FOUNDATION-AGNOSTIC METRIC DEPTH REFINEMENT VIA SPARSE SEMANTIC GUIDANCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advancements in depth estimation rely heavily on powerful foundation models. While these massive networks capture robust scene representations, their real-world applications still require sparse depth guidance to resolve absolute scale and local precision. However, integrating this guidance remains highly problematic because existing refinement paradigms suffer from severe architectural entanglement. They are tightly coupled to specific network designs and require costly retraining for any new foundation architecture. To break this barrier, we present FreeDepth, the first universally decoupled framework for foundation-agnostic metric depth refinement. Our method achieves true decoupling by mapping sparse inputs into a universal semantic space through anchor-relative log-depth representations. This unified signal then seamlessly modulates the shared spatial topology of any heterogeneous network without relying on specific architectural priors. To guarantee absolute robustness against localized feed-forward errors, this decoupled pipeline is complemented by test-time optimization gated by gradient compatibility, while keeping all primary weights strictly frozen. Extensive experiments across five datasets demonstrate that FreeDepth establishes a new state-of-the-art performance. Remarkably, despite being trained solely via frozen interfaces of DA3 and InfiniDepth, the identical corrector demonstrates strong zero-shot transferability to unseen foundation models, such as π^3 and VGGT. All code and models will be publicly released.

1 INTRODUCTION

Accurate metric depth estimation serves as the indispensable cornerstone for physically grounded applications, ranging from precise 3D reconstruction to embodied autonomous navigation. Driven by massive training corpora and visual foundation models, recent depth predictors demonstrate remarkable capabilities in recovering dense scene layouts and sharp object boundaries with unprecedented zero-shot generalization (Ranftl et al., 2021; Yang et al., 2024a;b). However, this triumph is largely confined to relative or affine-invariant depth. By discarding absolute scale and shift, these models produce visually pleasing but physically unscaled geometry, rendering them structurally inadequate for downstream tasks that demand strict metric precision. While pioneering efforts attempt to predict metric depth directly, they frequently struggle to maintain scale accuracy and cross-domain stability in open-world environments (Bhat et al., 2023; Yin et al., 2023; Hu et al., 2024; Piccinelli et al., 2024; Bochkovskii et al., 2025). The fundamental bottleneck lies in the inherent ambiguity of recovering absolute metric scale across diverse environments, especially when camera intrinsics and geometric priors vary drastically between training and deployment domains.

To resolve this fundamental ambiguity, a promising paradigm anchors the dense geometric priors of foundation models using sparse metric observations derived from LiDAR or Structure-from-Motion (SfM). These two signals exhibit natural complementarity. Sparse observations offer highly reliable absolute scale but lack fine-grained completeness, whereas pretrained foundation models capture rich structural details yet suffer from metric distortion. Leveraging this synergy, recent literature explores injecting sparse depth as decoder prompts, fusing incomplete metric priors with dense relative structures, or performing test-time optimization on internal model states (Lin et al., 2026; Wang et al., 2025d; Viola et al., 2025; Ke et al., 2026). These pioneering efforts demonstrate that even a minimal set of metric anchors can effectively calibrate dense geometric priors to yield complete and highly accurate metric depth maps.

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.

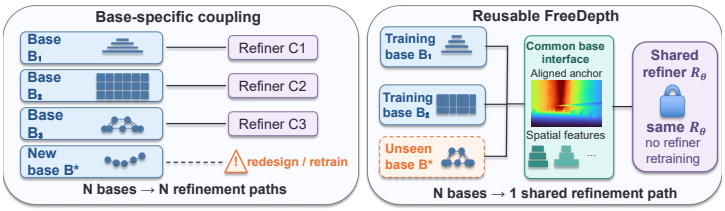

Figure 1: FreeDepth replaces base-specific refinement architectures with one shared corrector.

Despite their empirical success, existing sparse-guided methods are typically built around a specific foundation. These methods might inject sparse prompts into specific decoder layers, restrict adaptation parameters to particular tokens, network modules, or low-rank states, and implicitly assume the depth-output convention of that model for prompt definitions (Lin et al., 2026; Park et al., 2024a; Chung et al., 2025; Jeong et al., 2025). Consequently, when the original foundation is replaced by a new model with a different architecture, feature hierarchy, or output parameterization, the existing refiner usually cannot be reused directly and instead requires interface redesign, model retraining, or both. As illustrated in Fig. 1, although depth foundations such as VGGT, π^3, and DA3 continue to emerge (Wang et al., 2025a; 2026; Lin et al., 2025), this base-specific coupling prevents an existing sparse metric correction capability from directly benefiting from progress in new foundations.

To achieve true foundation-agnostic reusability, a universal refinement framework must navigate a complex triad of challenges. Diverse foundation models project depth into distinct normalized spaces, meaning an identical numerical residual carries entirely different geometric semantics across architectures. Furthermore, while relearning dense context from RGB imposes prohibitive computational overhead, extracting intermediate features from heterogeneous foundations introduces severe representation misalignment. Even when these issues are resolved, purely feed-forward networks inherently struggle to perfectly correct highly complex local boundaries. Motivated by these intertwined challenges, we introduce FreeDepth, a universally decoupled refinement architecture. We unify correction semantics across arbitrary bases by formulating sparse-aligned anchors and anchor-relative log-depth innovations. Rather than discarding misaligned foundation features, our CKA analysis (Kornblith et al., 2019) reveals that they retain a partially shared spatial topology. We exploit this intrinsic property through projected multiscale features and hierarchical FiLM modulation (Perez et al., 2017) to efficiently propagate metric information. Finally, for challenging samples that evade feed-forward correction, we design a gradient-compatibility gate. This mechanism selectively triggers LoRA-based test-time optimization (Hu et al., 2021) on those samples, keeping both the foundation model and the primary corrector strictly frozen.

Extensive empirical evaluations across five datasets encompassing indoor, outdoor, and autonomous driving scenarios validate the efficacy of FreeDepth. To rigorously assess its true cross-model reusability, we introduce a novel four base testing protocol. We train the shared corrector on Hypersim and VKITTI2 through frozen DA3 and InfiniDepth interfaces, then transfer it to held-out π^3 and VGGT without updating the corrector or its projections. Furthermore, a mixed-feature protocol successfully integrates representation stages from entirely different foundations. Empirically, FreeDepth establishes a new state of the art, securing an exceptional average rank of **1.3** across all benchmarks. It achieves top performance in seven out of ten primary evaluation metrics and secures second place in the remaining three. This zero-shot reuse improves the sparse-aligned predictions of both held-out foundations on ScanNet and KITTI.

In conclusion, our contributions are threefold:

- We identify the severe architectural entanglement inherent in existing sparse guided depth refinement paradigms and propose FreeDepth. This represents the first universally decoupled framework capable of seamlessly upgrading alongside diverse foundation models.
- We achieve foundation-agnostic refinement by separating sparse metric correction from the active foundation. A shared interface unifies correction semantics and spatial feature representations across heterogeneous models, enabling the same corrector to be reused without retraining.

- We propose a novel four foundation testing protocol to rigorously evaluate true cross-model reusability. Extensive experiments demonstrate that FreeDepth achieves an exceptional average rank of 1.3 across five diverse datasets and exhibits remarkable zero-shot transfer capabilities on unseen foundation architectures.

2 Related Work

Depth and visual-geometry foundation models. Monocular depth estimation has progressed from transformer predictors such as DPT to data-scaled Depth Anything models (Ranftl et al., 2021; Yang et al., 2024a;b). ZoeDepth, Metric3D, UniDepth, Depth Pro, and MoGe extend this progress toward metric scale, camera-aware prediction, and open-domain use (Bhat et al., 2023; Yin et al., 2023; Hu et al., 2024; Piccinelli et al., 2024; Bochkovskii et al., 2025; Wang et al., 2025b;c). Visual-geometry foundations further estimate point maps, cameras, and multiview structure, including VGGT, π^3, DA3, and InfiniDepth (Wang et al., 2025a; 2026; Lin et al., 2025; Yu et al., 2026). These foundations serve different objectives and use different architectures, thus, their predictions and intermediate representations naturally take different forms. But for sparse-guided depth, a corrector that directly depends on one foundation's prediction convention or feature layout cannot readily inherit improvements from a new foundation through simple substitution. This motivates a correction mechanism whose learned behavior remains reusable as foundations evolve.

Sparse-guided metric depth. Classical depth completion fuses RGB with irregular metric measurements through sparsity-aware convolutions, encoder-decoder fusion, or learned propagation (Uhrig et al., 2017; Ma & Karaman, 2018; Cheng et al., 2018; Park et al., 2020). CompletionFormer and subsequent methods strengthen global context, propagation, and robustness to varying point density (Youmin et al., 2023; Zuo & Deng, 2024; Park et al., 2024c; Hu et al., 2026). More recent methods preserve a pretrained dense prior while introducing sparse prompts, incomplete metric priors, or sparse-aware pretraining (Park et al., 2024b; Lin et al., 2026; Wang et al., 2025d; Ma et al., 2026). OMNI-DC learns multiresolution integration, whereas Marigold-DC and affine-prior alignment impose sparse constraints at test time (Zuo et al., 2025; Viola et al., 2025; Hyoseok et al., 2025). Prior Depth Anything permits replacement of its monocular prior, but its refinement pipeline still retains an external Depth Anything V2 encoder (Wang et al., 2025d; Yang et al., 2024b). These approaches establish the value of sparse guidance, yet their correction remains tied to a prompt interface, fusion architecture, auxiliary encoder, or per-image optimization procedure. Replacing the dense prior can consequently require redesign, retraining, or renewed optimization. In contrast, our FreeDepth externalizes sparse correction behind a shared interface to the aligned prediction and projected multiscale features, allowing one corrector to be reused as the dense prior changes.

3 Method

Premise. Sharing a corrector across heterogeneous foundations requires their projected features to capture similar content. We use linear CKA (Kornblith et al., 2019) for relative comparison, with higher scores in $[0, 1]$ indicating greater similarity. The CKA of 0.258 between InfiniDepth's shallow and deep features serves as our reference. Tab. 1 compares six pairs among DA3, InfiniDepth, π^3, and VGGT. Mean CKA at corresponding shallow, middle, and deep stages is 0.595, 0.771, and 0.742, respectively. These higher cross-model scores support a shared multiscale corrector. Held-out transfer and feature mixing later test its practical reusability.

Table 1: Linear CKA on KITTI (InfD: InfiniDepth). Cross-foundation columns compare corresponding stage groups. The last row is a within-foundation cross-depth reference.

Stage	π^3/VGGT	InfD/VGGT	DA3/VGGT	DA3/InfD	InfD/π^3	DA3/π^3
Shallow↑	0.895	0.385	0.563	0.741	0.428	0.556
Middle↑	0.781	0.766	0.770	0.862	0.697	0.749
Deep↑	0.813	0.769	0.689	0.645	0.833	0.704
InfiniDepth shallow–deep: 0.258						

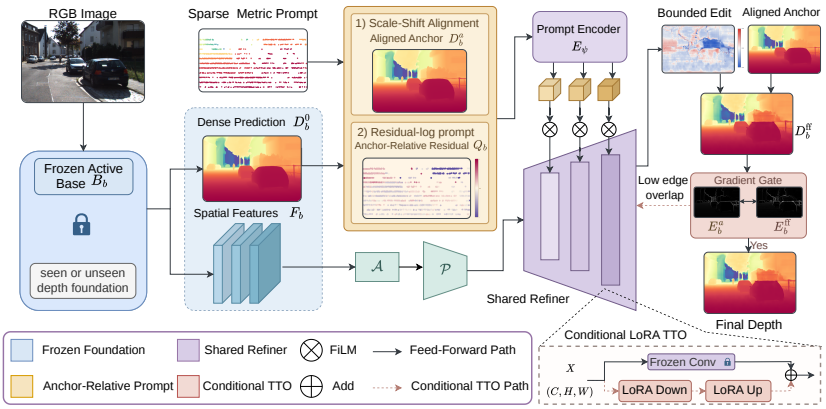

Figure 2: Overview. FreeDepth separates sparse metric correction from the active foundation. The feed-forward path combines an aligned depth anchor with projected multiscale features, and the compatibility gate routes difficult samples to LoRA TTO.

Method Overview. Given an RGB image and sparse metric depth observations, FreeDepth produces a dense metric depth map by correcting the prediction of an active, frozen depth foundation. As Fig. 2 illustrates, it aligns the foundation prediction with sparse observations, then combines the remaining discrepancy with projected multiscale features to predict a bounded dense correction. This feed-forward corrector is shared across foundations through a minimal prediction-and-feature interface. A gradient-based gate routes difficult samples to LoRA test-time optimization (TTO).

3.1 A Minimal Interface to Evolving Foundations

Let $\mathbb{B}_{\text{train}} = \{\mathcal{B}_k\}_{k=1}^{K}$ denote the frozen foundations, such as DA3, InfiniDepth, used to train the corrector. For an RGB image $I \in \mathbb{R}^{3\times H\times W}$, projecting sparse LiDAR measurements onto the camera image plane yields a 2D metric depth map $S \in \mathbb{R}^{H\times W}$ and validity mask $M \in \{0,1\}^{H\times W}$. We write $\Omega = \{p \mid M(p) = 1, S(p) > 0\}$ for the observed pixels. An active foundation $\mathcal{B}_b$ returns a dense prediction and L ordered feature maps:

$$\left(D_b^0, \mathcal{F}_b\right) = \mathcal{B}_b(I), \qquad \mathcal{F}_b = \{F_b^\ell\}_{\ell=1}^{L}, \tag{1}$$

where $D_b^0 \in \mathbb{R}^{H\times W}$ is the depth map after the standard output conversion of foundation b, and $F_b^\ell \in \mathbb{R}^{c_b^\ell}$ is its stage-ℓ feature. FreeDepth freezes $\mathcal{B}_b$ and uses only these outputs.

Because c_b^ℓ varies across foundations, a parameter-free adapter $\mathcal{A}_{c_b^\ell \to c_*}$ maps each stage to width c_* before the shared stage projection $\mathcal{P}^\ell$:

$$G_b^\ell = \mathcal{P}^\ell\left(\mathcal{A}_{c_b^\ell \to c_*}(F_b^\ell)\right), \tag{2}$$

where G_b^ℓ is the projected stage-ℓ feature used by the corrector. For wider inputs, the adapter selects c_* channels at random without replacement, following the idea of random filter selection (Wang et al., 2021). The selection is fixed across samples. For narrower inputs, it repeats channels evenly. The mapping may discard information but has no learned parameters. Sec. 4.3 tests whether the retained cues remain useful.

3.2 Multiscale Conditional Corrector

Anchor-relative sparse innovation. Raw predictions may differ in scale and offset. We therefore align D_b^0 to the sparse observations before predicting local corrections. Here $\mathcal{C}$ fits one scale and shift per image by least squares over Ω. The aligned anchor D_b^a retains the dense structure of the foundation prediction. With $\log_\epsilon x = \log(\max(x, \epsilon))$, the anchor and remaining discrepancy are

$$D_b^a = \mathcal{C}(D_b^0; S, M), \tag{3}$$

$$Q_b(p) = \begin{cases} \log_\epsilon S(p) - \log_\epsilon D_b^a(p), & p \in \Omega, \\ 0, & p \notin \Omega. \end{cases} \tag{4}$$

At an observed pixel, Q_b is the stabilized log ratio between measured and anchor depth, encoding correction direction and multiplicative magnitude independently of the original scale. It is zero elsewhere, with M distinguishing missing observations from a zero discrepancy. A prompt encoder $\mathcal{E}_\psi$ maps $X_b = [Q_b, M, \log_\epsilon D_b^a]$ to P_b. The shared corrector is

$$\widehat{D}_b^{\text{ff}} = \mathcal{R}_\theta\big(\{G_b^\ell\}_{\ell=1}^L, D_b^a, Q_b, M\big)\,, \tag{5}$$

where θ includes the prompt encoder but no foundation parameters. "ff" denotes pre-TTO output.

Partially shared structural conditioning. Sparse innovations specify depth corrections only at observed points. The corrector propagates these corrections to neighboring unobserved pixels using intermediate foundation features that provide cues about boundaries, surface layout, and object extent. Intermediate representations vary across foundations and can still capture similar spatial structure. We examine this correspondence with CKA (Kornblith et al., 2019) and test corrector reuse in the transfer experiments.

FreeDepth uses a DPT-style decoder (Ranftl et al., 2021) to fuse multiscale features. At stage ℓ, it transforms G_b^ℓ into H_b^ℓ and resizes and projects P_b to P_b^ℓ. FiLM (Perez et al., 2017) conditions these spatial features on the sparse correction prompt:

$$\widetilde{H}_b^\ell = H_b^\ell \odot \big(1 + \gamma_\ell(P_b^\ell)\big) + \beta_\ell(P_b^\ell), \tag{6}$$

where γ_ℓ and β_ℓ predict prompt-dependent scale and bias, and $\odot$ is element-wise multiplication. The decoder fuses these conditioned states from coarse to fine, linking the correction to scene layout and boundaries. Each training sample uses one complete foundation hierarchy, while the corrector is shared across foundations.

Bounded anchor edit. The decoder predicts a bounded log-depth edit, not depth from scratch:

$$\Delta_b = \delta_{\max} \tanh Z_b, \qquad \widehat{D}_b^{\text{ff}} = D_b^a \odot \exp(\Delta_b), \tag{7}$$

where Z_b is the decoder output and $\delta_{\max} > 0$ bounds the correction. The multiplicative form preserves positive depth and limits departure from the anchor.

Given ground-truth metric depth D^*, we train the corrector with

$$\mathcal{L}_{\text{train}} = \mathcal{L}_{\text{AbsRel}}(\widehat{D}_b^{\text{ff}}, D^*) + \lambda_g \mathcal{L}_{\text{grad}}(\widehat{D}_b^{\text{ff}}, D^*), \tag{8}$$

$$\mathcal{L}_{\text{grad}} = \Big\| \nabla \log_\epsilon \widehat{D}_b^{\text{ff}} - \nabla \log_\epsilon D^* \Big\|_1 . \tag{9}$$

Here $\mathcal{L}_{\text{AbsRel}}$ averages relative absolute error over valid pixels, $\lambda_g \geq 0$ weights the log-gradient loss, and ∇ is the two-dimensional spatial gradient. The two terms supervise metric accuracy and local structure.

3.3 Gradient-Gated LoRA Test-time Optimization

When sparse observations indicate structure weak or absent in the anchor, the required correction is harder. FreeDepth detects such cases from prediction–anchor structural disagreement and routes only them to a low-rank update.

Gradient-compatibility gate. Let E_b^{ff} and E_b^a contain the pixels with the top 20% log-depth gradient magnitudes in $\widehat{D}_b^{\text{ff}}$ and D_b^a. Their Jaccard overlap defines the gate:

$$g_b = \frac{|E_b^{\text{ff}} \cap E_b^a|}{|E_b^{\text{ff}} \cup E_b^a|}, \qquad \tau_b = \mathbb{I}[g_b < \eta]. \tag{10}$$

Here $g_b \in [0, 1]$ measures compatibility and $\tau_b \in \{0, 1\}$ is the routing decision at threshold η. Low overlap means that the sparse-guided output introduces substantial structure beyond the anchor, suggesting that the observations contain geometry missed by the foundation and impose a harder correction. The gate uses this disagreement as a difficulty signal without accessing ground truth.

Low-rank repair objective. When $\tau_b = 1$, the foundation and original corrector remain frozen. FreeDepth inserts convolutional LoRA branches into the prompt encoder and FiLM path (Hu et al., 2021):

$$Y = \mathrm{Conv}_{\mathrm{base}}(X) + \frac{\alpha}{r}\,\mathrm{Conv}_{\mathrm{up}}\big(\mathrm{Conv}_{\mathrm{down}}(X)\big). \tag{11}$$

Here X and Y are the layer input and output. Only the rank-r down–up branch is optimized. α/r scales its contribution, and its parameters are denoted by ϕ. Let $d_\phi = \log_\epsilon \widehat{D}_b^{\mathrm{tto}}(\phi)$, $d_a = \log_\epsilon D_b^a$, $d_f = \log_\epsilon \widehat{D}_b^{\mathrm{ff}}$, and $d_o(p) = \log_\epsilon S(p)$ on Ω. Let $\bar{\Omega}$ be the unobserved pixels. Defining $\ell_A(x, y) = |A|^{-1} \sum_{p\in A} |x(p) - y(p)|$, the objective is

$$\mathcal{L}_{\mathrm{TTO}}(\phi) = \ell_\Omega(d_\phi, d_o) + \lambda_a \ell_{\bar{\Omega}}(d_\phi, d_a) + \lambda_f \ell_{\bar{\Omega}}(d_\phi, d_f). \tag{12}$$

The first term enforces sparse measurements. On unobserved pixels, the other terms retain the anchor and feed-forward prediction, preventing TTO from satisfying sparse points by discarding either source. Separate region means balance sparse and dense support.

At TTO step t, we monitor the relative objective improvement

$$\upsilon_t = \frac{\mathcal{L}_{\mathrm{TTO}}^{t-1} - \mathcal{L}_{\mathrm{TTO}}^{t}}{|\mathcal{L}_{\mathrm{TTO}}^{t-1}| + \epsilon}. \tag{13}$$

where $\epsilon > 0$ prevents division by zero. After a minimum update budget, optimization stops upon sustained low improvement, sufficient sparse consistency, or the maximum step count. The temporary state ϕ is reset for each sample, leaving the foundation and reusable corrector unchanged.

3.4 Implementation Details

Training. FreeDepth is trained on Hypersim and VKITTI2 (Roberts et al., 2021; Cabon et al., 2020) through frozen DA3 and InfiniDepth interfaces (Lin et al., 2025; Yu et al., 2026). Each training pass uses four ordered stages from one active foundation. The parameter-free adapter maps every stage to $c_* = 1024$ channels before learned projection and coarse-to-fine DPT fusion (Ranftl et al., 2021). Only the corrector and stage projections are optimized. Foundations and the channel adapter remain fixed.

Test-time optimization. Only the LoRA parameters in the prompt encoder and FiLM path are optimized. The gate study uses AdamW with learning rate 5×10^{-4} and compatibility threshold $\eta = 0.5$ (Loshchilov & Hutter, 2019). See the appendix for update budgets and stopping thresholds.

4 Experiments

We test whether one FreeDepth corrector remains accurate across domains and reusable when the active foundation changes. After the main comparison, held-out foundations, mixed feature hierarchies, and channel-width variants stress different forms of reuse. Component and TTO studies then identify which design choices support accuracy and when per-sample adaptation is useful.

4.1 Experimental Setup

Datasets and sparse observations. We evaluate on DIODE, ScanNet, KITTI, Waymo, and NYUv2 (Vasiljevic et al., 2019; Dai et al., 2017; Geiger et al., 2013; Sun et al., 2020; Silberman et al., 2012), covering indoor, outdoor, and driving scenes. Sparse observations are synthesized from ground truth using a pattern that emulates projected 16-beam LiDAR scanlines, followed by small perturbations at valid samples. All methods are evaluated under the same protocol. Additional preprocessing details are provided in the appendix.

Training and transfer protocol. The shared corrector is trained only through frozen DA3 and InfiniDepth interfaces (Lin et al., 2025; Yu et al., 2026). For direct transfer, the same frozen corrector and projections are attached to held-out π^3 or VGGT interfaces without further learning (Wang et al., 2026; 2025a). We also train separate correctors through π^3 and VGGT to provide interface-trained references. Comparing each sparse-aligned anchor with its feed-forward output isolates correction beyond per-image metric alignment.

Table 2: Unified-protocol comparison on five benchmarks. FreeDepth uses one feed-forward pass without TTO. AbsRel is in percent. Rows follow average rank. Bold/underline mark first/second.

Method	DIODE		ScanNet		KITTI		Waymo		NYUv2		Rank↓
	Abs.↓	RMSE↓	Abs.↓	RMSE↓	Abs.↓	RMSE↓	Abs.↓	RMSE↓	Abs.↓	RMSE↓	
DA3 Metric	7.28	0.562	3.91	0.156	6.27	2.725	11.24	4.577	4.37	0.248	6.9
DA3-aligned	6.10	0.545	3.94	0.166	6.64	2.643	10.38	4.169	4.24	0.246	6.6
Marigold-DC	14.05	1.263	3.37	0.157	6.83	2.491	8.39	3.316	4.23	0.297	6.6
PriorDA	2.31	**0.324**	3.46	0.166	5.47	1.678	2.70	2.465	4.06	0.224	4.4
InfiniDepth	<u>2.23</u>	0.616	**1.16**	0.108	2.09	2.737	2.16	3.195	<u>1.15</u>	0.143	3.8
CAPA	2.59	0.380	1.39	<u>0.095</u>	2.05	<u>1.491</u>	3.21	2.603	1.39	<u>0.128</u>	3.3
OMNI-DC	2.24	0.373	1.54	0.112	<u>2.04</u>	1.718	**1.96**	<u>2.292</u>	1.57	0.162	<u>3.1</u>
FreeDepth	**2.16**	<u>0.329</u>	<u>1.30</u>	**0.091**	**1.74**	**1.262**	<u>2.00</u>	**1.656**	**1.03**	**0.095**	**1.3**

Evaluation protocol and metrics. The comparison covers metric foundations, feed-forward completion, and optimization-based methods (Yu et al., 2026; Ke et al., 2026; Zuo et al., 2025; Wang et al., 2025d; Viola et al., 2025). All methods use identical splits, masks, sparse observations, and metric code. Tab. 2 reports one feed-forward prediction per input with TTO disabled. We report AbsRel and RMSE, where lower is better, and average each method's rank over the ten dataset–metric cells. Distribution and TTO analyses additionally use $\delta_{\mathrm{rel}<1\%}$, the percentage of valid pixels with relative error below 1%.

4.2 Overall Sparse-guided Refinement Accuracy

Accuracy across domains. Tab. 2 asks whether reuse compromises refinement accuracy. FreeDepth ranks first in seven of ten dataset–metric cells and second in the other three, yielding the best average rank of 1.3 across indoor, outdoor, and driving scenes. Fig. 3 further shows that the corrected predictions retain thin structures and depth discontinuities. The shared feed-forward corrector is therefore competitive before any foundation is replaced.

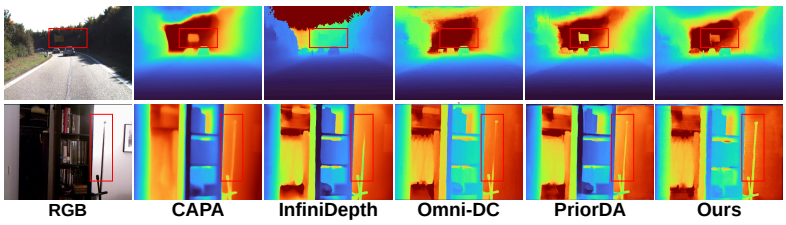

Figure 3: Qualitative results under the protocol of Tab. 2. Boxes mark thin structures, boundaries, and incomplete local geometry.

4.3 Reuse under Foundation Replacement

Held-out foundation interfaces. The left block of Tab. 3 gives interface-trained references. The right block attaches the DA3 and InfiniDepth-trained corrector to held-out π^3 and VGGT without updates. Every held-out anchor improves on DIODE, ScanNet, and KITTI, matching or surpassing the corresponding interface-trained references on ScanNet and KITTI. Because both the dense prediction and four feature stages change, these gains provide functional transfer evidence beyond CKA. Fig. 4 also shows that refinement preserves differences among foundation predictions rather than collapsing them to one output.

Diagnostic feature mixing. CKA establishes the representational premise for sharing the corrector. Feature mixing tests whether the corresponding content remains usable when a coherent single-foundation hierarchy is deliberately broken. In Tab. 4, the uppercase letter denotes the anchor and the four italic subscripts denote shallow-to-deep feature sources, where d, p, and v denote DA3, π^3, and VGGT. Thus D_{ddpp} uses a DA3 anchor, two shallow DA3 stages, and two deep π^3 stages.

Table 3: Cross-foundation AbsRel (%)↓. "Align." is the sparse-aligned anchor and "Final" is the feed-forward output. The right group directly reuses the frozen corrector on held-out foundations.

		Interface-trained references				*Held-out direct reuse*	
Dataset	Output	DA3	InfiniDepth	π^3	VGGT	π^3	VGGT
DIODE	Align.	6.10	22.19	7.58	9.61	7.58	9.61
	Final	2.16	2.42	2.75	2.77	2.91	2.86
ScanNet	Align.	3.94	21.40	2.29	2.88	2.29	2.88
	Final	1.30	2.14	1.24	1.22	1.07	1.12
KITTI	Align.	6.64	23.16	7.51	13.88	7.51	13.88
	Final	1.74	2.61	3.02	3.58	2.07	3.11

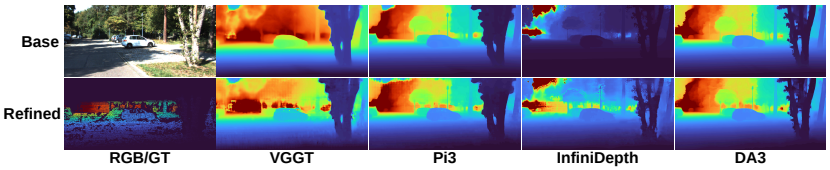

Figure 4: The same corrector retains foundation-specific geometry while improving predictions across active bases.

All mixtures remain functional without retraining, including the alternating D_{dpdp} hierarchy. D_{ppdd} leads on DIODE, KITTI, and Waymo, whereas V_{vvpp} leads on ScanNet, so feature sources are not interchangeable. Rather, the result shows that the corrector can use shared spatial content from heterogeneous projected features, complementing the statistical evidence from CKA.

Reusing projections across models with different channel widths. Tab. 5 varies the DA3 source width while fixing all learned projections and the corrector. All four variants connect without retraining, showing that channel selection or repetition retains usable spatial cues. DA3-G improves over DA3-S on every dataset and leads four of five, while DA3-L leads on KITTI. The corrector therefore supports width changes and generally benefits from stronger DA3 variants.

Table 4: AbsRel (%)↓ for fixed mixed-feature interfaces. Bold and underline denote first and second.

Interface	DIODE	ScanNet	KITTI	Waymo	NYUv2
D_{ddpp}	2.53	1.40	2.22	2.35	1.03
D_{dpdp}	<u>2.48</u>	1.38	<u>2.11</u>	2.20	**1.02**
D_{ppdd}	**2.44**	1.33	**1.95**	**2.11**	<u>1.02</u>
D_{ddvv}	3.52	1.70	3.04	2.43	1.33
D_{vvdd}	3.03	1.52	2.54	<u>2.16</u>	1.26
V_{ppvv}	3.34	<u>1.30</u>	3.26	4.50	1.13
V_{vvpp}	3.27	**1.21**	3.03	4.10	1.05

Table 5: AbsRel (%)↓ for four DA3 source widths mapped to the shared 1024-channel interface.

Dataset / Foundation	DA3-S	DA3-B	DA3-L	DA3-G
DIODE	3.39	3.28	<u>2.97</u>	**2.83**
ScanNet	1.83	1.70	<u>1.63</u>	**1.50**
KITTI	2.95	<u>2.38</u>	**2.23**	2.39
Waymo	2.82	2.65	<u>2.56</u>	**2.26**
NYUv2	1.44	1.32	<u>1.24</u>	**1.07**

4.4 Component Ablation

In Tab. 6, columns ③–⑤ vary prompt coordinates, columns ⑥–⑧ vary prompt injection strategy, and column ⑨ replaces the bounded edit with direct depth prediction. The anchor-only column ② separates metric alignment from learned correction.

Correction coordinates, conditioning, and output action. FreeDepth ① substantially improves over the aligned anchor ②, so alignment alone does not explain its accuracy. Among prompt coordinates, direct log or raw depth is unstable across ScanNet and KITTI, whereas the anchor-relative log ratio is consistently strongest. Hierarchical FiLM is better balanced across both datasets than output-only, additive, or concatenation variants, and direct depth prediction ⑨ performs worst. To-

Table 6: Component ablations on ScanNet and KITTI. Adjacent column groups vary prompt coordinates, feature injection, and output action. AbsRel is in percent.

				Prompt coordinates			Feature injection			Output
Dataset	Metric	① **Full FreeDepth**	② Anchor only	③ Direct log	④ Linear residual	⑤ Direct depth	⑥ Output-only FiLM	⑦ Additive	⑧ Concat fusion	⑨ Direct-depth
ScanNet	Abs.%↓	1.30	3.94	1.86	2.01	3.09	3.70	1.98	1.68	6.83
	RMSE↓	0.091	0.166	0.095	0.098	0.125	0.143	0.107	0.091	0.209
KITTI	Abs.%↓	1.74	6.64	9.46	4.16	38.54	3.42	3.09	3.82	41.22
	RMSE↓	1.262	2.643	3.382	2.057	15.193	1.689	1.854	1.794	7.911

gether, the results support anchor-relative prompting, multiscale conditioning, and bounded editing as complementary choices.

The gain holds across samples. Fig. 5 checks whether the mean gains are concentrated in favorable images. The 5th percentile of full FreeDepth exceeds the median of every ablation, while direct-log and direct-depth prompts show low medians and long upper tails. The full design therefore improves the lower tail as well as the average.

4.5 Effectiveness of Gradient-gated TTO

Selective optimization of difficult samples. Tab. 7 evaluates routing selectivity and update quality. Triggered on only 17.10–27.98% of inputs, the gate covers 62.38–78.79% of the worst pre-TTO decile and 50.25–59.54% of the worst quintile. On routed samples, TTO reduces AbsRel by 8.46–13.04%, raises $\delta_{\mathrm{rel}<1\%}$ by 4.94–5.63 points, and degrades only 1.64–4.68%. Thus, the gate targets useful adaptation to difficult samples. Ground truth is used only to compute these offline statistics.

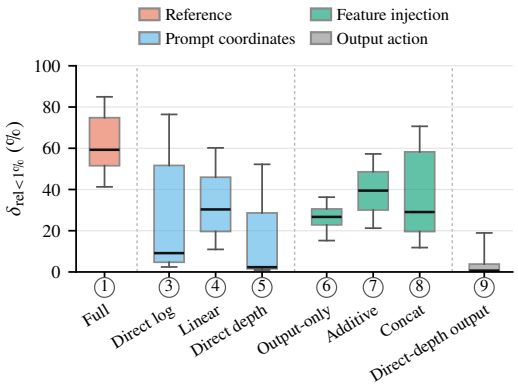

Figure 5: Per-sample $\delta_{\mathrm{rel}<1\%}$ for the independent variants in Tab. 6, pooled over ScanNet and KITTI ($n = 1800$ each). Boxes show quartiles and 5th–95th percentile whiskers.

Table 7: Gradient-gated LoRA TTO at learning rate 5×10^{-4}. Triggered statistics use only routed samples. Degradation means increased AbsRel after TTO. The last two rows report coverage of the worst pre-TTO subsets, and pp denotes percentage points.

Statistic	KITTI	NYUv2
Gate trigger rate (%)	17.10	27.98
Triggered AbsRel gain (%)	8.46	13.04
Triggered $\delta_{\mathrm{rel}<1\%}$ gain (pp)	5.63	4.94
Triggered AbsRel degradation rate (%)	4.68	1.64
Worst 10% gate rate (%)	62.38	78.79
Worst 20% gate rate (%)	50.25	59.54

5 Conclusion and Discussion

FreeDepth separates sparse-guided correction from the foundation that supplies dense geometry. Sparse alignment and anchor-relative prompts make the correction request consistent, while projected multiscale features provide spatial context for a bounded feed-forward edit. FreeDepth achieves the best average rank across five datasets, and transfers to held-out foundations, mixed hierarchies, and different channel widths without retraining the corrector. Gradient-gated LoRA further targets difficult samples, supporting sparse correction that can benefit from evolving foundations.

Reproducibility Statement

The main paper describes FreeDepth's sparse alignment, shared corrector, and optional test-time optimization. Appendix A provides the foundation interfaces, corrector architecture, and inference procedure. Appendix B describes the training setup, datasets, sparse observation protocol, evaluation metrics, and timing protocol. Additional results and diagnostic analyses are provided in Appendix C. Code and models will be publicly released.

Ethics Statement

FreeDepth refines depth predictions using existing foundation models and sparse observations. Its use in robotics or autonomous driving requires application-specific validation, as depth errors can affect downstream decisions. Scene imagery and reconstructed geometry may also reveal personal information or sensitive spatial layouts. Use and release of data, pretrained models, and derived artifacts should respect the source licenses, access conditions, and privacy constraints.

AI Use Statement

In this work, we used generative AI tools to assist with translation, implementation of parts of the core method code, experimental design, and data analysis. We also used these tools for language editing and LaTeX formatting. The authors take responsibility for the final content of this work, including all AI-assisted text, code, analyses, and claims.

References

Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth, 2023. URL https://arxiv.org/abs/2302.12288.

Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R. Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second, 2025. URL https://arxiv.org/abs/2410.02073.

Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2, 2020. URL https://arxiv.org/abs/2001.10773.

Xinjing Cheng, Peng Wang, and Ruigang Yang. Depth estimation via affinity learned with convolutional spatial propagation network, 2018. URL https://arxiv.org/abs/1808.00150.

Younjoon Chung, Hyoungseob Park, Patrick Rim, Xiaoran Zhang, Jihe He, Ziyao Zeng, Safa Cicek, Byung-Woo Hong, James S. Duncan, and Alex Wong. Eta: Energy-based test-time adaptation for depth completion, 2025. URL https://arxiv.org/abs/2508.05989.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes, 2017. URL https://arxiv.org/abs/1702.04405.

A Geiger, P Lenz, C Stiller, and R Urtasun. Vision meets robotics: The kitti dataset. *Int. J. Rob. Res.*, 32 (11):1231–1237, September 2013. ISSN 0278-3649. doi: 10.1177/0278364913491297. URL https://doi.org/10.1177/0278364913491297.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

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, December 2024. ISSN 1939-3539. doi: 10.1109/tpami.2024.3444912. URL http://dx.doi.org/10.1109/TPAMI.2024.3444912.

Weiming Hu, Zhipeng Zhang, Bing Li, Houwen Peng, and Stephen Maybank. Ocean: Object-aware anchor-free tracking with matching-relation learning. *Machine Intelligence Research*, 23(3):565–592, 2026. ISSN 2731-538X. doi: 10.1007/s11633-026-1634-0. URL https://www.mi-research.net/en/article/doi/10.1007/s11633-026-1634-0.

Lee Hyoseok, Kyeong Seon Kim, Kwon Byung-Ki, and Tae-Hyun Oh. Zero-shot depth completion via test-time alignment with affine-invariant depth prior, 2025. URL https://arxiv.org/abs/2502.06338.

Chanhwi Jeong, Inhwan Bae, Jin-Hwi Park, and Hae-Gon Jeon. Test-time prompt tuning for zero-shot depth completion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9443–9454, October 2025.

Bingxin Ke, Qunjie Zhou, Jiahui Huang, Xuanchi Ren, Tianchang Shen, Konrad Schindler, Laura Leal-Taixé, and Shengyu Huang. Depth completion as parameter-efficient test-time adaptation, 2026. URL https://arxiv.org/abs/2602.14751.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited, 2019. URL https://arxiv.org/abs/1905.00414.

Haotong Lin, Sili Chen, Junhao Liew, Donny Y. Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views, 2025. URL https://arxiv.org/abs/2511.10647.

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, 2026. URL https://arxiv.org/abs/2412.14015.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. URL https://arxiv.org/abs/1711.05101.

Baorui Ma, Jiahui Yang, Donglin Di, Xuancheng Zhang, Jianxun Cui, Hao Li, Yan Xie, and Wei Chen. Metricanything: Scaling metric depth pretraining with noisy heterogeneous sources, 2026. URL https://arxiv.org/abs/2601.22054.

Fangchang Ma and Sertac Karaman. Sparse-to-dense: Depth prediction from sparse depth samples and a single image, 2018. URL https://arxiv.org/abs/1709.07492.

Hyoungseob Park, Anjali Gupta, and Alex Wong. Test-time adaptation for depth completion, 2024a. URL https://arxiv.org/abs/2402.03312.

Jin-Hwi Park, Chanhwi Jeong, Junoh Lee, and Hae-Gon Jeon. Depth prompting for sensor-agnostic depth estimation, 2024b. URL https://arxiv.org/abs/2405.11867.

Jinhyung Park, Yu-Jhe Li, and Kris Kitani. Flexible depth completion for sparse and varying point densities. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21540–21550, 2024c. doi: 10.1109/CVPR52733.2024.02035.

Jinsun Park, Kyungdon Joo, Zhe Hu, Chi-Kuei Liu, and In So Kweon. Non-local spatial propagation network for depth completion. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (eds.), *Computer Vision – ECCV 2020*, pp. 120–136, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58601-0.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer, 2017. URL https://arxiv.org/abs/1709.07871.

Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. UniDepth: Universal monocular metric depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction, 2021. URL https://arxiv.org/abs/2103.13413.

Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M. Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding, 2021. URL https://arxiv.org/abs/2011.02523.

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In Andrew Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid (eds.), *Computer Vision – ECCV 2012*, pp. 746–760, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. ISBN 978-3-642-33715-4.

Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Sheng Zhao, Shuyang Cheng, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset, 2020. URL https://arxiv.org/abs/1912.04838.

Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. Sparsity invariant cnns, 2017. URL https://arxiv.org/abs/1708.06500.

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, 2019. URL https://arxiv.org/abs/1908.00463.

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, 2025. URL https://arxiv.org/abs/2412.13389.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5294–5306, 2025a. doi: 10.1109/CVPR52734.2025.00499.

Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision, 2025b. URL https://arxiv.org/abs/2410.19115.

Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. Moge-2: Accurate monocular geometry with metric scale and sharp details, 2025c. URL https://arxiv.org/abs/2507.02546.

Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Permutation-equivariant visual geometry learning, 2026. URL https://arxiv.org/abs/2507.13347.

Zehan Wang, Siyu Chen, Lihe Yang, Jialei Wang, Ziang Zhang, Hengshuang Zhao, and Zhou Zhao. Depth anything with any prior, 2025d. URL https://arxiv.org/abs/2505.10565.

Zi Wang, Chengcheng Li, and Xiangyang Wang. Convolutional neural network pruning with structural redundancy reduction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14913–14922, 2021. URL https://arxiv.org/abs/2104.03438.

Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, 2024a.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv:2406.09414*, 2024b.

Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. Metric3d: Towards zero-shot metric 3d prediction from a single image, 2023. URL https://arxiv.org/abs/2307.10984.

Zhang Youmin, Guo Xianda, Poggi Matteo, Zhu Zheng, Huang Guan, and Mattoccia Stefano. Completionformer: Depth completion with convolutions and vision transformers, 2023. URL https://arxiv.org/abs/2304.13030.

Hao Yu, Haotong Lin, Jiawei Wang, Jiaxin Li, Yida Wang, Xueyang Zhang, Yue Wang, Xiaowei Zhou, Ruizhen Hu, and Sida Peng. Infinidepth: Arbitrary-resolution and fine-grained depth estimation with neural implicit fields, 2026. URL https://arxiv.org/abs/2601.03252.

Yiming Zuo and Jia Deng. Ogni-dc: Robust depth completion with optimization-guided neural iterations, 2024. URL https://arxiv.org/abs/2406.11711.

Yiming Zuo, Willow Yang, Zeyu Ma, and Jia Deng. Omni-dc: Highly robust depth completion with multiresolution depth integration, 2025. URL https://arxiv.org/abs/2411.19278.

APPENDIX

This supplement gives more detail on FreeDepth's architecture, training setup, sparse-depth sampling, datasets, evaluation metrics, and runtime protocol. It also includes qualitative examples and additional results on alignment, foundation transfer, mixed and width-varying features, representation similarity, ablations, and the optional TTO path. Unless noted otherwise, all results use a single feed-forward pass with TTO disabled.

A METHOD DETAILS

A.1 FOUNDATION INTERFACE AND CHANNEL ADAPTATION

FreeDepth receives a positive dense depth prediction and four ordered spatial feature stages from the active foundation. Each wrapper returns patch-token tensors from shallow to deep and removes foundation-specific class, register, or camera tokens before passing them to the shared corrector. Tab. 8 lists the interfaces used in our experiments. We train the corrector with DA3 and InfiniDepth, then attach the frozen model directly to π^3 or VGGT for held-out evaluation, with no foundation-specific projection fitting.

Table 8: Foundation interfaces. "Train" marks foundations used to optimize the shared corrector. Held-out interfaces reuse it without learning.

Foundation	Train	Blocks	Status
DA3	yes	4, 11, 17, 23	seen
InfiniDepth	yes	4, 11, 17, 23	seen
π^3	no	3, 11, 23, 35	held out
VGGT	no	4, 11, 17, 23	held out

The shared stage projections expect width $c_* = 1024$. When an input has $c > c_*$ channels, the parameter-free adapter selects c_* of them without replacement using a deterministic seed-dependent permutation. When $c < c_*$, every channel is retained and the remaining positions are filled by balanced repetition. The seed affects only the order of repeated channels. One index map is shared across all spatial positions and samples, preserving the spatial grid without adding learned parameters. This fixed mapping remains effective across the tested range of 768–3072 channels.

A.2 EXACT CORRECTOR ARCHITECTURE

For each stage, FreeDepth removes non-patch tokens, applies layer normalization, and reshapes the remaining sequence to its two-dimensional patch grid. Tab. 9 traces the complete feed-forward path of the 1024-channel configuration. The four projected stages are resampled at factors 4, 2, 1, and $\frac{1}{2}$ relative to the patch grid, converted to a common width of 256, and fused from coarse to fine. The sparse path is computed once at image resolution and pooled separately for each decoder scale.

For a feature tensor with N_ℓ patch tokens, FreeDepth infers a two-dimensional grid whose aspect ratio matches the processed image, applies LayerNorm over channels, reshapes to $B \times c_* \times h_\ell \times w_\ell$, and then applies the stage-specific projection and resize operator in Tab. 9. The first two stages use transposed convolutions with scale factors four and two, the third keeps its grid, and the fourth uses a stride-two convolution. Lateral 3×3 convolutions map all four streams to 256 channels before top-down DPT fusion.

The sparse feature stays at image resolution until each decoder stage uses it. Adaptive average pooling captures smooth context, adaptive maximum pooling retains localized innovations, and a 1×1 convolution fuses the two responses. GroupNorm uses up to 16 groups, reducing the count when needed to divide the channel dimension, followed by SiLU. A global summary of the sparse feature predicts five sigmoid gates that scale the token maps before stage-specific Token FiLM. Zero-initialized Token FiLM projections and a learned residual scale initialized to 10^{-3} keep this path close to the identity at the start of training.

Table 9: Feed-forward corrector architecture. Spatial factors are relative to the patch grid.

Block	Architecture
Foundation stages	Four token hierarchies; non-patch tokens removed; LayerNorm; reshape to spatial grids
Stage projections	1×1 convolutions to 256, 512, 1024, and 1024 channels; spatial factors $4, 2, 1, \frac{1}{2}$
Decoder laterals	One 3×3 convolution per stage to a shared width of 256
Sparse encoder	Input $[Q, M, \log_\epsilon D^a]$; 3×3 convolution from 3 to 64 channels; two 64-channel residual blocks
Prompt pooling	Project 64 to 256 channels; adaptive average and maximum pooling; concatenate; 1×1 convolution from 512 to 256; GroupNorm and SiLU
Decoder fusion	Four 256-channel DPT fusion blocks, applied from the coarsest stage to the finest
Output head	256-channel residual block; 3×3 convolution to 128 channels; 1×1 convolution to depth edit and confidence

The prompt modulates all four projected stages and the final fused state, and the output head produces the bounded edit in Eq. 17. Excluding the frozen foundation, the corrector has 32,675,537 parameters. Routed TTO adds rank-8 LoRA branches to 27 convolutions in the prompt encoder, prompt projection, pooling fusion, and prompt FiLM. These branches contain 111,832 trainable parameters. The foundation and original corrector stay frozen.

A.3 Sparse Alignment and Prompt Construction

Let D^0 be the positive foundation prediction, and let S and M be the two-dimensional sparse metric map and validity mask obtained after projecting LiDAR measurements to the camera plane. The observed set is $\Omega = \{p \mid M(p) = 1, S(p) > 0\}$. FreeDepth fits a scale a and shift b independently for each image:

$$\begin{aligned}(a^*, b^*) &= \arg\min_{a,b} \sum_{p\in\Omega} \big(aD^0(p) + b - S(p)\big)^2, \\ D^a &= a^* D^0 + b.\end{aligned} \tag{14}$$

The sparse innovation is

$$Q = M \odot \big(\log_\epsilon S - \log_\epsilon D^a\big), \qquad \log_\epsilon x = \log(\max(x, \epsilon)). \tag{15}$$

We use $\epsilon = 10^{-6}$ for stabilized logarithms and denominator guards. The prompt encoder takes the three-channel tensor $[Q, M, \log_\epsilon D^a]$. A 3×3 stem convolution and two residual blocks produce a dense prompt feature, which is then projected to the decoder width. At each decoder scale, adaptive average and maximum pooling capture smooth context and strong local responses. Their outputs are concatenated and fused by a 1×1 projection.

A.4 Hierarchical Conditioning and Output Action

Each adapted foundation stage is normalized, mapped by a learned 1×1 projection, resized to its DPT scale, and fused from coarse to fine. The prompt feature modulates all four stages and the final decoder state through FiLM:

$$\widetilde{H}^\ell = H^\ell \odot \big(1 + \gamma_\ell(P^\ell)\big) + \beta_\ell(P^\ell). \tag{16}$$

The output head predicts a log-depth edit Z and confidence. The depth prediction is a bounded multiplicative action around the aligned anchor,

$$\widehat{D} = D^a \odot \exp(\delta_{\max} \tanh Z), \tag{17}$$

with $\delta_{\max} = 1.5$ in our experiments. This form preserves positive depth and makes the output action explicitly relative to the current foundation.

A.5 Gradient-Compatibility Gate and LoRA TTO

The feed-forward corrector is the default inference path and already performs well on its own. TTO is optional and can be disabled when single-pass latency or a simpler inference procedure is preferred.

The gate uses only the feed-forward output and sparse-aligned anchor. It computes horizontal and vertical gradients of log depth, retains the largest 20% of valid gradient magnitudes in each map, and forms binary edge sets E_f and E_a. Their Jaccard overlap is

$$g = \frac{|E_f \cap E_a|}{|E_f \cup E_a|}. \tag{18}$$

We use $\eta = 0.5$ and activate TTO when $g < \eta$, meaning that the prediction and anchor share fewer than half of the salient boundaries in their union. Such disagreement suggests that the sparse observations introduce structure missing from the anchor, making the correction more difficult. The gate uses this signal directly and requires no ground-truth error.

For routed samples, convolutional LoRA branches are inserted into the prompt encoder and prompt-modulation path. We use $r = 8$ and $\alpha = 4$, giving a LoRA scale of $\alpha/r = 0.5$. The TTO objective combines sparse consistency with anchor and feed-forward preservation on unobserved pixels, with $\lambda_a = \lambda_f = 0.5$. AdamW updates the LoRA parameters at learning rate 5×10^{-4} with zero weight decay.

For early stopping, let $d_\phi = \log_\epsilon \widehat{D}^{\mathrm{tto}}_\phi$ and $d_o = \log_\epsilon S$ on Ω. We denote the sparse-consistency term by

$$\mathcal{L}_s(\phi) = \ell_\Omega(d_\phi, d_o) = \frac{1}{|\Omega|} \sum_{p \in \Omega} |d_\phi(p) - d_o(p)|. \tag{19}$$

We set its stopping threshold to $\tau_s = 0.01$. Optimization begins from the checkpoint state for every sample and runs for at most 15 steps. After the first three steps, it stops when $\mathcal{L}_s < \tau_s$ or when relative objective improvement remains below 0.005 for two consecutive steps. The final prediction uses the parameter state with the lowest objective, after which the checkpoint state is restored.

Fixed choices. The reported configuration uses $K = 2$ training foundations, $L = 4$ feature stages, $c_* = 1024$, $\epsilon = 10^{-6}$, $\delta_{\max} = 1.5$, $\lambda_g = 2$, $\eta = 0.5$, $r = 8$, $\alpha = 4$, $\lambda_a = \lambda_f = 0.5$, and $\tau_s = 0.01$. The alignment scale and shift are fitted independently for each image. The FiLM scale and bias are learned functions of the conditioning input.

A.6 Feed-forward and TTO Execution

A feed-forward pass proceeds in four steps. First, the foundation processes the RGB image once and returns the dense prediction D^0 together with four frozen feature tensors. Second, the projected LiDAR observations form S and M. Per-image alignment produces the anchor D^a, and Eq. 15 converts the remaining sparse discrepancy into Q. Third, the deterministic channel adapter maps each feature stage to the shared 1024-channel interface. The corrector then fuses these features with the sparse prompt and predicts the unconstrained edit Z. Fourth, Eq. 17 bounds this edit and applies it multiplicatively to D^a, producing $\widehat{D}$. The prediction is then resized to the evaluation grid and restricted to the dataset-validity region.

The optional TTO path starts from this feed-forward result. The gate compares $\widehat{D}$ with D^a and either returns the feed-forward prediction or activates the sample-local LoRA branches. A routed sample reuses the previously computed foundation depth and feature tensors throughout optimization. Gradients update only the temporary LoRA state. The foundation, channel adapter, and original corrector remain fixed. The temporary state is discarded after the final prediction, so adaptation does not carry across samples.

B Experimental Details

B.1 Training Setup and Objective

We train the shared corrector on Hypersim and VKITTI2 using frozen DA3MONO and InfiniDepth interfaces. In the compact interface tables, DA3 refers to DA3MONO unless a size variant is named. Training runs for 200,000 steps on one NVIDIA A100 GPU with batch size 2 and learning rate 1×10^{-5}. Each batch uses one foundation interface and updates the shared corrector and learned stage projections.

For valid ground-truth pixels V, the training objective contains only AbsRel and log-depth gradient consistency, with $\lambda_g = 2$:

$$\mathcal{L}_{\text{train}} = \mathcal{L}_{\text{AbsRel}} + 2\mathcal{L}_{\text{grad}}, \tag{20}$$

$$\mathcal{L}_{\text{AbsRel}} = |V|^{-1} \sum_{p \in V} \frac{|\widehat{D}(p) - D^*(p)|}{D^*(p)}, \tag{21}$$

$$\mathcal{L}_{\text{grad}} = \sum_{q \in \{x,y\}} |V_q|^{-1} \sum_{p \in V_q} \left| \nabla_q \log_\epsilon \widehat{D}(p) - \nabla_q \log_\epsilon D^*(p) \right|. \tag{22}$$

Here V_q contains adjacent valid-pixel pairs along direction q. The AbsRel term enforces metric accuracy, and the gradient term preserves local depth changes and discontinuities.

B.2 Datasets and Evaluation Splits

Tab. 10 lists the evaluation sets, which span indoor scenes, mixed indoor and outdoor imagery, and autonomous-driving environments.

Table 10: Evaluation sets used by the experiments.

Dataset	Images	Evaluation content
DIODE	325	indoor depth
ScanNet	800	reconstructed indoor scenes
KITTI	1000	outdoor driving scenes
Waymo	496	front-camera projected LiDAR
NYUv2	654	indoor RGB-D scenes

B.3 Sparse Observation Protocol

For benchmarks with dense ground truth, we sample sparse depth using a LiDAR-16-style pattern that approximates scanlines from a low-beam LiDAR projected onto the camera plane. Sixteen nominal rows cover 10%–90% of the valid image height. Horizontally strided samples then receive small spatial and depth perturbations followed by random dropout. After invalid pixels are removed, the remaining observations form the 2D sparse map S and mask M. Every method uses the same samples within each benchmark.

B.4 Evaluation Protocol and Metrics

The experiments use nominal input resolutions of 1024×768 for DIODE, 640×480 for ScanNet and NYUv2, 1246×378 for KITTI, and 1920×1280 for Waymo. Metric computation is performed after mapping predictions to the common evaluation grid and valid-depth mask of each benchmark. Main-comparison methods share the same split, sparse observations, masks, and metric definitions.

For valid pixel set V, we report

$$\mathrm{AbsRel} = |V|^{-1} \sum_{p \in V} \frac{|\widehat{D}(p) - D^*(p)|}{D^*(p)}, \tag{23}$$

$$\mathrm{RMSE} = \sqrt{|V|^{-1} \sum_{p \in V} (\widehat{D}(p) - D^*(p))^2}, \tag{24}$$

$$\delta_{\mathrm{rel}<\tau} = |V|^{-1} \sum_{p \in V} \mathbf{1}\left[\frac{|\widehat{D}(p) - D^*(p)|}{D^*(p)} < \tau \right]. \tag{25}$$

We also report two threshold accuracies. The standard δ_1 counts pixels satisfying $\max(\frac{\widehat{D}}{D^*}, \frac{D^*}{\widehat{D}}) < 1.25$. For $\delta_{\mathrm{rel}<1\%}$, $\tau = 0.01$. The metric is therefore the percentage of valid pixels whose absolute relative error is below 1%. AbsRel and $\delta_{\mathrm{rel}<1\%}$ are displayed as percentages where indicated. Rankings use unrounded values.

B.5 Timing Protocol

All seven methods run on the same device with DIODE inputs. We measure model execution only, excluding data loading, preprocessing, metric computation, visualization, and logging.

B.6 Aggregation and Evidence Scope

CKA is evaluated on KITTI, and the gate analysis uses all 1000 KITTI and 654 NYUv2 samples. Paired statistics measure how consistently refinement helps across images. Transfer, mixing, fusion, and channel-width experiments examine different forms of interface reuse. Component ablations and gate records isolate design choices and conditional computation.

Metrics pool valid pixels before the final ratio or square root, and tables report per-dataset results unless a subset statistic is explicitly named. Gate records are matched by input index. Trigger rate uses every sample, and TTO gains are computed over routed samples. The offline difficulty statistics in Sec. C.11 never enter routing.

C Supplementary Results

C.1 Additional Qualitative Comparisons

Fig. 6 extends the main comparison with outdoor and indoor scenes containing thin structures, depth discontinuities, and weakly textured regions. The highlighted regions show that FreeDepth combines the dense scene layout of the active foundation with localized metric evidence: bicycle frames, poles, taps, and pipes remain spatially coherent while surrounding surfaces retain smooth depth variation.

C.2 Separating Alignment from Learned Refinement

Tabs. 11 and 12 separate the effects of sparse alignment and learned refinement. Alignment removes most of the global metric mismatch on every dataset. The corrector then further reduces AbsRel and increases high-precision coverage, adding local correction beyond scale and shift calibration.

The high-precision metric $\delta_{\mathrm{rel}<1\%}$ measures the fraction of valid pixels whose relative depth error is below 1% and directly reflects local metric precision. Tab. 12 shows that refinement raises this coverage by at least 40 percentage points on every benchmark and reaches 79.14% on NYUv2.

C.3 Image-Level Consistency of Learned Refinement

Tab. 13 pairs the sparse-aligned and final predictions for every evaluated image. The AbsRel columns summarize the per-image relative reduction by percentile, and the final column reports the paired change in high-precision coverage.

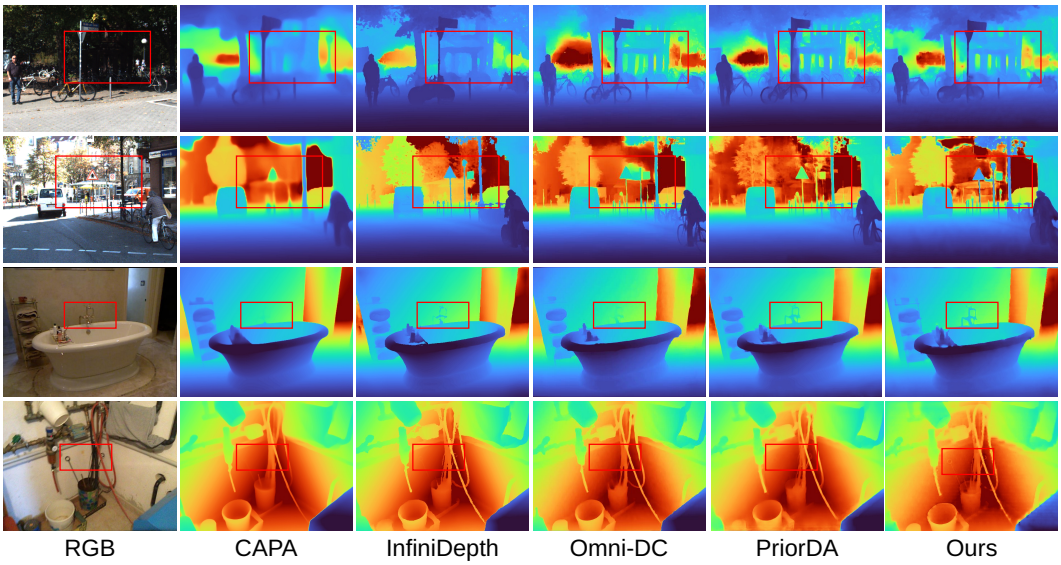

Figure 6: Additional indoor and outdoor comparisons. Red boxes mark thin structures and local depth discontinuities.

Table 11: Per-data AbsRel through raw prediction, sparse alignment, and learned refinement.

Dataset	Raw	Aligned	Refined
DIODE	64.46	6.10	2.16
ScanNet	41.52	3.94	1.30
KITTI	94.53	6.64	1.74
Waymo	96.27	10.38	2.00
NYUv2	52.52	4.24	1.03

Table 12: Per-dataset $\delta_{\mathrm{rel}<1\%}$ (%)↑ through the same three stages.

Dataset	Raw	Aligned	Refined
DIODE	0.14	28.97	77.08
ScanNet	0.57	27.29	74.88
KITTI	< 0.01	12.37	52.70
Waymo	< 0.01	8.38	52.58
NYUv2	0.08	27.08	79.14

Table 13: Paired change after refinement. AbsRel columns report per-image reduction percentiles. The final column is the median $\delta_{\mathrm{rel}<1\%}$ gain in percentage points.

		AbsRel reduction (%)↑			$\delta_{\mathrm{rel}<1\%}$ gain↑
Dataset	n	P_{10}	P_{50}	P_{90}	P_{50} pp
DIODE	325	30.75	53.96	83.21	50.56
ScanNet	800	47.27	62.92	77.17	48.88
KITTI	1000	58.17	69.61	80.87	39.81
Waymo	496	71.77	80.68	85.62	44.61
NYUv2	654	56.27	69.47	82.17	52.03

The 10th-percentile AbsRel reduction remains positive and ranges from 30.75% to 71.77%. The median high-precision gain ranges from 39.81 to 52.03 percentage points. These percentiles show consistent improvements across each benchmark, including images near the lower end of the gain distribution.

C.4 Additional Threshold Accuracy

Tab. 14 adds two threshold-based views to the mean errors. The conventional δ_1 allows a multiplicative error up to 1.25. The stricter $\delta_{\mathrm{rel}<1\%}$ requires relative error below 1% and emphasizes local metric precision. FreeDepth ranks first by mean δ_1 rank. InfiniDepth ranks first under the stricter criterion, with FreeDepth remaining competitive and leading on NYUv2.

Table 14: Threshold accuracy (↑) and mean per-metric rank (↓) under the unified protocol. Bold and underline mark the best and second-best unrounded results.

Method	DIODE		ScanNet		KITTI		Waymo		NYUv2		Mean rank↓	
	δ_1	$\delta_{<1\%}$%	δ_1	$\delta_{<1\%}$%	δ_1	$\delta_{<1\%}$%	δ_1	$\delta_{<1\%}$%	δ_1	$\delta_{<1\%}$%	δ_1	$\delta_{<1\%}$
DA3 Metric	0.9282	22.12	0.9804	26.45	0.9699	13.74	0.8964	8.10	0.9746	24.10	6.0	7.4
DA3-aligned	0.9437	28.97	0.9786	27.29	0.9656	12.37	0.9181	8.38	0.9727	27.08	6.4	6.8
Marigold-DC	0.8240	21.38	0.9766	39.37	0.9663	12.37	0.9304	11.34	0.9701	33.81	6.8	6.8
PriorDA	0.9800	75.11	0.9728	50.43	0.9514	32.53	0.9791	63.14	0.9628	50.25	6.0	4.6
InfiniDepth-metric	0.9732	**86.87**	0.9922	**83.68**	0.9906	**57.28**	0.9808	**82.13**	0.9941	78.91	3.4	**1.2**
CAPA	0.9752	79.01	0.9933	69.23	0.9931	49.27	0.9766	47.62	0.9950	64.31	3.0	4.0
OMNI-DC	0.9757	82.03	0.9900	73.98	0.9901	56.96	0.9819	79.76	0.9922	66.36	3.4	2.4
FreeDepth	**0.9822**	77.08	**0.9934**	74.88	**0.9950**	52.70	**0.9900**	52.58	**0.9964**	**79.14**	**1.0**	2.8

C.5 Additional Cross-Foundation Transfer

The main paper reports transfer results on DIODE, ScanNet, and KITTI. Table 15 further evaluates transfer on NYUv2 under both AbsRel and RMSE. A foundation-specific corrector provides a reference result for each interface.

In the held-out setting, the corrector trained with DA3 and InfiniDepth is attached directly to π^3 or VGGT. Every aligned anchor improves under both metrics. Direct reuse reaches 0.99% and 1.03% AbsRel for π^3 and VGGT, respectively, showing that the transfer behavior extends to another indoor dataset.

Table 15: NYUv2 transfer. Upper rows use foundation-specific correctors. Lower rows directly reuse the corrector.

Protocol	Foundation	NYUv2 AbsRel%↓ Align.	AbsRel%↓ Final	NYUv2 RMSE↓ Align.	RMSE↓ Final
Specific reference	DA3	4.24	1.03	0.246	0.095
	InfiniDepth	27.42	1.72	0.810	0.106
	π^3	3.18	1.26	0.212	0.114
	VGGT	3.77	1.17	0.272	0.116
Held out	π^3	3.17	0.99	0.212	0.103
	VGGT	3.77	1.03	0.272	0.102

Fig. 7 keeps each scene fixed while changing the active foundation. The base predictions differ in boundary placement, surface layout, and local detail. Applying the same sparse-guided corrector preserves these foundation-specific differences and recovers sharper metric structure around objects and depth transitions.

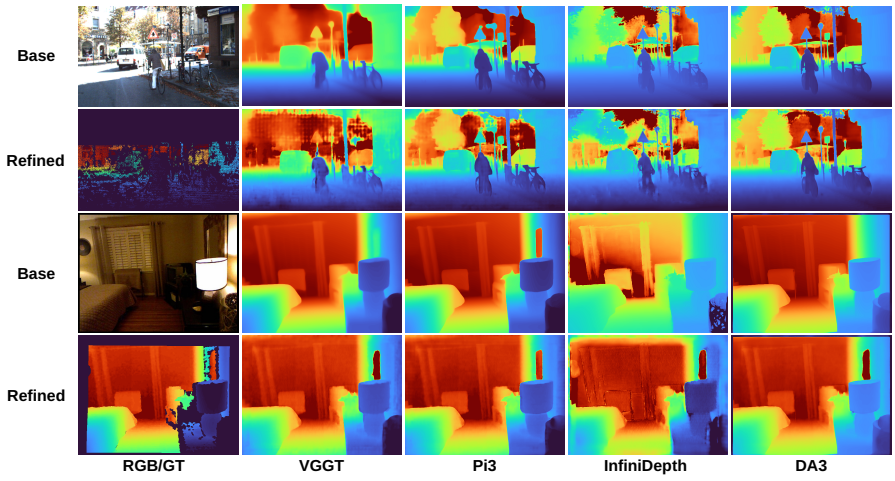

Figure 7: Qualitative reuse across foundations. Each scene shows RGB/GT followed by the base and refined outputs of VGGT, π^3, InfiniDepth, and DA3.

C.6 Output-Space Composition across Foundations

The held-out experiment evaluates each foundation separately. Tab. 16 tests direct composition by applying the same frozen corrector to DA3, InfiniDepth, π^3, and VGGT, then averaging the four refined depth maps with equal weights. The average uses no learned fusion or post-fusion sparse alignment.

Table 16: Equal-weight output composition. "Mean" averages member metrics. "Fused" evaluates the averaged refined depth map. No fusion parameters are trained.

	AbsRel%↓		RMSE↓		$\delta_{\text{rel}<1\%}$ (%)↑	
Dataset	Mean	Fused	Mean	Fused	Mean	Fused
DIODE	2.76	2.46	0.409	0.345	68.91	72.53
ScanNet	1.56	1.27	0.203	0.156	68.38	72.63
KITTI	2.50	2.05	1.886	1.410	41.60	45.05
NYUv2	1.24	1.01	0.105	0.083	71.07	75.65

Equal-weight fusion improves over the member metric mean in all 12 dataset–metric comparisons. Sparse alignment and anchor-relative correction place the different foundation outputs in a compatible metric depth space, allowing averaging to exploit their complementary errors. We use this experiment as a composition diagnostic. The main protocol remains a single active foundation and one corrector forward.

C.7 Representation Diagnostics

Linear centered kernel alignment (CKA) compares matched shallow, middle, and deep feature groups on KITTI. Higher CKA means that two feature maps preserve more similar spatial relationships after centering, even when their channel coordinates differ. The grouped results reveal recurring, depth-dependent correspondence across foundations. The following visualizations locate this correspondence in the image and trace it through the complete feature hierarchies.

Fig. 8 provides a spatial view of this correspondence. The three rows use matched shallow, middle, and deep interface stages: layers 4, 11, and 23 for DA3, VGGT, and InfiniDepth, and layers 3, 11, and 35 for π^3. We project the spatial tokens of each feature map onto three principal components and render them as PCA-RGB. PCA is fitted independently for each map, so color values are local to that map and visualize its spatial organization. Across the three scenes, the four foundations repeatedly emphasize road extent, horizon layout, object regions, and major boundaries at corresponding image locations. These shared spatial cues are the content consumed by the projected multiscale interface.

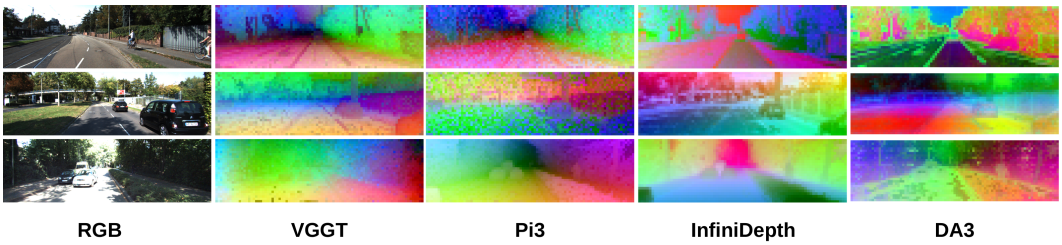

Figure 8: PCA-RGB of shallow, middle, and deep features on KITTI. Matched spatial responses recur across foundations.

The selected-stage summary gives only a partial view of correspondence along the feature hierarchy. Fig. 9 shows the mean 108×108 linear-CKA matrix on KITTI. Rows and columns contain DA3 layers 0–23, VGGT layers 0–23, π^3 layers 0–35, and InfiniDepth layers 0–23, ordered from shallow to deep within each model block. Strong near-diagonal structure appears within every foundation, and structured bands also cross model boundaries. The cross-model bands span intervals of layers, indicating correspondence across several stages.

C.8 Mixed-Feature Interfaces beyond AbsRel

The mixed-feature experiment tests the corrector with stages drawn from different foundation hierarchies. Tab. 17(a) adds RMSE and high-precision $\delta_{\mathrm{rel}<1\%}$ for these interfaces. The uppercase letter identifies the anchor foundation, and the four italic subscripts list feature sources from shallow to deep. `D`, `P`, and `V` denote DA3, π^3, and VGGT. Every combination produces a valid refinement. Four configurations achieve the lowest RMSE on at least one dataset. D_{ppdd} leads high-precision coverage on four datasets, and D_{dpdp} leads on NYUv2. The metric-dependent ordering is consistent with the AbsRel results and shows that the shared corrector can use spatial content from heterogeneous hierarchies.

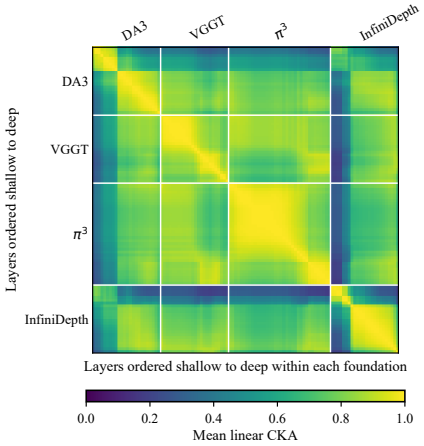

Figure 9: Mean layerwise CKA on KITTI. White lines separate foundations; layers run from shallow to deep.

C.9 Channel-Width Reuse beyond AbsRel

Tab. 17(b) gives additional metrics for attaching one corrector to DA3-S, DA3-B, DA3-L, and DA3-G. These models expose 768, 1536, 2048, and 3072 channels, respectively. The parameter-free map in Sec. A.1 presents 1024 channels to the corrector in every case. DA3-G leads on three of five datasets by RMSE and all five by high-precision coverage. DA3-L gives the lowest RMSE on ScanNet and KITTI. The fixed interface supports both channel selection and repetition with no model-specific retraining.

Table 17: Interface-reuse metrics. Bold and underline mark the best and second unrounded results.

Interface	DIODE	Scan.	KITTI	Waymo	NYU
			RMSE↓		
D_{ddpp}	**0.331**	0.088	1.394	1.773	0.093
D_{dpdp}	0.338	0.088	1.387	**1.710**	**0.093**
D_{ppdd}	0.347	0.087	**1.385**	1.719	0.093
D_{ddvv}	0.415	0.100	1.847	1.770	0.112
D_{vvdd}	0.420	0.098	1.829	1.715	0.116
V_{ppvv}	0.393	0.081	1.975	2.825	0.100
V_{vvpp}	0.377	**0.079**	1.779	2.469	0.100
			$\delta_{\mathrm{rel}<1\%}$ (%)↑		
D_{ddpp}	73.87	74.25	44.60	47.96	78.78
D_{dpdp}	74.82	74.66	45.24	52.15	**79.06**
D_{ppdd}	**76.69**	75.26	**47.86**	**52.60**	78.61
D_{ddvv}	50.57	66.09	34.31	46.75	69.39
D_{vvdd}	63.09	71.79	41.53	50.22	72.37
V_{ppvv}	63.42	73.33	35.54	23.50	75.63
V_{vvpp}	64.78	**75.83**	38.37	28.34	77.53

(a) Mixed-foundation interfaces.

Model	DIODE	Scan.	KITTI	Waymo	NYU
			RMSE↓		
DA3-S	0.396	0.100	1.665	1.862	0.104
DA3-B	0.447	0.103	1.641	1.823	0.102
DA3-L	0.376	**0.096**	**1.445**	1.765	0.100
DA3-G	**0.353**	0.097	1.524	**1.745**	**0.098**
			$\delta_{\mathrm{rel}<1\%}$ (%)↑		
DA3-S	65.98	64.05	34.54	35.41	67.57
DA3-B	67.65	67.78	40.68	43.83	71.60
DA3-L	70.65	67.84	40.59	41.96	72.57
DA3-G	**70.69**	**72.78**	**45.29**	**50.76**	**77.60**

(b) Fixed 1024-channel interface.

C.10 Additional-Dataset Component Ablation

Tab. 18(a) extends the component ablation to DIODE and NYUv2. The full model performs best on both datasets and improves clearly over the aligned anchor, confirming the contribution of learned local correction. Changing the prompt coordinates, injection strategy, or output action consistently lowers accuracy, matching the pattern observed in the main-paper ablation.

C.11 Difficulty Diagnostics of the Optional TTO Gate

We measure pre-TTO difficulty offline using the within-dataset percentile of feed-forward AbsRel and report the diagnostics in Tab. 18(b). AUC describes how well the continuous disagreement score separates the worst 20% from the remaining samples, with 0.5 corresponding to chance. For the actual threshold $\eta = 0.5$, we also report trigger lift: the trigger rate within the worst 20% divided by the overall trigger rate. A value of one means that difficult samples are no more likely to be selected than average. The AUC values are 0.875 and 0.790, and the trigger lifts are $2.94\times$ and $2.13\times$ on KITTI and NYUv2. Thus, difficult samples are more than twice as likely to be routed, and ordinary inputs continue through the feed-forward path.

C.12 Optimization Dynamics on Routed Samples

Tab. 19 summarizes the optional update. Edge overlap measures Jaccard agreement between prominent log-depth gradients in the prediction and anchor, with lower values indicating greater structural disagreement. Routed samples have lower median overlap than bypassed samples. KITTI more often reaches the 15-step limit. Most NYUv2 updates stop after the objective saturates.

Table 18: Component and gate diagnostics. Abs. denotes AbsRel. Ground truth is used only for the offline difficulty labels. Gate decisions use prediction–anchor agreement.

Variant	DIODE		NYUv2	
	Abs.%↓	RMSE↓	Abs.%↓	RMSE↓
Full FreeDepth	2.16	0.329	1.03	0.095
Anchor only	6.10	0.545	4.24	0.246
Direct log prompt	3.88	0.407	1.75	0.115
Linear residual	2.37	0.332	1.57	0.109
Direct depth prompt	6.49	0.690	3.71	0.192
Output-only FiLM	4.23	0.428	3.03	0.175
Additive injection	2.65	0.367	1.71	0.128
Concat fusion	2.42	0.352	1.57	0.110
Direct-depth output	3.60	0.391	7.57	0.306

(a) Component ablations.

Statistic	KITTI	NYUv2
Trigger rate (%)	17.10	27.98
Worst-20% difficulty AUC	0.875	0.790
Worst-20% trigger lift	$2.94\times$	$2.13\times$

(b) Gate difficulty diagnostics.

Table 19: Rank-8 gated-TTO dynamics at 5×10^{-4}. Statistics use routed samples and the full three-term log-depth objective.

Statistic	KITTI	NYUv2
Median overlap, routed	0.475	0.443
Median overlap, bypassed	0.584	0.577
Median TTO steps	15	8
Objective-saturation stop (%)	43.86	93.99
Maximum-budget stop (%)	56.14	6.01
Median objective reduction (%)	31.45	19.45

The median routed objective falls by 31.45% on KITTI and 19.45% on NYUv2, showing that the optional LoRA state can fit the additional sparse evidence for selected samples. All feed-forward results are obtained with this update disabled. The different stopping profiles mainly explain the routed latency gap in Tab. 20(b).

C.13 Runtime and Conditional-Compute Cost

Under the shared timing protocol, FreeDepth runs at 398.71 ms per image, close to OMNI-DC at 377.84 ms, as shown in Tab. 20(a). It is $2.38\times$, $25.29\times$, and $34.10\times$ faster than PriorDA, CAPA with VGGT, and Marigold-DC, respectively. DA3 Metric and InfiniDepth are faster.

Table 20: Runtime on common inputs and hardware. Panel (a) excludes preprocessing and evaluation. Panel (b) averages TTO cost over routed and bypassed samples.

Method	Runtime (ms)	FPS
DA3 Metric	68.83	14.528
InfiniDepth	179.15	5.582
OMNI-DC	377.84	2.647
FreeDepth	398.71	2.508
PriorDA	948.61	1.054
CAPA with VGGT	10081.83	0.099
Marigold-DC	13594.71	0.074

(a) Cross-method runtime on DIODE.

Statistic	KITTI	NYUv2
Measured samples	997	651
Route rate (%)	17.15	27.65
Feed-forward path (ms)	244.49	168.58
Bypass overhead (ms)	6.04	5.89
Routed overhead (ms)	10243.08	4073.72
Amortized TTO overhead (ms)	1761.84	1130.63
Amortized total (ms)	2006.34	1299.22
Throughput: FF; TTO (s^{-1})	4.090; 0.498	5.932; 0.770
Peak memory: FF; TTO (GiB)	4.51; 10.73	3.66; 7.97

(b) Feed-forward and gated-TTO cost.

Tab. 20(b) measures the optional gated path. Bypassed samples add 6.04 ms on KITTI and 5.89 ms on NYUv2. Routed samples execute multiple LoRA steps and average 10.24 s and 4.07 s. Averaged over every input, the TTO overhead is 1761.84 and 1130.63 ms, giving total latencies of 2006.34 and 1299.22 ms per image. TTO provides an optional accuracy–compute trade-off. The standard FreeDepth result remains single-pass.