to ground a strong monocular depth prior. Our results show that substantial metric gains can emerge from only a few observations, but that this extreme scarcity also makes each observation disproportionately influential and therefore potentially harmful when unreliable. We introduced Reliable Few-Point Metric Grounding (RFMG), which estimates which observations to trust and separates global metric calibration from spatially varying local refinement according to the amount of reliable evidence available. Across multiple frozen monocular depth backbones, datasets, observation budgets, and measurement-quality conditions, RFMG provides effective few-point grounding while retaining substantial benefit under unreliable observations. These results support a broader view of sparse metric sensing: when a strong dense geometric prior is already available, sparse measurements need not reconstruct scene geometry themselves, but can instead serve as compact metric evidence for grounding the existing representation.

AI Use Statement

Generative AI tools were used to assist with language editing, clarification of technical descriptions, code debugging, and review of mathematical and experimental presentation during the development of this work. All model designs, experiments, quantitative results, interpretations, and claims were reviewed and verified by the authors. The authors take responsibility for the final content of the paper, including all text, code, claims, and artifacts produced with the aid of generative AI.

Reproducibility Statement

We provide detailed method formulations, training objectives, corruption procedures, and implementation details in the appendix. Dataset splits, preprocessing, sparse-observation sampling, corruption settings, evaluation protocols, and backbone-specific configurations are also documented there. The main paper reports the model architecture, training setup, observation budgets, and evaluation metrics required to interpret and reproduce the reported experiments.

Ethics Statement

This work uses publicly available datasets and does not involve human subjects, personally identifiable information, or evaluation in safety-critical deployment settings. The study focuses on methodological evaluation of sparse metric depth grounding under controlled corruption settings.

References

Andrea Conti, Matteo Poggi, and Stefano Mattoccia. Sparsity agnostic depth completion. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5871–5880, 2023.

Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

Jakub Gregorek, Paraskevas Pegios, Nando Metzger, Konrad Schindler, Theodora Kontogianni, and Lazaros Nalpantidis. Need for speed: Zero-shot depth completion with single-step diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1861–1872, 2026.

Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen. Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10579–10596, 2024. doi: 10.1109/TPAMI.2024.3444912.

Haotong Lin, Sida Peng, Jingxiao Chen, Songyou Peng, Jiaming Sun, Minghuan Liu, Hujun Bao, Jiashi Feng, Xiaowei Zhou, and Bingyi Kang. Prompting depth anything for 4k resolution accurate metric depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17070–17080, 2025.

Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Yang Zhao, Sida Peng, Hengkai Guo, Xiaowei Zhou, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. In *International Conference on Learning Representations*, 2026.

Luigi Piccinelli, Christos Sakaridis, Yung-Hsu Yang, Mattia Segu, Siyuan Li, Wim Abbeloos, and Luc Van Gool. Unidepthv2: Universal monocular metric depth estimation made simpler. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(3):2354–2367, 2026. doi: 10.1109/TPAMI.2025.3628473.

Thomas Schöps, Johannes L. Schönberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3260–3269, 2017.

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European Conference on Computer Vision (ECCV)*, pp. 746–760, 2012.

Igor Vasiljevic, Nick Kolkin, Shanyi Zhang, Ruotian Luo, Haochen Wang, Falcon Z. Dai, Andrea F. Daniele, Mohammadreza Mostajabi, Steven Basart, Matthew R. Walter, and Gregory Shakhnarovich. DIODE: A dense indoor and outdoor depth dataset. *arXiv preprint arXiv:1908.00463*, 2019.

Massimiliano Viola, Kevin Qu, Nando Metzger, Bingxin Ke, Alexander Becker, Konrad Schindler, and Anton Obukhov. Marigold-dc: Zero-shot monocular depth completion with guided diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5359–5370, 2025.

Zehan Wang, Siyu Chen, Lihe Yang, Jialei Wang, Ziang Zhang, Hengshuang Zhao, and Zhou Zhao. Depth anything with any prior. In *International Conference on Learning Representations*, 2026.

Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10371–10381, 2024.

Zhu Yu, Zhengyi Zhao, Runmin Zhang, Lingteng Qiu, Kejie Qiu, Yisheng He, Siyu Zhu, Zilong Dong, Si-Yuan Cao, and Hui-Liang Shen. Large depth completion model from sparse observations. In *International Conference on Learning Representations*, 2026.

Youmin Zhang, Xianda Guo, Matteo Poggi, Zheng Zhu, Guan Huang, and Stefano Mattoccia. Completionformer: Depth completion with convolutions and vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18527–18536, 2023.

Yiming Zuo and Jia Deng. Ogni-dc: Robust depth completion with optimization-guided neural iterations. In *European Conference on Computer Vision*, 2024.

A Additional RFMG Formulation

A.1 Observation Encoding and Reliability Estimation

The observation representation defined in Sec. 3.2 has dimension 278 and is projected to a 256-D embedding using a two-layer MLP. The sparse observation set is contextualized using a single Transformer encoder block with four self-attention heads (head dimension 64). A two-layer reliability head with hidden dimension 128 maps each contextualized representation h_i to a scalar logit $g(h_i)$, followed by

$$q_i = \sigma(g(h_i)). \tag{14}$$

The resulting reliability scores are shared by the global calibration and local refinement stages.

A.2 Reliability-Weighted Global Calibration

This section gives the exact bounded transformations and numerical safeguards used by the global calibration stage in Sec. 3.3. For observation i, we use

$$t_i = \mathrm{clip}\left(\frac{d_i}{\max(z_i, 10^{-4})}, 0.2, 5.0\right), \qquad \tilde{c}_i = \mathrm{clip}(c_i, 0.1, 10.0), \tag{15}$$

followed by

$$w_i = q_i \tilde{c}_i, \qquad W = \sum_{i=1}^{N} w_i, \qquad \bar{t} = \begin{cases} \dfrac{\sum_i w_i t_i}{W + 10^{-6}}, & W > 10^{-6}, \\ 1, & \text{otherwise.} \end{cases} \tag{16}$$

The shrunk scale estimate is

$$\tilde{s} = 1 + \frac{W}{\tau + W + 10^{-6}}(\bar{t} - 1), \qquad \tau = \mathrm{softplus}(\theta_\tau) + 10^{-4}, \tag{17}$$

and is smoothly bounded with $\Delta = 0.30$:

$$s = 1 + \Delta \tanh\left(\frac{\tilde{s} - 1}{\Delta}\right), \qquad Z_s = sZ_0. \tag{18}$$

Thus, when weighted support is small, the estimate remains close to the frozen prior scale; when $W \to 0$, $s \to 1$.

A.3 Evidence-Aware Local Refinement

After global calibration, the sparse observations are re-encoded relative to Z_s. At observation location u_i,

$$z_{s,i} = sz_i, \qquad r_i^{(s)} = \mathrm{clip}\big(\log\max(d_i, 10^{-4}) - \log\max(z_{s,i}, 10^{-4}), -5, 5\big). \tag{19}$$

The resulting sparse tokens a_i are used as keys/values, while dense backbone features F_j are queries. For attention head h,

$$\ell_{ji}^h = \mathrm{clip}\left(\frac{(W_Q^h F_j)^\top (W_K^h a_i)}{\sqrt{64}}, -20, 20\right). \tag{20}$$

To allow abstention, we introduce a positive null mass

$$\alpha_{ji}^h = \frac{q_i \exp(\ell_{ji}^h)}{\kappa + \sum_{k=1}^N q_k \exp(\ell_{jk}^h)}, \qquad \kappa = \mathrm{softplus}(\theta_\kappa) + 10^{-4}. \tag{21}$$

The spatial support map is the attention mass assigned to real observations, averaged over the four heads:

$$m_j = \frac{1}{4}\sum_{h=1}^{4}\sum_{i=1}^{N} \alpha_{ji}^h \in [0, 1). \tag{22}$$

The value-weighted attention output is added to the frozen feature to obtain a grounded feature F_g. After bilinear upsampling to image resolution, F_g is concatenated with the backbone uncertainty cue and passed through a lightweight three-layer convolutional head to produce a dense residual R. The local correction is masked by the upsampled support map m.

The amount of reliable evidence is

$$N_{\text{eff}} = \sum_{i=1}^{N} q_i, \tag{23}$$

and the correction-authority gate is

$$\beta = \tanh(\gamma \max(N_{\text{eff}} - 1, 0)), \qquad \gamma = \mathrm{softplus}(\theta_\gamma) + 10^{-4}. \tag{24}$$

The final prediction is

$$\widehat{Z} = Z_s + \beta\,(m \odot R). \tag{25}$$

For $N = 1$, $N_{\text{eff}} = q_1 \le 1$, hence $\beta = 0$ and the model reduces exactly to global calibration.

B Training Details

B.1 Sparse Observation Sampling and Synthetic Corruption

For each training image, $N \sim \mathcal{U}\{1, \dots, 16\}$ valid ground-truth pixels are sampled uniformly without replacement. A corruption prevalence $\rho \sim \mathcal{U}(0, 0.8)$ is sampled, and each observation is independently corrupted with $b_i \sim \mathrm{Bernoulli}(\rho)$. The reliability target is $y_i^{\text{rel}} = 1 - b_i$.

Clean observations use $d_i = d_i^*$. Corrupted observations follow the depth-dependent perturbation

$$\tilde{u}_i = \mathrm{clip}\left(\frac{d_i^* - d_{\min}}{d_{\max} - d_{\min}}, 0, 1\right), \tag{26}$$

$$\sigma_i = 0.20\, d_i^* \left[1 + 2(2\tilde{u}_i - 1)^2\right], \tag{27}$$

$$d_i = \mathrm{clip}(d_i^* + \sigma_i \xi_i, d_{\min}, d_{\max}), \qquad \xi_i \sim \mathcal{N}(0, 1). \tag{28}$$

For NYUv2, $d_{\min} = 0.5$ m and $d_{\max} = 10$ m.

B.2 Training Objective and Optimization

For valid pixels $\mathcal{V}$, define

$$e_p = \log \max(\widehat{Z}_p, 10^{-6}) - \log \max(Z_p^*, 10^{-6}). \tag{29}$$

The dense reconstruction loss is

$$\mathcal{D}(\widehat{Z}, Z^*) = \sqrt{\frac{1}{|\mathcal{V}|}\sum_{p\in\mathcal{V}} e_p^2 - \frac{0.85}{|\mathcal{V}|^2}\left(\sum_{p\in\mathcal{V}} e_p\right)^2} + 0.1\frac{1}{|\mathcal{V}|}\sum_{p\in\mathcal{V}} |\widehat{Z}_p - Z_p^*|. \tag{30}$$

The full objective is

$$\mathcal{L} = \mathcal{D}(\widehat{Z}, Z^*) + 0.5\,\mathcal{D}(Z_s, Z^*) + 0.1\,\mathcal{L}_{\text{rel}}, \tag{31}$$

where $\mathcal{L}_{\text{rel}}$ is binary cross-entropy on the clean/corrupted reliability targets.

All RFMG models are optimized with AdamW (learning rate 10^{-4}, weight decay 10^{-4}, $\beta_1 = 0.9$, $\beta_2 = 0.999$) using cosine annealing with 5% linear warmup. We use an effective batch size of 8. Models are trained for 100 epochs on NYUv2, ETH3D, and DIODE, and for 20 epochs on KITTI. Checkpoints are selected by validation AbsRel. NYUv2 training uses 480×640 resolution; horizontal flipping ($p = 0.5$) and photometric color jitter ($p = 0.2$) are applied during training.

C Backbone-Specific Implementation

All monocular backbones remain frozen. The backbone-specific representations and preprocessing are:

- **UniDepthV2 ViT-L/14:** native 480×640 evaluation; 256-D depth features, native metric depth, and the native uncertainty-related cue C.
- **Depth Anything 3 Metric-Large:** images resized to 378×504 with ImageNet normalization; 256-D features are taken immediately before the final depth projection; metric depth is obtained using the resized focal length divided by 300.
- **Metric3D v2 ViT-L:** aspect-preserving resize to 616×1064 with padding; decoder features are cropped to remove padded borders and sliced to 256 channels; depth is unpadded, resized, and scaled by focal length divided by 1000. Because no native uncertainty cue is available, an edge-aware smoothness proxy derived from depth gradients is used.

D Experimental Protocol

D.1 Datasets and Evaluation Setup

- **NYUv2:** 700 training, 95 validation, and 654 test images; valid depth range $[0.5, 10.0]$ m.
- **ETH3D:** evaluated over valid depth range $[0.5, 10.0]$ m.
- **KITTI:** evaluated over valid depth range $[0.001, 80.0]$ m using the standard Garg crop.
- **DIODE:** valid ranges $[0.5, 10.0]$ m for indoor scenes and $[1.0, 80.0]$ m for outdoor scenes.

Sparse observations are sampled uniformly without replacement from valid ground-truth pixels. The main benchmark uses $N \in \{1, 8, 16\}$, while the scaling analysis uses $N \in \{1, 2, 4, 8, 16, 32, 64\}$. Evaluation uses Clean, 40%, and 80% corruption, with the same corruption model as Appendix B.1.

D.2 Evaluation Metrics

The primary metric is Absolute Relative Error,

$$\text{AbsRel} = \frac{1}{|\mathcal{V}|}\sum_{p\in\mathcal{V}} \frac{|\widehat{Z}_p - Z_p^*|}{Z_p^*}. \tag{32}$$

We additionally compute RMSE and the standard δ_k accuracy thresholds ($k \in \{1, 2, 3\}$) over the same valid mask. No test-time metric-scale alignment or post-processing is applied.

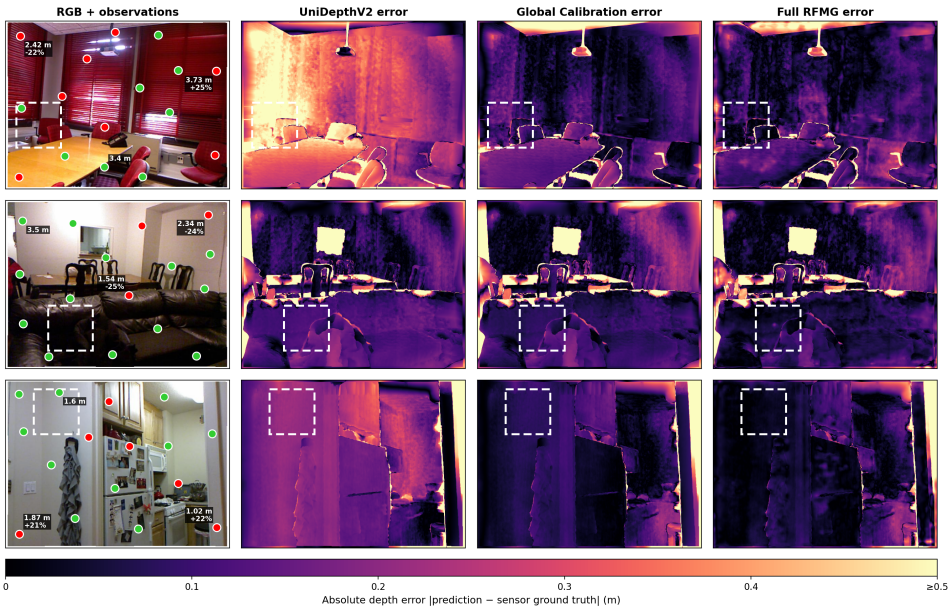


Figure 4: **Qualitative error analysis of few-point metric grounding under unreliable evidence.** Examples use UniDepthV2 with $N = 16$ observations under 40% corruption. Green and red markers denote clean and corrupted observations, respectively. We visualize absolute metric depth error for the frozen monocular prior, global calibration only, and full RFMG using a shared color scale, where darker regions indicate lower error. Global calibration removes a substantial portion of the scene-level metric error, while full RFMG further reduces spatially varying residual errors through evidence-aware local refinement. Dashed boxes highlight representative regions where these local improvements are visible.

E Qualitative Error Analysis

Figure 4 visualizes the progression from the frozen monocular prior Z_0, through global calibration Z_s, to the final RFMG prediction $\widehat{Z}$. Examples use UniDepthV2 with $N = 16$ sparse observations under 40% corruption. Global calibration removes a substantial portion of the scene-level metric error, while evidence-aware local refinement further reduces spatially varying residual errors in the highlighted regions.