FROM RELATIVE TO METRIC: GROUNDING DIFFUSION DEPTH WITH SCENE-STRUCTURED LANGUAGE

Anonymous authors
Paper under double-blind review

ABSTRACT

Monocular relative depth models generalize well across diverse scenes, but their predictions are defined only up to an unknown metric gauge. Recovering physical depth therefore requires not only reliable scene geometry, but also a mechanism for grounding that geometry in metric scale. We investigate whether the semantic priors already encoded in diffusion-based depth models can support such grounding without camera metadata, external geometric measurements, or test-time ground-truth alignment. Our key observation is that scale-relevant semantic cues are not geometrically interchangeable: foreground objects, midground structures, and background perspective provide complementary evidence that plays different spatial roles in metric calibration. Based on this insight, we propose a hierarchical scene-language-guided calibration framework for frozen diffusion-depth models. The framework reuses both the relative disparity and intermediate spatial features of the frozen backbone, and introduces depth-structured descriptions for foreground, midground, and background regions to guide spatially varying scale–shift prediction through shared cross-attention. We further propose Scene-Scale Anchor Alignment (SSAA), which complements local calibration by organizing the learned representation according to scene-level metric scale. Experiments with Marigold, Lotus, and E2E-FT across multiple metric-depth benchmarks show consistent improvements and favorable performance against existing relative-to-metric methods with a lightweight trainable module.

1 INTRODUCTION

A single image often constrains relative scene geometry much more strongly than absolute physical scale. Monocular metric depth estimation must therefore solve two coupled problems: recovering scene structure and assigning that structure a plausible metric gauge. Models trained with metric supervision can learn such a gauge, but their predictions may become entangled with the camera configurations, scene statistics, and depth ranges represented in the training data. Relative-depth models instead learn scale- and shift-invariant geometry from heterogeneous datasets and consequently transfer more robustly across domains Ranftl et al. (2020), at the cost of discarding the metric information required to recover physical distances. This complementarity motivates a different route to metric depth: preserve transferable relative geometry and learn to *metrically ground* it using image-derived semantic priors, without camera metadata, external geometric measurements, or test-time ground-truth alignment.

Diffusion-based depth models provide a compelling substrate for this problem. Methods such as Marigold Ke et al. (2024), GeoWizard Fu et al. (2024), DepthFM Gui et al. (2025), and Lotus He et al. (2025) adapt pretrained text-to-image diffusion models to depth estimation, inheriting rich semantic priors while learning detailed spatial geometry. Their strong transferability, however, is commonly evaluated under affine-invariant protocols, where each prediction is optimally aligned to the ground truth by an image-specific scale and shift Ranftl et al. (2020). Such evaluation measures relative geometric accuracy while explicitly removing the metric gauge. Iris Zeng et al. (2026) further shows that image-specific language can improve the affine-invariant predictions of Marigold, Lotus, and E2E-FT Garcia et al. (2025), demonstrating that semantic information complements diffusion-based geometry. Yet whether the joint geometric and semantic representations of a frozen diffusion-depth model can be directly exploited for *metric grounding* remains largely unexplored.

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.

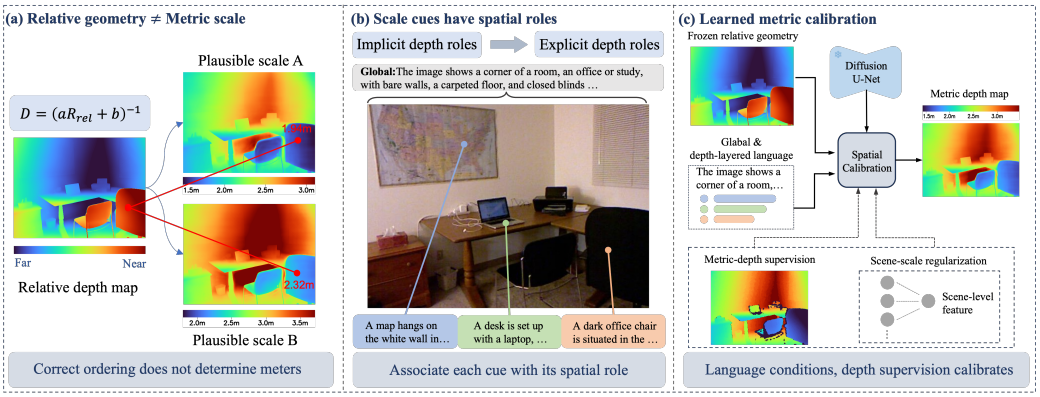


Figure 1: **Motivation and overview.** (a) Relative geometry alone does not determine the metric gauge: multiple physical scales can be consistent with the same relative prediction. (b) Semantic scale cues play different geometric roles. Foreground objects, midground structures, and background perspective provide complementary evidence that is entangled in a single global caption. (c) We retain a frozen diffusion-depth backbone and learn a lightweight calibration branch that combines its relative geometry with depth-structured language and scene-scale anchor alignment (SSAA) to predict metric depth.

Language provides a natural source of metric priors because scene categories, object identities and canonical sizes, spatial relations, and perspective layouts constrain the range of physically plausible scene scales. Existing methods exploit such priors in different ways. WorDepth Zeng et al. (2024a) models image captions as distributions over plausible metric scenes, RSA Zeng et al. (2024b) predicts global inverse-depth calibration from language, and TR2M Cui et al. (2026) combines visual and textual features to estimate spatially varying calibration. CLIP-based approaches likewise associate semantic representations with distance prompts, depth intervals, or image-adaptive depth bins Zhang et al. (2022); Auty & Mikolajczyk (2023); Son & Lee (2024); Kim et al. (2025). These studies establish that vision-language representations contain useful metric priors. The remaining question, however, is not simply whether language contains scale information, but *how scale-relevant semantics should be organized with respect to scene geometry*.

Our key observation is that semantic scale cues are not geometrically interchangeable. The identity and apparent extent of a foreground object, the arrangement of midground structures, and the perspective layout of a distant background provide qualitatively different evidence for metric calibration. A single image-level caption may contain all of these cues, but leaves their depth-dependent roles implicit. A spatial calibration model must therefore simultaneously discover *what* semantic evidence is relevant and *where* that evidence should affect the prediction. We instead make this coarse geometric organization explicit by structuring language according to scene depth. Importantly, this is not intended to provide numerical depth through text; rather, it exposes the distinct semantic roles of near-, intermediate-, and far-range scene content so that they can interact with an already geometry-aware representation.

To this end, we propose a *hierarchical scene-language-guided metric calibration framework* for frozen diffusion-depth models (Fig. 1). Each image is associated with one global description and three local descriptions corresponding to foreground, midground, and background content. These descriptions contain semantic entities and spatial relations but no metric depths or camera parameters. The global description conditions the frozen diffusion backbone through its native text-conditioning pathway. The local descriptions are encoded by the frozen text encoder, augmented with learnable depth-layer identities, and fused with an intermediate spatial feature of the diffusion-depth model through a shared cross-attention module. The diffusion backbone thus serves two roles without being updated: its relative disparity provides an explicit geometric scaffold, while its intermediate feature provides spatial queries through which depth-structured semantic evidence can affect calibration. A lightweight decoder combines these representations to predict spatially varying scale and shift maps in the inverse-depth domain, from which metric depth is obtained. Unlike approaches that introduce a separate calibration-specific image encoder, our formulation directly reuses the geometric representation already learned by the frozen depth model.

Spatially resolved calibration addresses local metric corrections, but pixel-wise depth supervision alone does not explicitly organize the shared representation according to the overall metric regime of a scene. We therefore introduce *Scene-Scale Anchor Alignment* (SSAA) as complementary scene-level supervision. SSAA compares a globally pooled calibration feature with a small set of learnable feature anchors and supervises their assignments using soft targets derived from the log-median metric depth of each training image. The soft assignment preserves the continuous nature of scene scale: scenes lying between representative depth ranges can activate multiple anchors rather than being forced into a discrete scale category. SSAA therefore encourages the calibration representation to retain coarse scene-level metric information while the depth-structured language supports spatially varying semantic calibration. Ground-truth scene-scale statistics are required only for training; inference uses neither ground-truth depth for anchor assignment nor per-image fitted calibration parameters.

We evaluate the proposed framework on four frozen diffusion-depth variants—Marigold, E2E-FT Garcia et al. (2025), Lotus-G, and Lotus-D—that share a Stable Diffusion v2 foundation but differ in their depth adaptation and prediction formulations. Only a 23.4M-parameter calibration branch is trained for each variant. With Lotus-D, our method reduces AbsRel over TR2M by 6.1% on NYUv2 and 12.1% on KITTI. We further train a single Lotus-D-based model jointly on NYUv2, KITTI, and VOID and directly evaluate it on eight unseen datasets spanning indoor, outdoor, real, and synthetic scenes, without target-domain adaptation or test-time ground-truth alignment; the model achieves an average rank of 2.00 among the five compared methods. Controlled ablations further reveal complementary effects of the proposed components: depth-structured local descriptions improve metric reconstruction, while SSAA reduces zero-shot average AbsRel by 3.0% and simultaneously improves RMSE and threshold accuracy. These results indicate that spatially organized semantic priors and scene-level metric supervision provide complementary signals for grounding transferable diffusion-based geometry in metric space.

Our contributions are summarized as follows:

- We introduce a relative-to-metric calibration framework that reuses both the relative disparity and spatial geometry representation of a frozen diffusion-depth model, enabling metric prediction without an additional calibration-specific image encoder or test-time ground-truth alignment.
- We develop a hierarchical scene-language representation and depth-layered fusion mechanism that explicitly organize semantic scale cues according to their coarse geometric roles, allowing spatial diffusion features to selectively query foreground, midground, and background information for metric calibration.
- We propose SSAA, which complements spatial calibration with soft scene-level supervision and organizes the learned calibration representation according to continuous scene-scale statistics.

2 METHODOLOGY

Overview We introduce a hierarchical scene-language guided calibration framework that uses the relative geometry of a frozen diffusion depth model as an explicit reference for metric depth prediction with a lightweight trainable decoder. Given an RGB image $I_i \in \mathbb{R}^{3\times H\times W}$, we construct

$$\mathcal{T}_i = \left(T_i^g, T_i^f, T_i^m, T_i^b\right), \tag{1}$$

where T_i^g describes the complete scene and T_i^f, T_i^m, and T_i^b describe the foreground, midground, and background, respectively. The four descriptions are constructed by a frozen vision-language model under a fixed structured schema, as detailed in supplementary materials. The global description conditions the frozen backbone to produce relative disparity R_i and an intermediate geometry feature F_i^{16}; the local descriptions condition the calibration decoder, which predicts pixel-wise scale and shift maps A_i and B_i. Metric depth is obtained in the inverse-depth domain:

$$D_{m,i} = \frac{1}{A_i \odot R_i + B_i}. \tag{2}$$

Thus, global language supplies scene-level context, while local language guides region-dependent metric calibration. The frozen backbone is not modified, and the relative disparity is retained as an explicit geometric reference rather than as a guarantee of preserved depth structure.

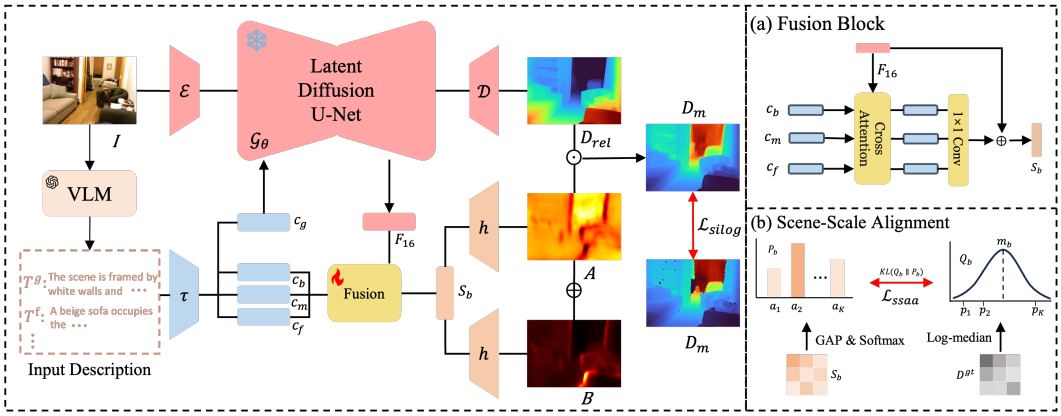


Figure 2: Overview of the proposed hierarchical scene-language guided metric depth estimation framework. Given an RGB image, a frozen vision-language model constructs one global description and three depth-aware local descriptions associated with the foreground, midground, and background. The global description conditions a frozen diffusion depth backbone, which produces an affine-invariant relative disparity map and an intermediate geometry feature. A lightweight calibration decoder injects the three local descriptions through depth-layered cross-attention and predicts spatially varying scale and shift maps. Scene-Scale Anchor Alignment further regularizes the shared calibration feature using image-level metric-scale statistics. Only the calibration decoder, text projection layers, scale–shift heads, and scene-scale anchors are optimized.

Spatial Metric Calibration During the frozen diffusion backbone's forward pass that predicts the relative disparity map R_i, we simultaneously extract the activation from its second U-Net up-sampling block:

$$F_i^{16} = \mathrm{Hook}_{16}(\mathcal{G}_{\bar{\theta}}) \in \mathbb{R}^{C_{16}\times H/16\times W/16}, \tag{3}$$

where $C_{16} = 1280$ for the adopted SD-2 U-Net. Both R_i and F_i^{16} are detached before entering the calibration branch. The decoder combines this feature with the three local descriptions:

$$S_i = f_{\mathrm{dec}}\left(F_i^{16}, T_i^f, T_i^m, T_i^b\right) \in \mathbb{R}^{256\times H/16\times W/16}, \tag{4}$$

where S_i is the shared calibration feature. Two independent heads upsample it to the input resolution and produce $A_i = h_{\mathrm{scale}}(S_i)$ and $B_i = h_{\mathrm{shift}}(S_i)$, both in $\mathbb{R}^{1\times H\times W}$. The calibrated inverse depth is

$$\widehat{R}_i = A_i \odot R_i + B_i, \tag{5}$$

and $D_{m,i} = 1/\widehat{R}_i$. Because A_i and B_i vary spatially, the model can correct region-dependent deviations while retaining R_i as its geometric reference.

Depth-Layered Text Fusion Let $X_i \in \mathbb{R}^{d\times h\times w}$ denote the projected decoder feature. Each local summary is encoded by the frozen text encoder, augmented with a learnable depth-layer embedding, and projected to the decoder width:

$$U_i^r = P\left[\tau_{\bar{\theta}}(T_i^r) + \mathbf{E}_{\mathrm{layer}}^r\right] \in \mathbb{R}^{L\times d}, \qquad r \in \{f, m, b\}, \tag{6}$$

where $\mathbf{E}_{\mathrm{layer}}^r \in \mathbb{R}^{L\times d_t}$ is a learnable embedding associated with depth layer r and P is a learned projection. We employ a shared cross-attention mechanism to fuse the spatial visual tokens with the three local text representations. The visual tokens serve as queries, while the local text representations provide keys and values, producing the depth-layer responses Y_i^f, Y_i^m, and Y_i^b. These responses are then reshaped, concatenated, and fused through a residual projection:

$$\widetilde{X}_i = X_i + W_c\left[\mathrm{Reshape}(Y_i^f) \,\|\, \mathrm{Reshape}(Y_i^m) \,\|\, \mathrm{Reshape}(Y_i^b)\right]. \tag{7}$$

The attention parameters are shared across the three layers, and the 1×1 convolution W_c is zero-initialized. The module therefore begins as a visual-only decoder and learns spatially adaptive language modulation without explicit region masks.

Scene-Scale Anchor Alignment Pixel-wise depth supervision does not explicitly organize the shared representation by image-level scene scale. We therefore apply Scene-Scale Anchor Alignment (SSAA) to $v_i = \mathrm{GAP}(S_i) \in \mathbb{R}^{256}$. Given K_a learnable feature anchors $\{a_k\}_{k=1}^{K_a}$, the predicted assignment is

$$P_{i,k} = \frac{\exp(\cos(v_i, a_k)/\tau)}{\sum_{j=1}^{K_a} \exp(\cos(v_i, a_j)/\tau)}, \tag{8}$$

where τ is a temperature parameter. From valid ground-truth pixels M_i, we compute $m_i = \log[\mathrm{median}_{p \in M_i} D_i^{\mathrm{gt}}(p)]$. Fixed centers $\{p_k\}_{k=1}^{K_a}$ estimated from the training-set distribution define the soft target

$$Q_{i,k} = \frac{\exp\left[-(m_i - p_k)^2/(2\sigma^2)\right]}{\sum_{j=1}^{K_a} \exp[-(m_i - p_j)^2/(2\sigma^2)]}, \tag{9}$$

where σ controls the overlap between adjacent scale regimes. The alignment loss is

$$\mathcal{L}_{\mathrm{ssaa}} = \frac{1}{|\mathcal{B}_v|} \sum_{i \in \mathcal{B}_v} \sum_{k=1}^{K_a} Q_{i,k} \left(\log Q_{i,k} - \log P_{i,k}\right), \tag{10}$$

where $\mathcal{B}_v$ contains images with sufficiently many valid pixels. The soft targets preserve continuity between neighboring scale ranges, so SSAA regularizes the shared feature without treating scene scale as a hard category.

Training Objectives For each valid pixel $p \in M_i$, let $e_i(p) = \log[\max(D_{m,i}(p), d_{\min})] - \log[\max(D_i^{\mathrm{gt}}(p), d_{\min})]$. We use the scale-invariant logarithmic loss Eigen et al. (2014):

$$\mathcal{L}_{\mathrm{silog}} = \frac{\alpha}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \sqrt{\frac{1}{|M_i|} \sum_{p \in M_i} e_i(p)^2 - \lambda \left(\frac{1}{|M_i|} \sum_{p \in M_i} e_i(p) \right)^2}, \tag{11}$$

where $\alpha = 10$ and $\lambda = 0.85$. We also use an edge-aware smoothness loss $\mathcal{L}_{\mathrm{smooth}}$ on A_i and B_i, and the pseudo scale–shift loss $\mathcal{L}_{\mathrm{pseudo}}$ of Cui et al. (2026). The complete objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{silog}} + \lambda_1 \mathcal{L}_{\mathrm{smooth}} + \lambda_2 \mathcal{L}_{\mathrm{pseudo}} + \lambda_3 \mathcal{L}_{\mathrm{ssaa}}, \tag{12}$$

where λ_1, λ_2, and λ_3 balance the three auxiliary objectives. The diffusion backbone, VAE, backbone text encoder, and vision-language model remain frozen; only the calibration branch and scene-scale anchors are optimized. Detailed settings are provided in the supplementary material.

3 Experiments

3.1 Experimental Protocol

Training and evaluation data. We jointly train on NYUv2 Silberman et al. (2012), KITTI Geiger et al. (2012), and VOID Wong et al. (2020), using valid depth ranges of $[10^{-3}, 10]$, $[10^{-3}, 80]$, and $[0.2, 5]$ meters, respectively. KITTI follows the Eigen split and benchmark crop. Each image is paired with one global and three local descriptions corresponding to foreground, midground, and background. The descriptions contain no numerical depth, camera intrinsics, or dataset-specific scale information; depth maps provide the only numerical scale supervision.

We evaluate held-out NYUv2 and KITTI for seen-domain calibration and eight unseen test sets for zero-shot transfer: iBims-1 Koch et al. (2018), SUN RGB-D Song et al. (2015), DIODE Indoor/Outdoor Vasiljevic et al. (2019), HyperSim Roberts et al. (2021), DDAD Guizilini et al. (2020), DIML Outdoor Kim et al. (2018), and Virtual KITTI 2 Cabon et al. (2020). Unless explicitly stated otherwise, predictions are evaluated directly in metric space without median scaling, ground-truth fitting, or target-domain adaptation. Following RSA Zeng et al. (2024b) and TR2M Cui et al. (2026), we report standard depth metrics; the zero-shot tables use AbsRel and δ_1.

Implementation. We instantiate the framework on four frozen diffusion-depth predictors: Marigold Ke et al. (2024), E2E-FT Garcia et al. (2025), Lotus-G, and Lotus-D He et al. (2025). Only the projections, depth-layer embeddings, cross-attention module, calibration decoder, scale–shift

Table 1: Quantitative results on NYUv2. I, L, R, and M denote image, language, relative depth, and metric depth, respectively. Domain Adap denotes domain adaptation. Median denotes per-image median scaling using ground truth. Linear Fit estimates an image-level scale and shift using test-set ground truth, while Global uses one scale and shift for the entire dataset. * denotes results reproduced in our environment. Among the indirect metric-depth methods that do not use test-time ground-truth alignment, the best, second-best, and third-best distinct results are highlighted by dark, medium, and light orange, respectively.

Model	Type	Scaling	Train Params	$\delta_1\uparrow$	$\delta_2\uparrow$	$\delta_3\uparrow$	AbsRel$\downarrow$	$\log_{10}\downarrow$	RMSE$\downarrow$
Direct metric depth estimation methods									
DA2	I$\to$M	Domain Adap	25M	0.961	0.996	0.999	0.073	0.032	0.261
ZoeDepth	I$\to$M	Domain Adap	345M	0.951	0.994	0.999	0.077	0.033	0.282
Metric3Dv2*	I$\to$M	Zero-shot	1011M	0.980	0.997	0.999	0.067	0.030	0.261
UniDepth*	I$\to$M	Zero-shot	347M	0.981	0.998	1.000	0.072	0.030	0.229
Indirect metric depth estimation methods									
DA	R$\to$M	Linear Fit	25M	0.969	0.995	0.999	0.055	0.024	0.260
Marigold	R$\to$M	Linear Fit	865M	0.962	0.990	0.998	0.057	0.024	0.226
Lotus-D	R$\to$M	Linear Fit	865M	0.965	0.991	0.998	0.057	0.024	0.221
Lotus-D	R$\to$M	Global	865M	0.491	0.804	0.940	0.301	0.118	0.865
Lotus-D	R$\to$M	Median	865M	0.446	0.675	0.791	0.349	0.201	1.111
DepthCLIP	I,L$\to$M	Zero-shot	-	0.394	0.683	0.851	0.388	0.156	1.167
DepthLM	I,L$\to$M	Zero-shot	-	0.824	0.933	0.990	0.134	0.068	0.401
SPACE-CLIP	I$\to$M	Fine-tuning	-	0.896	0.984	0.997	0.104	0.045	0.385
ScaleDepth*	I,L$\to$M	Domain Adap	109M	0.940	0.993	**0.999**	0.083	0.036	0.293
SharpDepth*	I,M,R$\to$M	Zero-shot	865M	**0.956**	0.989	0.996	**0.075**	**0.032**	0.286
RSA	I,L,R$\to$M	Rescale factors	4.7M	0.752	0.964	0.992	0.156	0.071	0.528
Zhan et al. (2026)	I,L,R$\to$M	Rescale factors	-	0.782	0.976	0.996	0.145	0.064	0.489
TR2M	I,L,R$\to$M	Rescale maps	19M	0.954	**0.996**	**0.999**	0.082	0.035	0.293
Marigold+Ours	I,L,R$\to$M	Rescale maps	23.4M	0.938	0.991	0.998	0.086	0.037	0.314
Lotus-G+Ours	I,L,R$\to$M	Rescale maps	23.4M	0.945	0.993	**0.999**	0.082	0.035	0.298
E2E-FT+Ours	I,L,R$\to$M	Rescale maps	23.4M	0.954	0.995	**0.999**	0.076	0.033	0.283
Lotus-D+Ours	I,L,R$\to$M	Rescale maps	23.4M	**0.956**	0.995	**0.999**	0.077	0.033	**0.278**

heads, and three SSAA anchors are optimized, totaling 23.4M trainable parameters. Each model is trained for 40,000 steps with a global batch size of 16. The learning rate decreases linearly from 3×10^{-5} to 1×10^{-5} over the first 20,000 steps and remains fixed at 1×10^{-5} for the remaining 20,000 steps. We set $\lambda_1 = 0.01$, $\lambda_2 = 0.5$, and $\lambda_3 = 0.1$ for edge-aware smoothness, pseudo scale–shift supervision, and SSAA, respectively. Local text features have width $d = 512$, and horizontal flipping is disabled to preserve spatial references. Unless otherwise stated, zero-shot and ablation experiments use Lotus-D.

3.2 Metric Calibration on Seen Domains

Comparison protocol. Tables 1 and 2 compare direct metric-depth methods with relative-to-metric approaches. *Linear Fit* and *Median* use test-image ground truth and are included only as oracle alignment references; *Global* applies one scale–shift pair across an evaluation set. Since methods differ in backbones and training data, these tables provide system-level comparisons, while controlled component effects are analyzed below.

Metric grounding without test-time alignment. On NYUv2, Lotus-D + Ours achieves an AbsRel of 0.077 and RMSE of 0.278, reducing the corresponding TR2M errors by 6.1% and 5.1%. E2E-FT + Ours further reaches an AbsRel of 0.076. On KITTI, Lotus-D + Ours lowers AbsRel from 0.066 to 0.058 (12.1%) and $\mathrm{RMSE}_{\log}$ from 0.093 to 0.088. These results show that the frozen relative prediction can provide a stable geometric scaffold for learned metric calibration.

Compatibility across frozen backbones. The same calibration architecture applies to Marigold, Lotus-G, E2E-FT, and Lotus-D without updating the diffusion parameters. Across the four variants, AbsRel ranges from 0.076–0.086 on NYUv2 and 0.058–0.067 on KITTI. Thus, the proposed calibration is not tied to a particular diffusion-depth adaptation. The 23.4M trainable parameters account for only about 2.7% of an 865M-parameter diffusion backbone.

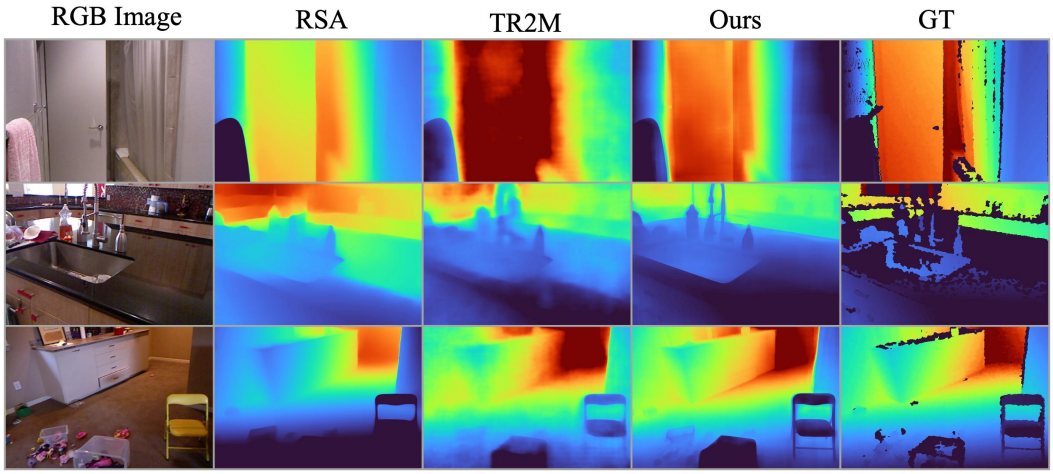


Figure 3: Qualitative results on NYUv2. More results are in the Supplementary Material.

Table 2: Quantitative results on KITTI. Detailed notations are given in Table 1.

Model	Type	Scaling	Train Params	$\delta_1\uparrow$	$\delta_2\uparrow$	$\delta_3\uparrow$	AbsRel↓	$\text{RMSE}_{\log}\downarrow$	RMSE↓
Direct metric depth estimation methods									
DA2	I→M	Domain Adap	25M	0.973	0.997	0.999	0.053	0.081	2.235
ZoeDepth	I→M	Domain Adap	345M	0.971	0.996	0.999	0.054	0.082	2.281
Metric3Dv2*	I→M	Zero-shot	1011M	0.977	0.996	0.999	0.051	0.080	2.401
UniDepth*	I→M	Zero-shot	347M	0.986	0.998	0.999	0.043	0.067	1.810
Indirect metric depth estimation methods									
DA	R→M	Linear Fit	25M	0.944	0.991	0.998	0.077	0.122	3.190
Marigold	R→M	Linear Fit	865M	0.936	0.991	0.998	0.088	0.121	2.642
Lotus-D	R→M	Linear Fit	865M	0.946	0.993	0.998	0.087	0.116	2.513
Lotus-D	R→M	Global	865M	0.731	0.935	0.979	0.198	0.205	4.120
Lotus-D	R→M	Median	865M	0.434	0.718	0.868	0.324	0.456	8.314
DepthCLIP	I,L→M	Zero-shot	-	0.465	0.713	0.867	0.343	0.305	7.583
DepthLM	I,L→M	Zero-shot	-	0.765	0.912	0.988	0.156	0.193	4.356
SPACE-CLIP	I→M	Fine-tuning	-	0.909	0.981	0.995	0.090	0.153	3.845
ScaleDepth*	I,L→M	Domain Adap	109M	**0.978**	**0.998**	**0.999**	**0.051**	**0.076**	**2.006**
SharpDepth*	I,M,R→M	Zero-shot	865M	0.965	0.996	**0.999**	0.068	0.096	2.596
RSA	I,L,R→M	Rescale factors	4.7M	0.786	0.967	0.995	0.147	0.179	4.143
Zhan et al. (2026)	I,L,R→M	Rescale factors	-	0.763	0.959	0.988	0.154	0.188	4.380
TR2M	I,L,R→M	Rescale maps	19M	0.965	0.996	**0.999**	0.066	0.093	2.328
Marigold+Ours	I,L,R→M	Rescale maps	23.4M	0.946	0.990	0.996	0.067	0.104	2.766
Lotus-G+Ours	I,L,R→M	Rescale maps	23.4M	0.956	0.994	**0.999**	0.064	0.097	2.560
E2E-FT+Ours	I,L,R→M	Rescale maps	23.4M	0.961	0.995	**0.999**	0.060	0.093	2.523
Lotus-D+Ours	I,L,R→M	Rescale maps	23.4M	0.964	0.996	**0.999**	0.058	0.088	2.387

Qualitative results. Figure 3 shows that our predictions retain coherent relative structure while reducing region-dependent metric distortion compared with RSA and TR2M, consistent with the intended role of spatially varying calibration.

3.3 Zero-Shot Cross-Dataset Generalization

We evaluate one jointly trained model on all eight unseen test sets without target-domain supervision, adaptation, or ground-truth alignment. Depth Anything V1 and ZoeDepth use their NYUv2-trained variants for indoor datasets and KITTI-trained variants for outdoor datasets, whereas RSA, TR2M, and our method use a single model across all targets. Since the systems use different backbones, this experiment evaluates complete-system transfer. Across the 16 dataset–metric pairs in Tables 3 and 4, our method ranks first on eight and second on four, giving the best overall average rank of 2.00. For indoor scenes, it obtains an average rank of 1.63, with the best results on both metrics for iBims-1 and HyperSim and the best AbsRel on SUN RGB-D. For outdoor scenes, it ties Depth Anything V1 for the best average rank (2.38). On DIML Outdoor, AbsRel decreases from TR2M's 0.5791 to 0.3008 (48.1%), while δ_1 increases from 0.2453 to 0.4759; our method also achieves the lowest AbsRel on

Table 3: Zero-shot metric depth evaluation on indoor datasets. We highlight **Best** and second-best results. Avg. Rank reports the average rank over the eight metrics.

Method	Backbone	iBims-1		SUN RGB-D		DIODE Indoor		HyperSim		Avg. Rank
		AbsRel ↓	δ_1 ↑	AbsRel ↓	δ_1 ↑	AbsRel ↓	δ_1 ↑	AbsRel ↓	δ_1 ↑	
DA	ViT-Large	0.1548	0.6936	0.3615	**0.6796**	0.3268	0.2998	0.3968	0.3220	2.50
ZoeDepth	BeiT_{384}-Large	0.1732	0.6412	0.3901	0.5656	0.3278	**0.3786**	0.4228	0.2939	3.50
RSA	ViT-Small	0.2266	0.5362	0.5444	0.3917	0.3541	0.3593	0.4673	0.2313	4.63
TR2M	ViT-Small	0.1562	0.7262	0.3718	0.6011	**0.2955**	0.3589	0.4118	0.2860	2.75
Lotus-D+Ours	SD U-Net	**0.1503**	**0.7581**	**0.3428**	0.6685	0.3204	0.3237	**0.3776**	**0.3804**	**1.63**

Table 4: Zero-shot metric depth evaluation on outdoor datasets.

Method	Backbone	DDAD		DIML Outdoor		DIODE Outdoor		Virtual KITTI 2		Avg. Rank
		AbsRel ↓	δ_1 ↑	AbsRel ↓	δ_1 ↑	AbsRel ↓	δ_1 ↑	AbsRel ↓	δ_1 ↑	
DA	ViT-Large	**0.1053**	**0.8866**	1.6105	0.0028	0.7993	**0.2886**	0.0922	**0.9067**	2.38
ZoeDepth	BeiT_{384}-Large	0.1281	0.8390	1.9143	0.0031	0.8551	0.2681	0.1075	0.8546	3.88
RSA	ViT-Small	0.2269	0.4983	1.2233	0.0417	0.6887	0.2809	0.1820	0.6974	3.88
TR2M	ViT-Small	0.1128	0.8686	0.5791	0.2453	**0.5046**	0.2783	0.1229	0.8461	2.50
Lotus-D+Ours	SD U-Net	0.1379	0.8398	**0.3008**	**0.4759**	0.5854	0.1898	**0.0883**	0.8981	**2.38**

Virtual KITTI 2. The gains are not uniform: Depth Anything V1 and TR2M remain stronger on DDAD, and our δ_1 on DIODE Outdoor is lower. Nevertheless, the aggregate results indicate that the learned metric calibration transfers beyond the NYUv2–KITTI–VOID training distributions.

3.4 Ablation and Diagnostic Analysis

Unless otherwise stated, all ablations use Lotus-D and the same NYUv2–KITTI–VOID training protocol. "Zero-shot Avg." denotes the macro-average over the eight unseen test sets. More ablation experiments in the Supplementary Material.

Text description compatibility Table 5 compares no text, Iris captions Zeng et al. (2026), and our global description T^g. Across Lotus-D and Marigold on NYUv2 and KITTI, introducing T^g changes AbsRel by at most 0.0012. These results indicate that global text has a limited effect on affine-aligned relative-depth accuracy under the evaluated settings.

Table 5: Ablation study on the effect of text on diffusion-based depth estimation. All numbers use per-image ground-truth affine alignment.

		NYUv2		KITTI	
Backbone	Inference text	δ_1 ↑	AbsRel ↓	δ_1 ↑	AbsRel ↓
Lotus-D	none	**0.9664**	0.0537	0.9499	0.0844
Lotus-D	Iris caption	0.9662	**0.0536**	0.9503	0.0835
Lotus-D	ours: T^g	0.9658	0.0546	**0.9504**	**0.0832**
Marigold	none	0.9581	**0.0609**	**0.8978**	**0.1092**
Marigold	Iris caption	**0.9589**	0.0610	0.8890	0.1108
Marigold	ours: T^g	0.9586	0.0612	0.8914	0.1095

Contributions of hierarchical language and SSAA. Table 6 progressively adds global language, depth-structured local descriptions, and SSAA. Global language alone causes only modest changes. Adding foreground, midground, and background descriptions improves eight of the nine reported quantities: zero-shot RMSE decreases from 4.152 to 4.107 and δ_1 increases from 0.543 to 0.555, while AbsRel decreases from 0.309 to 0.297. This behavior is consistent with local language providing spatially differentiated calibration cues. SSAA primarily improves cross-domain robustness. Relative to hierarchical language alone, it reduces zero-shot AbsRel from 0.297 to 0.288 (3.0%), RMSE from 4.107 to 3.951, and increases δ_1 from 0.555 to 0.567, while seen-domain changes remain smaller. This supports the intended division of roles: local language guides spatial corrections, whereas SSAA organizes scene-level metric scale. Supplementary Material D.7 compares semantic, spatial, and depth-based partitioning strategies: semantic grouping provides only marginal gains, spatial grouping performs better, and depth-aligned grouping performs best.

Effectiveness of diffusion feature scale Table 7 compares U-Net features at spatial scales F^8, F^{16}, F^{32}, and F^{64}. The intermediate F^{16} feature performs best on all reported seen-domain and zero-shot quantities, with a zero-shot AbsRel of 0.288 versus 0.314, 0.329, and 0.312 for the other scales. This suggests that metric calibration benefits from an intermediate representation that retains spatial detail while providing sufficient semantic context. Figure 4 shows the same trend qualitatively.

Table 6: Ablation study on the main modules. Zero-shot Avg denotes the unweighted average over the eight zero-shot test sets.

Image	L_g	L_{local}	SSAA	NYUv2 AbsRel ↓	NYUv2 RMSE ↓	NYUv2 δ_1 ↑	KITTI AbsRel ↓	KITTI RMSE ↓	KITTI δ_1 ↑	Zero-shot Avg AbsRel ↓	Zero-shot Avg RMSE ↓	Zero-shot Avg δ_1 ↑
✓				0.080	0.295	0.953	0.059	2.526	0.962	0.313	4.189	0.541
✓	✓			0.079	0.292	0.952	0.060	2.510	0.961	0.309	4.152	0.543
✓	✓	✓		0.078	0.286	0.954	0.060	2.470	0.963	0.297	4.107	0.555
✓	✓	✓	✓	**0.077**	**0.278**	**0.956**	**0.058**	**2.387**	**0.964**	**0.288**	**3.951**	**0.567**

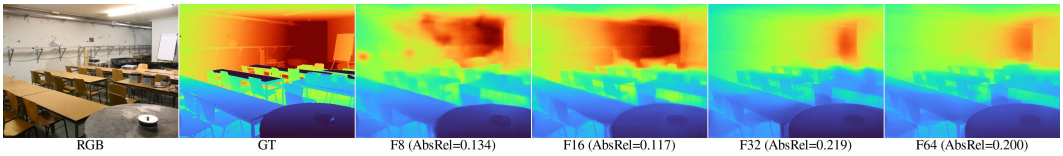


Figure 4: Qualitative comparison of U-Net feature scales on an iBims-1 scene. The columns show the RGB image, ground-truth depth, and predictions obtained from F^8, F^{16}, F^{32}, and F^{64}, respectively. The corresponding AbsRel values are reported below each prediction.

Table 7: Ablation of diffusion feature sources on the U-Net.

F^8	F^{16}	F^{32}	F^{64}	NYUv2 AbsRel ↓	NYUv2 RMSE ↓	NYUv2 δ_1 ↑	KITTI AbsRel ↓	KITTI RMSE ↓	KITTI δ_1 ↑	Zero-shot Avg AbsRel ↓	Zero-shot Avg RMSE ↓	Zero-shot Avg δ_1 ↑
✓				0.097	0.346	0.919	0.068	2.769	0.953	0.314	4.846	0.491
	✓			**0.077**	**0.278**	**0.956**	**0.058**	**2.387**	**0.964**	**0.288**	**3.951**	**0.567**
		✓		0.105	0.359	0.906	0.070	2.789	0.948	0.329	4.918	0.476
			✓	0.095	0.328	0.927	0.064	2.411	0.960	0.312	4.184	0.521

4 Conclusion

We presented a relative-to-metric calibration framework that metrically grounds the transferable geometry of frozen diffusion-depth models without camera metadata, external geometric measurements, or test-time ground-truth alignment. Our central observation is that scale-relevant semantic cues play different geometric roles across a scene and should therefore be organized according to their spatial context rather than compressed into a single global description. Based on this insight, we introduced hierarchical scene language to associate foreground, midground, and background semantics with spatial diffusion features, while retaining the frozen relative disparity as an explicit geometric scaffold. We further proposed Scene-Scale Anchor Alignment (SSAA) to complement local calibration with scene-level metric supervision. Experiments across four diffusion-depth variants and multiple metric-depth benchmarks demonstrate consistent improvements in metric calibration and strong transfer to unseen indoor, outdoor, real, and synthetic domains.

Limitations. Although our structured language dataset contains objects, relationships, coarse locations, and regional descriptions, the current model primarily uses the regional descriptions as implicit conditioning signals. The correspondence between each local description and its relevant image regions is learned only through metric-depth supervision, without explicit region-level alignment or depth-layer constraints. Consequently, the model may not fully exploit the fine-grained geometric information contained in the structured annotations. Future work could introduce text-derived region masks, explicit text–depth alignment objectives, or confidence-aware layer-wise fusion to more effectively utilize the structured language.

AI use statement

Generative AI tools were used to produce structured scene descriptions and to assist literature discovery and manuscript drafting. The authors are responsible for verifying the descriptions, technical claims, citations, code, and final text. This statement will be updated before submission to identify the final set of tools and the exact tasks for which they were used.

REFERENCES

Dylan Auty and Krystian Mikolajczyk. Learning to prompt clip for monocular depth estimation: Exploring the limits of human language. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 2031–2049. IEEE, 2023.

Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2. *arXiv preprint arXiv:2001.10773*, 2020.

Beilei Cui, Yiming Huang, Long Bai, and Hongliang Ren. Tr2m: Transferring monocular relative depth to metric depth with language descriptions and dual-level scale-oriented contrast. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 34181–34192, 2026.

David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27, 2014.

Xiao Fu, Wei Yin, Mu Hu, Kaixuan Wang, Yuexin Ma, Ping Tan, Shaojie Shen, Dahua Lin, and Xiaoxiao Long. Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. In *European Conference on Computer Vision*, pp. 241–258. Springer, 2024.

Gonzalo Martin Garcia, Karim Abou Zeid, Christian Schmidt, Daan De Geus, Alexander Hermans, and Bastian Leibe. Fine-tuning image-conditional diffusion models is easier than you think. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pp. 753–762, 2025.

Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 3354–3361. IEEE, 2012.

Ming Gui, Johannes Schusterbauer, Ulrich Prestel, Pingchuan Ma, Dmytro Kotovenko, Olga Grebenkova, Stefan Andreas Baumann, Vincent Tao Hu, and Björn Ommer. Depthfm: Fast generative monocular depth estimation with flow matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 3203–3211, 2025.

Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In *CVPR*, pp. 2485–2494, 2020.

Jing He, Haodong Li, Wei Yin, Yixun Liang, Leheng Li, Kaiqiang Zhou, Hongbo Zhang, Bingbing Liu, and YingCong Chen. Lotus: Diffusion-based visual foundation model for high-quality dense prediction. In *International Conference on Learning Representations*, volume 2025, pp. 89454–89467, 2025.

Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *CVPR*, pp. 9492–9502, 2024.

Sohee Kim, Jisu Kang, Dunam Kim, and Seokju Lee. Clip can understand depth. *Pattern Recognition*, pp. 112475, 2025.

Youngjung Kim, Hyungjoo Jung, Dongbo Min, and Kwanghoon Sohn. Deep monocular depth estimation via integration of global and local predictions. *IEEE transactions on Image Processing*, 27(8):4131–4144, 2018.

Tobias Koch, Lukas Liebel, Friedrich Fraundorfer, and Marco Körner. Evaluation of cnn-based single-image depth estimation methods. In *European Conference on Computer Vision*, pp. 331–348. Springer, 2018.

René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3):1623–1637, 2020.

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. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10892–10902. IEEE, 2021.

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European conference on computer vision*, pp. 746–760. Springer, 2012.

Eunjin Son and Sang Jun Lee. Cabins: Clip-based adaptive bins for monocular depth estimation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 4557–4567. IEEE, 2024.

Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 567–576, 2015.

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

Alex Wong, Xiaohan Fei, Stephanie Tsuei, and Stefano Soatto. Unsupervised depth completion from visual inertial odometry. *IEEE Robotics and Automation Letters*, 5(2):1899–1906, 2020.

Ziyao Zeng, Daniel Wang, Fengyu Yang, Hyoungseob Park, Stefano Soatto, Dong Lao, and Alex Wong. Wordepth: Variational language prior for monocular depth estimation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9708–9719. IEEE, 2024a.

Ziyao Zeng, Yangchao Wu, Hyoungseob Park, Daniel Wang, Fengyu Yang, Stefano Soatto, Dong Lao, Byung-Woo Hong, and Alex Wong. Rsa: Resolving scale ambiguities in monocular depth estimators through language descriptions. *Advances in neural information processing systems*, 37: 112684–112705, 2024b.

Ziyao Zeng, Jingcheng Ni, Daniel Wang, Patrick Rim, Younjoon Chung, Fengyu Yang, Byung-Woo Hong, and Alex Wong. Iris: Integrating language into diffusion-based monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 34193–34205, 2026.

Mingxia Zhan, Li Zhang, Beibei Wang, Yingjie Wang, and Zenglin Shi. Language as prior, vision as calibration: Metric scale recovery for monocular depth estimation. *arXiv preprint arXiv:2601.01457*, 2026.

Renrui Zhang, Ziyao Zeng, Ziyu Guo, and Yafeng Li. Can language understand depth? In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 6868–6874, 2022.