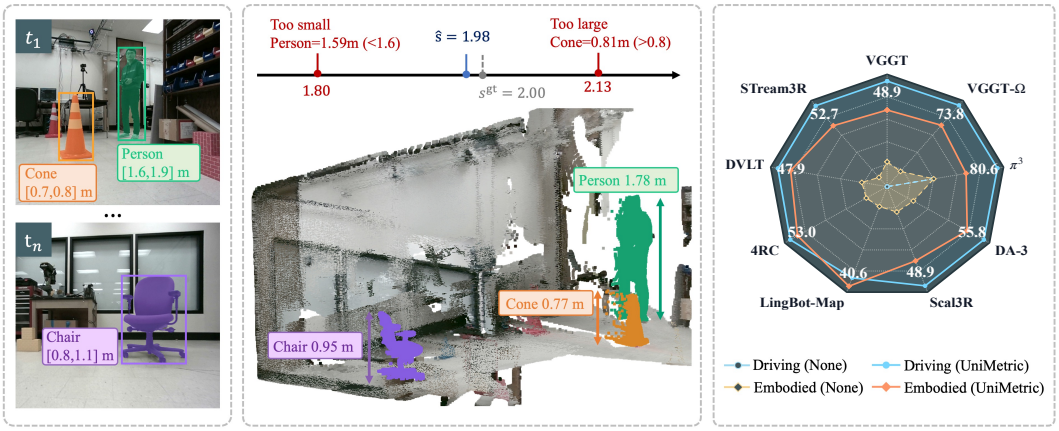
UNIMETRIC: UNIVERSAL METRIC SCALE RECOVERY FOR VISUAL GEOMETRY FOUNDATION MODELS

Anonymous authors
Paper under double-blind review



(a) Coarse Size Priors (b) Jointly Constrained Scene Scale (c) Performance

Figure 1: **Uncertain object-size priors jointly constrain metric scale.** (a) A vision–language model provides plausible height intervals for familiar objects across frames. (b) Combined with reconstructed object heights, these intervals constrain a shared scale. A candidate consistent with one object may violate another's prior; their joint constraints narrow the admissible range. Here, $\hat{s}$ and s^{gt} denote the estimated and reference scales. (c) UniMetric improves metric depth accuracy across nine geometry foundation models on MetricBench's driving-and-infrastructure and embodied-AI suites. Curves compare original (None) and scale-corrected (UniMetric) predictions using δ_1 (%, higher is better), averaged equally over six subsets per suite. Evaluation uses no ground-truth scale alignment.

ABSTRACT

Recent advances in multimodal foundation models, exemplified by GPT-6 Astra, raise the prospect of general-purpose embodied reasoning and planning. Using Internet videos for embodied pretraining calls for reliable physical-scale annotations. Yet visual geometry foundation models often recover scene structure and camera motion without reliable metric scale. We show that uncertain size priors of familiar objects can jointly constrain a reconstruction's shared scale. Building on this insight, we present UniMetric, a universal and training-free metric scale recovery framework for geometry foundation models, without additional sensors or task-specific metric supervision. Specifically, Metric Prior Acquisition elicits object height intervals from a vision–language model. Semantic–Geometric Alignment measures reconstructed object heights along the estimated support-surface normal in each frame and converts the priors into scale intervals. Global Scale Recovery reconciles these intervals through a convex objective, yielding a jointly feasible scale when they overlap and a unique compromise otherwise. The estimated sequence-level scale is applied uniformly to reconstructed points and camera translations. To evaluate metric accuracy without ground-truth scale alignment, we introduce MetricBench, covering twelve embodied, driving, and roadside subsets with sensor-derived ground truth. Across nine geometry foundation models, UniMetric reduces mean depth Abs Rel, reconstruction Chamfer distance, and trajectory error by 54.2%, 83.8%, and 65.6%, respectively, improving every model on all three tasks at the suite level. Our codebase and the MetricBench benchmark will be released.

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.

1 INTRODUCTION

Recent advances in general-purpose multimodal foundation models, exemplified by GPT-6 Astra (OpenAI, 2026), raise the prospect of a general-purpose embodied "brain" for reasoning and planning across diverse physical platforms (Cheng et al., 2026). Grounding these capabilities in the physical world requires understanding real dimensions, distances, and motion. This motivates enriching Internet videos with reliable physical-scale annotations to support embodied training at scale. Visual geometry foundation models (GFMs) (Wang et al., 2025a;b) can recover dense structure and camera motion from such videos, yet their reconstructions often lack reliable metric scale. Bridging this gap is therefore key to turning large-scale visual experience into metric data for embodied foundation-model pretraining.

The missing scale is commonly supplied by additional sensor measurements (Campos et al., 2021; Qin et al., 2018; Tian et al., 2021) or learned from metric supervision (Bhat et al., 2023; Hu et al., 2024; Piccinelli et al., 2024; Keetha et al., 2025). These routes provide metric information through the acquisition setup or metric-labelled training data. For a pretrained geometry model, however, a practical goal is to recover the physical unit without introducing new sensing requirements or training an additional metric predictor. This raises the question we study: **can the content of a scene provide enough information to recover the metric scale of its reconstruction?**

Our motivation is that uncertainty about individual object sizes need not imply equally large uncertainty about a scene's scale. Familiar objects such as cars, doors, and pedestrians have plausible physical size ranges, even when their exact dimensions are unknown. Each object alone provides only a coarse reference. Within a reconstruction, however, these references constrain the same scale: a factor that makes the car plausible must also make the door and the pedestrian plausible. A scale acceptable for one object may be ruled out by another (Fig. 1(b)). Their joint compatibility can therefore narrow the range of plausible scene scales without requiring any individual object to have an exactly known size. The opportunity lies in using uncertain references to constrain shared scale.

Building on this motivation, we present UniMetric (Sec. 3.2), a universal and training-free metric scale recovery framework for geometry foundation models. **Metric Prior Acquisition** queries a pretrained vision–language model (VLM) to identify familiar objects and provide their plausible height ranges. **Semantic–Geometric Alignment** measures each object in the reconstruction along the support direction estimated for its frame and converts its height range into an interval of admissible scales. **Global Scale Recovery** combines these intervals across objects and frames by minimizing a convex objective whose zero set is exactly their common intersection. The recovered scalar uniformly rescales the pointmaps and camera translations, preserving relative geometry without modifying the underlying GFM.

Evaluating scale recovery requires retaining the scale error rather than removing it during alignment. A protocol that fits scale to ground truth (Umeyama, 1991; Eigen et al., 2014) cannot reveal whether a prediction already has the correct physical unit. We therefore introduce MetricBench (Sec. 4.1), comprising twelve dataset subsets across driving, roadside, and embodied scenarios with sensor-derived metric references. Its protocol evaluates depth, 3D reconstruction, and camera trajectories where applicable, allowing rigid registration when needed but never ground-truth scale fitting. Across nine GFMs, UniMetric reduces mean depth absolute relative error (Abs Rel), reconstruction Chamfer distance (CD), and mean absolute trajectory error (ATE Mean) by 54.2%, 83.8%, and 65.6%, respectively, relative to uncorrected predictions. Every evaluated model improves its suite-averaged performance on all three tasks (Table 3), demonstrating the practical benefit of a shared scale correction across scene geometry and camera motion.

Our contributions are threefold:

1. We recast metric scale recovery as joint estimation from uncertain semantic size priors, and show that coarse object-size knowledge, imprecise for any single object, yields a consistent scale estimate when many objects are combined.
2. We propose UniMetric, a universal and training-free framework that recovers metric scale for any geometry foundation model by reconciling semantic size intervals with its reconstruction.
3. We introduce MetricBench, a cross-platform benchmark that measures metric scale directly, on which UniMetric reduces depth, reconstruction, and trajectory errors by 54.2%, 83.8%, and 65.6% on average across nine geometry foundation models.

2 Related Work

Visual Geometry Foundation Models. Building on the pairwise pointmaps of DUSt3R (Wang et al., 2024), VGGT (Wang et al., 2025a), π^3 (Wang et al., 2025b), and Depth Anything 3 (Lin et al., 2025) infer structure and cameras for many views in one forward pass, and later work extends the formulation to longer, streaming, and dynamic inputs (Wang et al., 2026; Xie et al., 2026; Burzio et al., 2026; Lan et al., 2025; Chen et al., 2026; Luo et al., 2026). Their training targets are scale-normalized and their results are reported after per-image median scaling (Eigen et al., 2014) or similarity alignment (Umeyama, 1991), so the metric scale of their output is neither constrained nor measured (Sec. 4.1). UniMetric leaves their geometry untouched and supplies the scale.

Metric Scale from Sensors and Supervision. Metric scale is conventionally observed by calibrated stereo, depth, or inertial sensors (Campos et al., 2021; Qin et al., 2018; Salih et al., 2026) or by a known camera height (Tian et al., 2021), or learned from metric labels in single-image depth (Bhat et al., 2023; Yin et al., 2023; Hu et al., 2024; Piccinelli et al., 2024; Bochkovskiy et al., 2025) and its video and multi-view extensions (Chen et al., 2025; Keetha et al., 2025); language-derived scale predictors (Zeng et al., 2024; Wu & Chen, 2025) are trained in the same way. The first route requires hardware; the second inherits the scale statistics of its training data. UniMetric uses neither: it combines the reconstruction of an existing model with the size knowledge of an existing vision–language model, without training. Object size priors have previously scaled filtering-based monocular SLAM, using height priors for a small set of predefined detector classes (Sucar & Hayet, 2017); UniMetric instead obtains open-vocabulary height intervals from a VLM and applies to any geometry foundation model.

3 Method

Given an RGB sequence $\mathcal{I} = \{I_t\}_{t=1}^{T}$ and a pretrained geometry foundation model (GFM) $\mathcal{G}$, we estimate a positive scale $\hat{s}$ that converts reconstructed points and camera translations from a shared unknown unit into metres. RGB images are the only external observations, and the correction preserves relative geometry.

3.1 Metric Scale Recovery from Uncertain Semantic Size Priors

We study how aggregating uncertain object-size priors affects scale accuracy when relative object heights are exact. We first analyze point estimates, then evaluate real VLM height intervals in a controlled study.

Mathematical analysis of joint scale estimation. For observation i, let $H_i > 0$ be the physical height in metres and $h_i > 0$ its relative height, i.e., the height in the arbitrary unit of the reconstruction. They satisfy $H_i = s_\star h_i$, where $s_\star > 0$ is the shared scale to recover. A semantic height estimate $\widetilde{H}_i$ therefore provides a scale observation

$$y_i = \frac{\widetilde{H}_i}{h_i} = s_\star + \epsilon_i, \qquad \epsilon_i = \frac{\widetilde{H}_i - H_i}{h_i}. \tag{1}$$

Although objects differ in height, these ratios estimate the same quantity. We combine N observations using their mean, $\bar{s}_N = N^{-1} \sum_{i=1}^{N} y_i$, with expectations taken over semantic errors while the relative heights are held fixed.

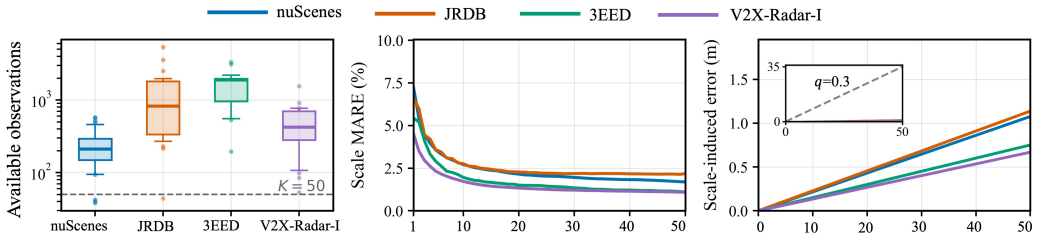
Proposition 1 (Mean-square consistency of joint scale estimation). *Suppose the errors $\{\epsilon_i\}_{i\geq 1}$ have zero means, are pairwise uncorrelated, and satisfy $\mathrm{Var}(\epsilon_i) \leq \sigma^2 < \infty$ for all i. Then*

$$\mathbb{E}\big[(\bar{s}_N - s_\star)^2\big] = \frac{1}{N^2} \sum_{i=1}^{N} \mathrm{Var}(\epsilon_i) \leq \frac{\sigma^2}{N}, \tag{2}$$

and $\bar{s}_N$ converges in mean square to $s_\star$ (proof in Appendix A.1).

Under these assumptions, the error bound decreases as observations accumulate, even when individual priors remain noisy. This analysis concerns point priors, the special case of UniMetric's interval

objective in which every interval collapses to a point (Appendix A.2); we next evaluate the interval estimator with real VLM priors.



(a) Anchor availability (b) Joint scale estimation (c) Distance error, K=20

Figure 2: **Joint scale estimation with real VLM height priors and exact relative heights.** (a) Available frame–object observations per sequence; the dashed line marks the budget of 50 observations. (b) Mean absolute relative scale error as the observation budget increases. (c) Scale-induced distance error computed from the results in (b) at $K = 20$; the inset shows the uncorrected reference.

Controlled evaluation with VLM height intervals. We construct relative heights from annotated object heights as $h_i = qH_i$, with $q = 0.3$, so that the true recovery factor is $s_\star = 1/q$. For each annotated object, the VLM provides a height interval and a confidence without access to H_i or q. These priors are combined with the relative heights using UniMetric's weighted interval estimator (Sec. 3.2).

The study covers 193 sequences from the MetricBench subsets of nuScenes, JRDB, 3EED, and V2X-Radar-I. We vary the budget K from 1 to 50 frame–object observations per sequence, using all available observations when fewer than K exist. We report mean absolute relative scale error (MARE), obtained by averaging $100\,|\hat{s}/s_\star - 1|$ over sampling trials and sequences within each subset (Appendix B).

Increasing the observation budget from $K = 1$ to 50 lowers MARE from 4.5–7.2% to 1.1–2.2% across the four subsets (Fig. 2(b); Table S2). These results support joint scale estimation with real semantic intervals. Sec. 3.2 next replaces the annotation-derived relative heights with measurements from GFM reconstructions.

3.2 UniMetric

Overall Architecture. UniMetric combines semantic size priors with the relative geometry that a GFM reconstructs from the same RGB sequence $\mathcal{I} = \{I_t\}_{t=1}^{T}$ (Fig. 3). The GFM provides camera-frame pointmaps $X_t \in \mathbb{R}^{H\times W\times 3}$, depth maps D_t, and world-to-camera poses $[R_t \mid \mathbf{t}_t]$, whose points and translations we assume to share one unknown unit across frames. The three modules below turn object height priors into scale intervals and estimate the scalar that converts the reconstruction to metric units.

Metric Prior Acquisition. A pretrained vision–language model $\mathcal{V}$, Seed1.8 (ByteDance Seed, 2026) in our experiments, identifies familiar objects and provides their plausible physical heights. For each frame I_t, it returns candidate anchors and the dominant support surface:

$$\Big(\big\{(c_{t,i},\, b_{t,i},\, [H^-_{t,i}, H^+_{t,i}],\, q_{t,i})\big\}_{i=1}^{N_t},\; b^s_t\Big) = \mathcal{V}(I_t). \tag{3}$$

Here, $c_{t,i}$ and $b_{t,i}$ denote the object category and bounding box, $[H^-_{t,i}, H^+_{t,i}]$ is its height interval in metres, $q_{t,i} \in [0, 1]$ is the reported confidence, and b^s_t localizes the support surface. Retaining both height bounds preserves uncertainty about the instance's dimensions rather than assigning a single nominal size (prompt templates in Appendix C.1).

We retain non-truncated anchors standing upright on the selected support surface and remove duplicate detections within each frame. Box-prompted Segmentation then converts the retained object and support boxes into pixel masks:

$$M_{t,i} = \mathcal{S}(I_t,\, b_{t,i}), \qquad M^s_t = \mathcal{S}(I_t,\, b^s_t), \tag{4}$$

where $\mathcal{S}$ is a pretrained segmentation model (Ravi et al., 2024). This module therefore supplies both metric height intervals and the image regions needed to associate them with reconstructed objects.

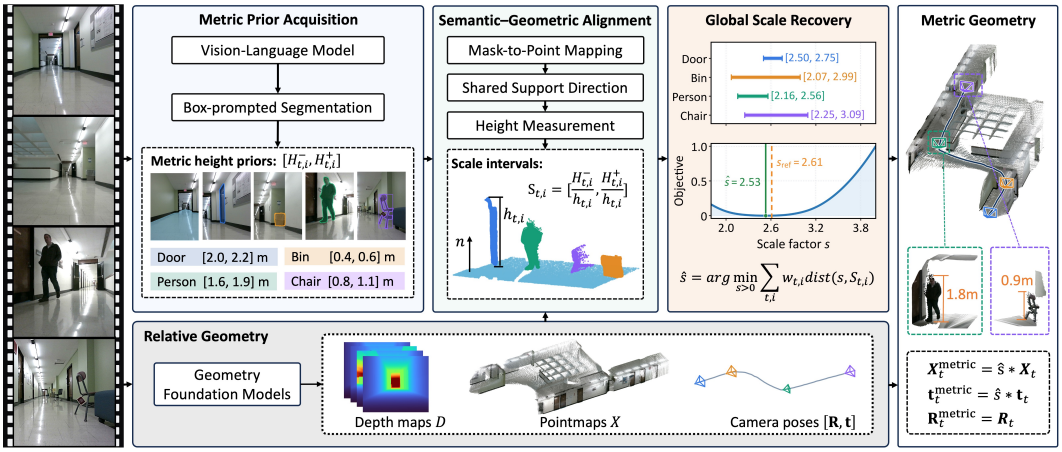


Figure 3: **Overview of UniMetric.** Metric Prior Acquisition obtains object height intervals and masks from RGB images. Semantic–Geometric Alignment measures reconstructed heights along frame-specific support directions and converts the priors into scale intervals. Global Scale Recovery combines these intervals through a weighted convex objective, yielding a shared scale for reconstructed geometry and camera translations.

Semantic–Geometric Alignment. This module connects the semantic and geometric branches through the three operations shown in Fig. 3. Its goal is to measure the reconstructed extent corresponding to each height prior and express their relationship as a constraint on scale.

Mask-to-Point Mapping. The object and support masks select the corresponding samples from each pointmap. Using the predicted camera pose, we transform these samples into the common reconstruction frame:

$$\begin{aligned} \mathcal{P}_{t,i} &= \left\{ R_t^\top \left(X_t(\mathbf{u}) - \mathbf{t}_t \right) \;\middle|\; M_{t,i}(\mathbf{u}) = 1 \right\}, \\ \mathcal{P}_t^s &= \left\{ R_t^\top \left(X_t(\mathbf{u}) - \mathbf{t}_t \right) \;\middle|\; M_t^s(\mathbf{u}) = 1 \right\}, \end{aligned} \tag{5}$$

where $\mathbf{u}$ indexes pixels. The resulting point sets associate each semantic anchor and its support surface with their reconstructed geometry.

Shared Support Direction. To measure height without assuming that a camera axis is vertical, we fit a plane to the support points and use its unit normal:

$$\mathbf{n}_t = \mathrm{PlaneFit}\left(\mathcal{P}_t^s\right). \tag{6}$$

This direction is shared by the anchors within frame t, but estimated independently across frames. The method therefore does not impose a single support plane on the entire sequence.

Height Measurement. For anchors whose height axes align with $\mathbf{n}_t$, the reconstructed height is the extent of their points along this direction:

$$h_{t,i} = \max_{\mathbf{p} \in \mathcal{P}_{t,i}} \mathbf{n}_t^\top \mathbf{p} - \min_{\mathbf{p} \in \mathcal{P}_{t,i}} \mathbf{n}_t^\top \mathbf{p}. \tag{7}$$

The support plane supplies only the direction; both endpoints come from the object's own points, which must cover its vertical extent. The resulting $h_{t,i}$ and the semantic prior describe the same physical height in reconstruction and metric units, respectively. For $h_{t,i} > 0$, a scale s is admissible when the rescaled height lies within the prior:

$$s\, h_{t,i} \in \left[H_{t,i}^-,\, H_{t,i}^+ \right] \iff s \in S_{t,i} := \left[\frac{H_{t,i}^-}{h_{t,i}},\, \frac{H_{t,i}^+}{h_{t,i}} \right]. \tag{8}$$

Each retained frame–object observation thus constrains the same sequence-level conversion to metres.

Global Scale Recovery. We collect the scale intervals across objects and frames and jointly estimate a single scalar (Fig. 3). Each observation receives a weight

$$w_{t,i} = q_{t,i}\, \alpha_{c_{t,i}}, \tag{9}$$

where $\alpha_c \in [0, 1]$ is a category-specific reliability weight, and $\mathcal{A}$ denotes the retained observations with positive weights. A candidate scale is penalized only outside an interval, by its distance to the nearest bound: for $S = [a, b]$,

$$\mathrm{dist}(s, S) = \max\{a - s,\ 0,\ s - b\}, \tag{10}$$

and minimize the weighted sum of squared interval distances:

$$\hat{s} \in \arg\min_{s>0}\ J(s), \qquad J(s) = \sum_{(t,i)\in\mathcal{A}} w_{t,i}\ \mathrm{dist}(s, S_{t,i})^2. \tag{11}$$

The weighted interval objective is convex and piecewise quadratic. Its minimizers coincide with the common intersection when one exists; otherwise, it admits a unique minimizer. We solve this one-dimensional problem by golden-section search (Kiefer, 1953) over the interval hull (Appendix A.2).

When all intervals collapse to points and the weights are equal, Eq. (11) reduces to the mean estimator analyzed in Sec. 3.1.

Finally, we apply the recovered scale to the reconstruction:

$$X_t^m = \hat{s}\, X_t, \qquad \mathbf{t}_t^m = \hat{s}\, \mathbf{t}_t, \qquad R_t^m = R_t, \qquad t = 1, \ldots, T. \tag{12}$$

The same correction applies to scene structure and camera motion, leaving rotations, intrinsics, and relative geometry unchanged. The associated depth maps inherit the scaling, $D_t^m = \hat{s}\, D_t$, through their camera-frame pointmaps.

4 Experiments

4.1 MetricBench

We introduce **MetricBench**, a cross-platform benchmark for evaluating metric geometry across diverse Physical AI settings. It brings together **twelve dataset subsets comprising 694 sequences and 51,339 frames**, organized into **driving-and-infrastructure** and **embodied-AI** suites (Table 1). The subsets are drawn from KITTI (Geiger et al., 2012), nuScenes (Caesar et al., 2020), Waymo (Sun et al., 2020), Argoverse (Chang et al., 2019), V2X-Radar-I (Yang et al., 2025), DAIR-V2X-I (Yu et al., 2022), CEAR (Zhu et al., 2024), JRDB (Martin-Martin et al., 2021), FusionPortable (Jiao et al., 2022), DROID (Khazatsky et al., 2024), Humanoid Everyday (Zhao et al., 2025), and 3EED (Li et al., 2025). MetricBench combines broad platform and scene coverage with a unified evaluation of depth, 3D reconstruction, and ego pose.

Table 1: **MetricBench: hardware platforms and evaluation subset sizes.** Seq.: sequences; UGV: unmanned ground vehicle. † marks partial subsets (Appendix D.1).

Suite	Datasets	Hardware platforms	Data size
Driving & infrastructure	KITTI, nuScenes, Waymo, Argoverse, V2X-Radar-I, DAIR-V2X-I	Vehicle; roadside unit	Seq.: 547 Frames: 27,198
Embodied AI	CEAR, JRDB, FusionPortable, DROID, Humanoid Everyday†, 3EED†	Quadruped; wheeled robot/UGV; handheld rig; robot arm; humanoid; drone	Seq.: 147 Frames: 24,141

Platform and scene coverage. The benchmark spans vehicles, roadside units, wheeled and quadruped robots, robotic arms, humanoids, drones, and handheld rigs, covering diverse viewpoints and acquisition setups. Its two suites encompass driving and roadside perception alongside indoor and outdoor embodied scenes, including close-range manipulation and aerial observations. Beyond platform diversity, the subsets vary substantially in physical scale: median reference depth ranges from 0.695 m on Humanoid Everyday to 36.865 m on V2X-Radar-I, representing more than a $50\times$ variation in characteristic scene depth.

Reference data and evaluation. Each evaluated frame contains a matched RGB image, a calibrated metric point cloud, and camera metadata. Reference geometry comes from LiDAR or depth cameras, with dataset-provided metric poses supplying trajectory references where applicable. We report Abs Rel and δ_1 (Eigen et al., 2014) for depth, Chamfer distance (CD) for 3D reconstruction, and mean absolute trajectory error (ATE Mean; Sturm et al., 2012) for ego pose. Evaluation permits rigid $SE(3)$ registration where coordinate frames differ, but never fits scale to ground truth, retaining errors in both physical scale and geometry. The two static roadside subsets are excluded from ego-pose evaluation. Suite-level results weight the applicable subsets equally (Appendix D).

Table 2: **Metric depth estimation on MetricBench.** Absolute relative depth error (Abs Rel, lower is better) before (None) and after UniMetric correction, without ground-truth scale alignment. Bold marks the better result within each None/UniMetric pair. DA-3 denotes Depth Anything 3.

Model	Scale Recovery	Autonomous Driving and Infrastructure						Embodied AI					
		KITTI	nuScenes	Waymo	Argoverse	V2X-Radar-I	DAIR-V2X-I	CEAR	JRDB	Fusion Portable	DROID	Humanoid Everyday	3EED
VGGT	None	0.974	0.983	0.983	0.974	0.943	0.965	0.850	0.805	0.949	0.439	0.359	0.936
	UniMetric	**0.722**	**0.316**	**0.311**	**0.311**	**0.392**	**0.866**	**0.551**	**0.520**	**0.676**	**0.332**	**0.335**	**0.534**
VGGT-Ω	None	0.971	0.979	0.983	0.972	0.949	0.966	0.786	0.790	0.885	0.294	0.311	0.937
	UniMetric	**0.347**	**0.152**	**0.154**	**0.244**	**0.123**	**0.542**	**0.538**	**0.312**	**0.331**	**0.284**	**0.221**	**0.343**
π^3	None	0.190	0.297	0.165	0.282	0.341	0.704	0.328	0.310	0.330	0.315	1.009	0.354
	UniMetric	**0.188**	**0.098**	**0.095**	**0.140**	**0.147**	**0.471**	**0.322**	**0.253**	**0.219**	**0.221**	**0.480**	**0.295**
DA-3	None	0.967	0.948	0.962	0.937	0.922	0.956	0.754	0.788	0.894	0.335	0.482	0.915
	UniMetric	**0.372**	**0.331**	**0.162**	**0.425**	**0.168**	**0.531**	**0.406**	**0.345**	**0.358**	**0.327**	**0.364**	**0.436**
Scal3R	None	0.966	0.971	0.974	0.959	0.938	0.967	0.763	0.770	0.904	0.338	0.320	0.898
	UniMetric	**0.231**	**0.234**	**0.241**	**0.287**	**0.411**	**0.856**	**0.361**	**0.528**	**0.478**	**0.336**	**0.263**	**0.489**
LingBot-Map	None	0.958	0.971	0.975	0.964	0.952	0.973	0.749	0.799	0.887	0.356	0.381	0.929
	UniMetric	**0.244**	**0.272**	**0.325**	**0.343**	**0.351**	**0.658**	**0.283**	**0.338**	**0.403**	**0.297**	**0.282**	**0.347**
4RC	None	0.973	0.979	0.985	0.976	0.938	0.960	0.777	0.789	0.906	0.324	0.436	0.932
	UniMetric	**0.326**	**0.310**	**0.192**	**0.480**	**0.151**	**0.524**	**0.579**	**0.287**	**0.370**	**0.302**	**0.270**	**0.340**
DVLT	None	0.957	0.968	0.974	0.962	0.881	0.926	0.672	0.569	0.867	1.001	0.722	0.898
	UniMetric	**0.435**	**0.220**	**0.156**	**0.245**	**0.171**	**0.576**	**0.448**	**0.392**	**0.447**	**0.264**	**0.348**	**0.538**
STream3R	None	0.960	0.971	0.970	0.961	0.956	0.970	0.719	0.796	0.858	0.421	0.379	0.932
	UniMetric	**0.515**	**0.205**	**0.242**	**0.250**	**0.497**	**0.685**	**0.380**	**0.330**	**0.734**	**0.241**	**0.317**	**0.449**

4.2 Experimental Settings

Models and comparison settings. We evaluate UniMetric with nine geometry foundation models: VGGT (Wang et al., 2025a), VGGT-Ω (Wang et al., 2026), π^3 (Wang et al., 2025b), Depth Anything 3 (DA-3; Lin et al., 2025), Scal3R (Xie et al., 2026), LingBot-Map (Chen et al., 2026), 4RC (Luo et al., 2026), Déjà View (DVLT; Burzio et al., 2026), and STream3R (Lan et al., 2025). Each GFM uses its released weights without additional training. For each sequence, *None* and *UniMetric* share the same GFM predictions and differ only in the scale applied to depth, reconstructed points, and camera translations. The VLM (Seed1.8) and segmentation model (SAM 2.1) are shared across GFMs, while geometric filtering follows the dataset–GFM configuration.

4.3 Quantitative Results

Table 3: **Suite-averaged metric geometry on MetricBench.** Metrics are averaged equally over applicable subsets; Average further averages over the nine models. δ_1 denotes depth accuracy within a factor of 1.25 (%), CD denotes Chamfer distance (m), and ATE denotes mean absolute trajectory error (m). Roadside subsets are excluded from ATE. Bold marks the better result within each None/UniMetric pair.

Model	Scale Recovery	Autonomous Driving and Infrastructure				Embodied AI			
		Abs Rel↓	δ_1(%)↑	CD (m)↓	ATE (m)↓	Abs Rel↓	δ_1(%)↑	CD (m)↓	ATE (m)↓
VGGT	None	0.970	0.0	20.186	42.574	0.723	11.5	5.354	10.302
	UniMetric	**0.486**	**48.9**	**4.038**	**14.878**	**0.491**	**35.5**	**1.189**	**6.406**
VGGT-Ω	None	0.970	0.0	19.167	42.332	0.667	13.8	4.763	9.885
	UniMetric	**0.260**	**73.8**	**1.292**	**6.746**	**0.338**	**55.7**	**0.628**	**2.037**
π^3	None	0.330	33.0	2.506	8.869	0.441	34.1	0.880	3.298
	UniMetric	**0.190**	**80.6**	**1.018**	**4.417**	**0.298**	**57.7**	**0.691**	**1.969**
DA-3	None	0.949	0.0	18.716	41.834	0.695	15.1	4.891	9.864
	UniMetric	**0.331**	**55.8**	**2.239**	**14.285**	**0.373**	**46.2**	**0.918**	**3.708**
Scal3R	None	0.962	0.0	19.892	42.162	0.665	12.5	5.189	9.947
	UniMetric	**0.377**	**48.9**	**3.426**	**10.219**	**0.409**	**36.6**	**1.587**	**4.118**
LingBot-Map	None	0.966	0.0	18.815	41.769	0.683	8.1	4.654	9.899
	UniMetric	**0.365**	**37.2**	**2.724**	**11.071**	**0.325**	**40.6**	**0.679**	**3.311**
4RC	None	0.969	0.0	20.052	42.941	0.694	11.4	4.944	9.721
	UniMetric	**0.331**	**53.0**	**2.282**	**13.709**	**0.358**	**49.0**	**0.828**	**2.711**
DVLT	None	0.945	0.1	18.399	42.667	0.788	11.1	4.865	9.920
	UniMetric	**0.300**	**47.9**	**2.366**	**14.187**	**0.406**	**42.3**	**0.850**	**3.674**
STream3R	None	0.965	0.0	19.508	42.558	0.684	6.0	4.020	9.506
	UniMetric	**0.399**	**52.7**	**3.370**	**19.556**	**0.409**	**39.7**	**1.707**	**7.507**
Average	None	0.892	3.7	17.471	38.634	0.671	13.7	4.396	9.149
	UniMetric	**0.338**	**55.4**	**2.528**	**12.119**	**0.379**	**44.8**	**1.008**	**3.938**

Metric depth across models and scenes. UniMetric reduces Abs Rel in all 108 model–subset pairs in Table 2, with a 54.2% reduction in the mean error. Improvements span both driving and infrastructure scenes and embodied-AI environments. As summarized in Table 3, mean Abs Rel decreases from 0.892 to 0.338 in the driving suite and from 0.671 to 0.379 in the embodied suite. The larger reduction in driving scenes accompanies their larger uncorrected metric errors.

Shared correction across geometric tasks. The same estimated scale also improves reconstructed geometry and camera trajectories. Every GFM improves its suite-average performance on all three tasks in both suites (Table 3). Across model–subset pairs, CD decreases in 107 of 108 comparisons and ATE Mean in 81 of 90. On the driving suite, mean CD decreases from 17.471 to 2.528 m, and mean ATE decreases from 38.634 to 12.119 m, corresponding to reductions of 14.943 and 26.515 m, respectively. These improvements show that a scale estimated from semantic object priors can benefit multiple outputs of the same reconstruction without task-specific scale fitting. Per-subset gains nevertheless vary, reflecting the geometric errors that remain after rescaling.

4.4 Ablation Studies

We compare metric-depth calibration methods and examine UniMetric's estimation design using VGGT-Ω on nuScenes and FusionPortable.

Table 4: **Scale calibration of VGGT-Ω predictions.** All rows rescale the same GFM outputs by one factor per sequence, estimated from metric depth (Appendix C.2) or by UniMetric. Oracle (gray) fits the factor to ground truth separately for each metric. **Bold** and underline mark the best and second-best non-oracle results.

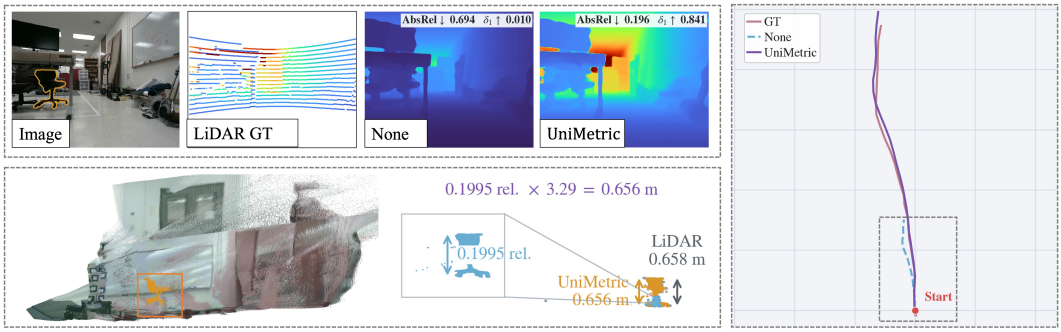
Scale calibration	nuScenes				FusionPortable			
	Abs Rel↓	δ_1 ↑	CD↓	ATE↓	Abs Rel↓	δ_1 ↑	CD↓	ATE↓
None	0.979	0.0	14.594	23.792	0.885	0.9	15.848	39.863
Depth Pro	0.212	55.6	1.107	5.134	0.543	22.6	9.378	66.352
UniDepth v2	0.278	53.1	1.638	13.515	0.485	14.8	1.450	23.485
Depth Anything V2	0.347	29.0	1.705	11.423	0.396	47.0	0.702	10.855
ZoeDepth-NK	0.218	65.9	1.044	6.965	0.432	49.4	3.940	56.668
UniMetric (Ours)	**0.152**	**82.0**	**0.726**	**2.476**	**0.331**	**52.2**	**0.653**	**4.238**
Oracle	0.087	92.9	0.546	1.407	0.219	80.1	0.384	2.972

Table 5: **Design ablations of UniMetric with VGGT-Ω.** The variants replace scale intervals with their midpoints (Point priors), assign equal weights to valid observations (Uniform weights), or retain only the lowest-weight observation (Single anchor). **Bold** and underline mark the best and second-best results.

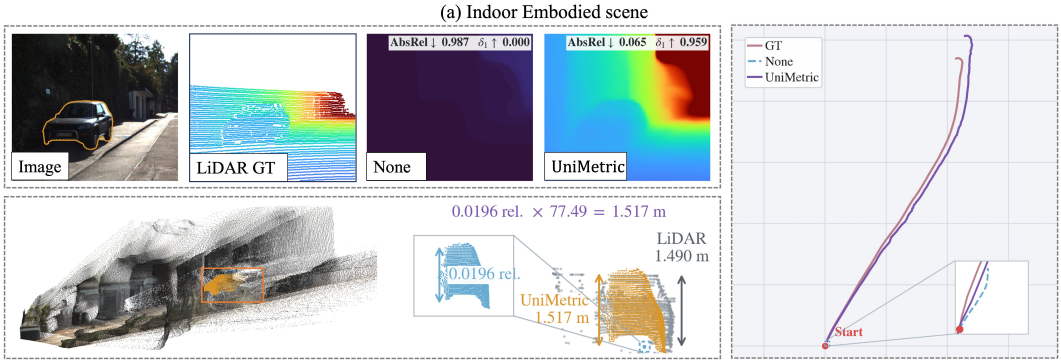
Variant	nuScenes				FusionPortable			
	Abs Rel↓	δ_1 ↑	CD↓	ATE↓	Abs Rel↓	δ_1 ↑	CD↓	ATE↓
Point priors	0.172	79.1	0.989	5.231	0.452	47.4	1.806	19.430
Uniform weights	0.180	67.5	1.235	5.902	0.469	46.9	1.828	22.636
Single anchor	0.334	45.4	1.983	21.643	0.683	18.0	6.086	48.649
UniMetric (full)	**0.152**	**82.0**	**0.726**	**2.476**	**0.331**	**52.2**	**0.653**	**4.238**

Comparison with metric-depth methods. We compare UniMetric with calibrating the same VGGT-Ω reconstruction using four pretrained metric depth estimators: Depth Pro (Bochkovskiy et al., 2025), UniDepth v2 (Piccinelli et al., 2025), Depth Anything V2 (Yang et al., 2024), and ZoeDepth (Bhat et al., 2023). Each baseline estimates one scale per sequence from its metric depth maps, matching the granularity of UniMetric (Appendix C.2). UniMetric achieves the best non-oracle results across all four metrics on both nuScenes and FusionPortable (Table 4). Oracle fits a ground-truth scale separately for each metric, indicating further room for improvement through global rescaling.

Ablation of key designs. The full method achieves the highest δ_1 accuracy on both nuScenes (**82.0%**) and FusionPortable (**52.2%**) (Table 5). It improves over midpoint priors by 2.9 and 4.8 percentage points, and over uniform weights by 14.5 and 5.3 percentage points, respectively. These gains support preserving interval uncertainty and weighting observations by reliability. Retaining only the lowest-weight observation yields δ_1 scores of 45.4% and 18.0%, respectively, showing that a single unreliable reference is not sufficient.



(a) Indoor Embodied scene



(b) Outdoor driving scenes

Figure 4: **Qualitative scale recovery with VGGT-Ω on CEAR (top) and KITTI (bottom).** Depth maps and camera trajectories compare predictions before (None) and after (UniMetric) scale correction. Point-cloud views show the highlighted regions and their measured heights alongside LiDAR references. These measurements cover visible regions rather than complete objects. GT denotes ground truth; rel. denotes the original reconstruction unit.

4.5 Qualitative Results

On an indoor CEAR sequence and an outdoor KITTI drive, UniMetric recovers scale factors of 3.29 and 77.49 for VGGT-Ω (Fig. 4). Both bring the highlighted anchors to within 3 cm of their LiDAR heights: the chair measures 0.656 m against 0.658 m, and the car 1.517 m against 1.490 m. On the displayed frames, Abs Rel falls from 0.694 to 0.196 on CEAR and from 0.987 to 0.065 on KITTI, and δ_1 rises from 1.0% to 84.1% and from 0.0% to 95.9%. The uncorrected trajectories are compressed to a small fraction of the travelled distance; the same factors bring them onto the reference paths, although local deviations remain. Depth visualizations on all twelve subsets and camera trajectories of eight GFMs are provided in Appendix F.

5 Conclusion

We presented UniMetric, a training-free framework for recovering metric scale in visual geometry foundation models. Its central insight is that uncertain size priors of familiar objects can jointly constrain a shared scale for scene geometry and camera motion. Theoretical analysis and controlled experiments support the benefit of aggregating multiple uncertain observations. On MetricBench, evaluation against sensor-derived metric references without ground-truth scale alignment demonstrates improvements in suite-averaged depth, reconstruction, and trajectory accuracy across nine models and twelve dataset subsets. Semantic size knowledge is thus a practical source of metric information, which could enrich large-scale visual data with metric supervision for embodied foundation-model pretraining.

Limitations. UniMetric depends on the quality of semantic priors and reconstructed object measurements, including the availability of anchors with measurable heights. A shared scale corrects global scale discrepancies but cannot resolve local geometric distortions or trajectory drift.

AI USE STATEMENT

OpenAI Codex was used to assist with language polishing and LaTeX formatting of the manuscript, including improvements to grammar, phrasing, readability, and formatting consistency. Its use was limited to editorial and formatting tasks and did not extend to the research methodology, experiments, or interpretation of results. The authors take full responsibility for the final manuscript, including its scientific claims and reported results.

ETHICS STATEMENT

This work does not collect new data from human or animal subjects and does not involve human-subject experiments or interaction studies. We evaluate on existing datasets containing synchronized visual and sensor measurements. The source imagery and sensor data remain under their original licenses; our benchmark release will provide MetricBench data, evaluation code, and a machine-readable manifest without redistributing the original recordings. We acknowledge that semantic size priors may vary across object categories, environments, and cultures, and that these variations can affect metric-scale estimates. The corresponding limitations are discussed in the paper and supplementary material.

REPRODUCIBILITY STATEMENT

All experiments use fixed GFM predictions and the evaluation protocol described in the paper and supplementary material. We will release the UniMetric implementation, MetricBench data and subset manifest, evaluation scripts, and the cached VLM-derived priors and segmentation outputs used in the reported experiments. These materials allow the reported tables and figures to be regenerated without additional training or ground-truth scale fitting. The released benchmark data and source components will follow the licenses of the corresponding datasets.

REFERENCES

Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. ZoeDepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023.

Alexey Bochkovskiy, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan Richter, and Vladlen Koltun. Depth Pro: Sharp monocular metric depth in less than a second. In *International Conference on Learning Representations*, 2025.

Alessandro Burzio, Tobias Fischer, Sven Elflein, Qunjie Zhou, Riccardo de Lutio, Jiawei Ren, Jiahui Huang, Shengyu Huang, Marc Pollefeys, Laura Leal-Taixé, Zan Gojcic, and Haithem Turki. Déjà View: Looping transformers for multi-view 3D reconstruction. *arXiv preprint arXiv:2605.30215*, 2026.

ByteDance Seed. Seed1.8 model card: Towards generalized real-world agency. *arXiv preprint arXiv:2603.20633*, 2026.

Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11621–11631, 2020.

Carlos Campos, Richard Elvira, Juan J. Gómez Rodríguez, José M. M. Montiel, and Juan D. Tardós. ORB-SLAM3: An accurate open-source library for visual, visual-inertial, and multimap SLAM. *IEEE Transactions on Robotics*, 37(6):1874–1890, 2021.

Ming-Fang Chang, John Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, et al. Argoverse: 3d tracking and forecasting with rich maps. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8748–8757, 2019.

Lin-Zhuo Chen, Jian Gao, Shangzhan Zhang, Yihang Chen, Ka Leong Cheng, Yipengjing Sun, Liangxiao Hu, Nan Xue, Xing Zhu, Yujun Shen, Yao Yao, and Yinghao Xu. LingBot-Map: Geometric context transformer for streaming 3D reconstruction. *arXiv preprint arXiv:2604.14141*, 2026.

Sili Chen, Hengkai Guo, Shengnan Zhu, Feihu Zhang, Zilong Huang, Jiashi Feng, and Bingyi Kang. Video Depth Anything: Consistent depth estimation for super-long videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22831–22840, 2025.

Dongzhou Cheng, Taoran Yi, Ye Fang, Xingwu Zhang, Fan Feng, Yixuan Li, Gengxiong Zhuang, Rongze Wang, Shuai Yang, Wei Song, Weizhi Xue, Minyan Wu, Jie Gui, Jiaqi Wang, and Tong Wu. In-context robot learning with VLM agents. *arXiv preprint arXiv:2609.19138*, 2026.

David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in Neural Information Processing Systems*, 2014.

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.

Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Kaixuan Wang, Hao Chen, Gang Yu, Chunhua Shen, and Shaojie Shen. Metric3Dv2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10579–10596, 2024.

Jianhao Jiao, Hexiang Wei, Tianshuai Hu, Xiangcheng Hu, Yilong Zhu, Zhijian He, Jin Wu, Jingwen Yu, Xupeng Xie, Huaiyang Huang, et al. Fusionportable: A multi-sensor campus-scene dataset for evaluation of localization and mapping accuracy on diverse platforms. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3851–3856. IEEE, 2022.

Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, et al. Mapanything: Universal feed-forward metric 3d reconstruction. *arXiv preprint arXiv:2509.13414*, 2025.

Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, et al. DROID: A large-scale in-the-wild robot manipulation dataset. *arXiv preprint arXiv:2403.12945*, 2024.

Jack Kiefer. Sequential minimax search for a maximum. *Proceedings of the American Mathematical Society*, 4(3):502–506, 1953.

Yushi Lan, Yihang Luo, Fangzhou Hong, Shangchen Zhou, Honghua Chen, Zhaoyang Lyu, Shuai Yang, Bo Dai, Chen Change Loy, and Xingang Pan. STream3R: Scalable sequential 3D reconstruction with causal transformer. *arXiv preprint arXiv:2508.10893*, 2025.

Rong Li, Yuhao Dong, Tianshuai Hu, Ao Liang, Youquan Liu, Dongyue Lu, Liang Pan, Lingdong Kong, Junwei Liang, and Ziwei Liu. 3EED: Ground everything everywhere in 3d. *arXiv preprint arXiv:2511.01755*, 2025.

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. *arXiv preprint arXiv:2511.10647*, 2025.

Yihang Luo, Shangchen Zhou, Yushi Lan, Xingang Pan, and Chen Change Loy. 4RC: 4D reconstruction via conditional querying anytime and anywhere. In *International Conference on Machine Learning*, 2026.

Roberto Martin-Martin, Mihir Patel, Hamid Rezatofighi, Abhijeet Shenoi, JunYoung Gwak, Eric Frankel, Amir Sadeghian, and Silvio Savarese. Jrdb: A dataset and benchmark of egocentric robot visual perception of humans in built environments. *IEEE transactions on pattern analysis and machine intelligence*, 45(6):6748–6765, 2021.

OpenAI. GPT-6 Astra system card. `https://deploymentsafety.openai.com/gpt-6-astra`, 2026.

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*, pp. 10106–10116, 2024.

Luigi Piccinelli, Christos Sakaridis, Yung-Hsu Yang, Mattia Segu, Siyuan Li, Wim Abbeloos, and Luc Van Gool. UniDepthV2: Universal monocular metric depth estimation made simpler. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Tong Qin, Peiliang Li, and Shaojie Shen. VINS-Mono: A robust and versatile monocular visual-inertial state estimator. *IEEE Transactions on Robotics*, 34(4):1004–1020, 2018.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

Diyari Mohammed Salih, Lingxiang Hu, Naima AitOufroukh-Mammar, and Fabien Bonardi. VIDAR: Visual-inertial dense alignment and reconstruction via a geometric foundation model. *arXiv preprint arXiv:2607.17171*, 2026.

Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of RGB-D SLAM systems. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 573–580, 2012.

Edgar Sucar and Jean-Bernard Hayet. Probabilistic global scale estimation for MonoSLAM based on generic object detection. In *CVPR Workshops*, 2017.

Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2446–2454, 2020.

Rui Tian, Yunzhou Zhang, Delong Zhu, Shiwen Liang, Sonya Coleman, and Dermot Kerr. Accurate and robust scale recovery for monocular visual odometry based on plane geometry. In *IEEE International Conference on Robotics and Automation*, pp. 5296–5302, 2021.

Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(4):376–380, 1991.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5294–5306, 2025a.

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. VGGT-Ω. *arXiv preprint arXiv:2605.15195*, 2026.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. DUSt3R: Geometric 3D vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20697–20709, 2024.

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. *arXiv preprint arXiv:2507.13347*, 2025b.

Bojin Wu and Jing Chen. VGLD: Visually-guided linguistic disambiguation for monocular depth scale recovery. *arXiv preprint arXiv:2505.02704*, 2025.

Tao Xie, Peishan Yang, Yudong Jin, Yingfeng Cai, Wei Yin, Weiqiang Ren, Qian Zhang, Wei Hua, Sida Peng, Xiaoyang Guo, and Xiaowei Zhou. Scal3R: Scalable test-time training for large-scale 3D reconstruction. *arXiv preprint arXiv:2604.08542*, 2026.

Lei Yang, Xinyu Zhang, Jun Li, Chen Wang, Jiaqi Ma, Zhiying Song, Tong Zhao, Ziying Song, Li Wang, Mo Zhou, et al. V2x-radar: A multi-modal dataset with 4d radar for cooperative perception. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth Anything V2. In *Advances in Neural Information Processing Systems*, 2024.

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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9043–9053, 2023.

Haibao Yu, Yizhen Luo, Mao Shu, Yiyi Huo, Zebang Yang, Yifeng Shi, Zhenglong Guo, Hanyu Li, Xing Hu, Jirui Yuan, et al. Dair-v2x: A large-scale dataset for vehicle-infrastructure cooperative 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 21361–21370, 2022.

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. In *NeurIPS*, 2024.

Zhenyu Zhao, Hongyi Jing, Xiawei Liu, Jiageng Mao, Abha Jha, Hanwen Yang, Rong Xue, Sergey Zakharor, Vitor Guizilini, and Yue Wang. Humanoid everyday: A comprehensive robotic dataset for open-world humanoid manipulation. *arXiv preprint arXiv:2510.08807*, 2025.

Shifan Zhu, Zixun Xiong, and Donghyun Kim. Cear: Comprehensive event camera dataset for rapid perception of agile quadruped robots. *IEEE Robotics and Automation Letters*, 9(10):8999–9006, 2024.