DEEP PRIOR LEARNING FOR EMBODIED PERCEPTION

Anonymous authors
Paper under double-blind review

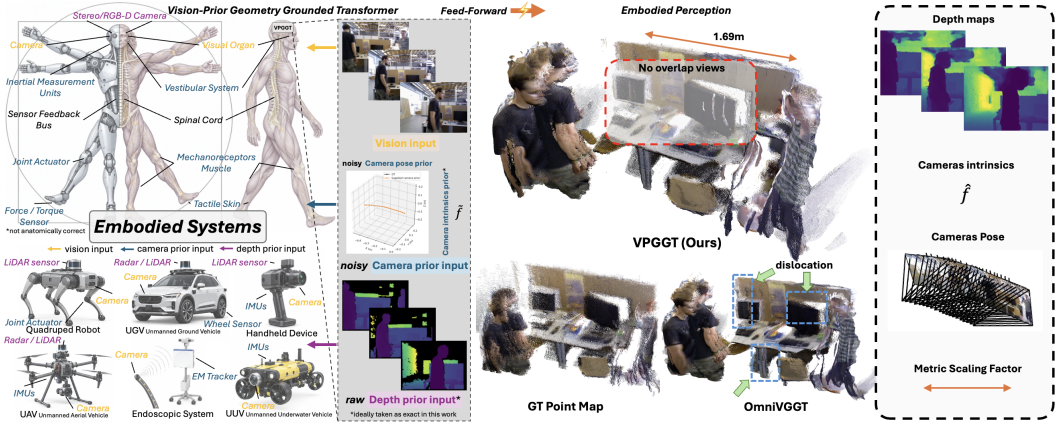


Figure 1: **VPGGT** is an embodied perception foundation model that combines images with optional camera parameters and depth observations to predict camera parameters, depth maps, and pointmaps with a shared metric scale. By leveraging geometric priors supplied by upstream sensing or estimation, the model circumvents the direct processing of raw sensor streams.

ABSTRACT

Embodied systems need geometric perception that exploits available observations beyond images alone. Recent feed-forward 3D models incorporate geometric priors, including camera poses, intrinsics, and depth. However, handling noisy poses, preserving accurate priors, and recovering physical scale require more than simply accepting these inputs. We introduce *Vision-Prior Geometry Grounded Transformer* (VPGGT), a VGGT-based framework that extends OmniVGGT for prior-aware embodied perception. We formulate sensor-motivated pose corruptions from ground-truth trajectories for training and introduce a parameter-free *prior residual connection* (PRC) to mitigate *prior dilution*, where predictions are less accurate than their supplied pose priors. Our noise formulation targets camera poses; supplied intrinsics and depth receive no additional corruption. We further introduce *Metric Global Attention*, which conditions a global scale token on available pose and depth scales and predicts a shared metric scaling factor for the geometric outputs. Experiments across four datasets show that *PRC* improves translation-direction accuracy and joint pose AUC over a matched training baseline when camera priors are provided for all views, under both exact and corrupted poses. These results support explicit prior access during refinement as a useful addition to feature-level conditioning.

1 INTRODUCTION

We study feed-forward 3D perception from a set of images and optional camera or depth priors (Fig. 1). DUSt3R (Wang et al., 2024b) and VGGT (Wang et al., 2025) have advanced image-based reconstruction by predicting multiple geometric attributes within a shared network. Their image-only inputs, however, do not exploit the camera estimates and depth observations often available to embodied systems. Pow3R (Jang et al., 2025), OmniVGGT (Peng et al., 2026), and MapAnything (Keetha et al., 2025) extend feed-forward reconstruction to incorporate such geometric information alongside images, providing a common model for different input configurations.

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.

Several challenges remain when these models are used with observations from embodied platforms. First, camera priors are estimates rather than ground truth, and their errors can accumulate over a trajectory or vary independently between frames. Although previous work has examined incorrect camera inputs, training and evaluation under structured pose errors require further study. Second, a model may produce camera estimates that are less accurate than the supplied priors, even when those priors are exact. We refer to this shortfall as *prior dilution* (Sec. 3.2). Third, embodied systems require geometry in physical units, while normalizing geometric inputs removes their original scale. MapAnything (Keetha et al., 2025) addresses metric reconstruction with a factored representation. Here, we study scale recovery while retaining the camera and dense-prediction heads of a pretrained VGGT-derived model.

In this paper, we introduce *Vision-Prior Geometry Grounded Transformer* (VPGGT), a feed-forward model that combines images with optional camera and depth priors. First, we formulate sensor-motivated pose corruptions from existing ground-truth trajectories for training (Sec. 5).[1] Second, we introduce a parameter-free *Prior Residual Connection* (PRC) that places scale-aligned priors in the camera head's refinement state to mitigate prior dilution (Sec. 4.2). Third, *Metric Global Attention* conditions a global token on available pose and depth scales and predicts a shared metric factor (Sec. 4.3). The model retains the image-only setting when geometric priors are absent and a common interface for further platform-specific training.

Comprehensive experiments and systematic architecture ablations examine the effects of prior quality and availability on camera and depth prediction. Under matched training configurations, *PRC* improves translation-direction accuracy and joint pose AUC on all four datasets when camera priors cover every view, for both exact and corrupted poses. The improvements are largest on TUM RGB-D and C3VD, while results at partial coverage reveal remaining limitations. We study these effects separately from metric-scale recovery to distinguish better camera refinement from better estimation of physical scale.

2 Related Work

Vision-Based 3D Reconstruction Traditional reconstruction systems combine correspondence estimation, camera recovery, and sparse or dense mapping through task-specific modules (Schonberger & Frahm, 2016; Furukawa et al., 2010; Campos et al., 2021). Learned features, matchers, depth networks, and differentiable optimization improve these components(DeTone et al., 2018; Sarlin et al., 2020; Sun et al., 2021; Teed & Deng, 2021; Yao et al., 2018; Wang et al., 2024a). VGGSfM (Wang et al., 2024a), for example, integrates learned correspondences and camera estimation with differentiable bundle adjustment. In contrast, DUSt3R (Wang et al., 2024b) predicts dense pointmaps from image pairs, while VGGT (Wang et al., 2025) jointly predicts camera parameters and dense geometry from multiple views. These models reduce the need for separately designed reconstruction stages, but their image-only interfaces do not use external camera or depth observations.

Embodied Perception Embodied platforms may combine cameras with IMUs, wheel encoders, LiDAR, and depth sensors(Zhao et al., 2025; Frey et al., 2026; Mao & Gu, 2025; Cullen, 2023). Visual–inertial and multi-sensor SLAM fuse their measurements through state estimation and geometric optimization (Qin et al., 2018; Campos et al., 2021). Their observation models, uncertainty estimates, and calibration describe how individual sensors constrain the estimated motion. Prior-conditioned feed-forward models operate at a different level, taking geometric estimates rather than raw sensor measurements. Pow3R (Jang et al., 2025), OmniVGGT (Peng et al., 2026), and MapAnything (Keetha et al., 2025) incorporate camera and depth information into dense reconstruction, with MapAnything also predicting metric scale. Pow3R examines responses to incorrect camera parameters, and MapAnything identifies noise and uncertainty in geometric inputs as a remaining limitation. Our work studies structured pose corruption, preservation of informative camera priors during refinement, and scale conditioning within a pretrained multi-view model.

[1] We corrupt camera poses only. Supplied intrinsics and depth receive no additional corruption in this work.

3 PRELIMINARIES AND PROBLEM DEFINITION

3.1 PRELIMINARIES

Our implementation builds on OmniVGGT (Peng et al., 2026), retaining the backbone and prediction heads inherited from VGGT (Wang et al., 2025). Each image I_i is encoded into patch tokens $\mathbf{t}_i^I$, one camera token $\mathbf{t}_i^g$, and four register tokens $\mathbf{t}_i^R$. The backbone contains $L = 24$ pairs of frame-wise and global self-attention blocks. OmniVGGT's GeoAdapters add normalized depth features to image tokens and camera features to camera tokens, with stochastic prior availability during training. Depth and pointmap heads predict dense geometry, while the camera head refines a nine-dimensional camera estimate over four iterations. We retain these components and add a camera-state prior connection and a metric-scale prediction branch.

3.2 PRIOR DILUTION

We use *prior* to denote a supplied geometric estimate. For a higher-is-better pose metric M, let M_{prior} and M_{out} be the scores of the input prior and model prediction on the same eligible image pairs. We define their gap as $d_M = M_{\text{prior}} - M_{\text{out}}$ and refer to $d_M > 0$ as *prior dilution*. When the prior outperforms the matched no-camera-prior reference M_0, we define

$$\kappa_M = \frac{M_{\text{out}} - M_0}{M_{\text{prior}} - M_0}, \qquad \mathcal{D}_M = 1 - \kappa_M. \quad (1)$$

Here κ_M measures closure of the baseline-to-prior score gap. For exact priors, the angular-score reference is 100. For noisy priors, $\kappa_M > 1$ indicates an output better than the supplied estimate. At partial coverage, prior retention is evaluated only on provided–provided pairs, with all compared scores using the same pair support.

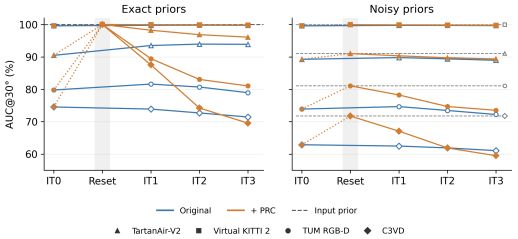


Figure 2: **Prior dilution during camera refinement.** A0 is released OmniVGGT; A1 uses the same checkpoint with *PRC* enabled, without additional training. All views receive camera priors; no depth is supplied. Colors distinguish methods and markers distinguish datasets. Dashed lines show input-prior scores. Reset is a state intervention, not a learned iteration.

Figure 2 compares the released camera head with and without PRC, without additional training. The two variants have identical IT0 scores. Reset recovers the exact-prior score of 100 on all four datasets, but each subsequent refinement update reduces AUC@30°. This effect is strongest on TUM RGB-D and C3VD, where A1 finishes at 81.03 and 69.54, respectively. On C3VD, the latter is also below the original model's 71.44. The same post-reset decline occurs with noisy priors. Thus, direct access to a prior state alone does not ensure that the existing refinement operator preserves its accuracy. We train the model with PRC in Sec. 4.2.

4 METHODOLOGY

VPGGT estimates scene geometry from images and optional camera and depth priors. We describe its inputs and outputs in Sec. 4.1, followed by *Prior Residual Connection* in Sec. 4.2 and *Metric Global Attention* in Sec. 4.3. The training procedure is given in Sec. 4.4.

4.1 OVERVIEW OF VPGGT

Given images $\mathcal{I} = \{I_i\}_{i=1}^N$, the optional inputs comprise camera observations $\mathcal{C} = \{(K_j, G_j^p)\}_{j\in\mathcal{Q}}$ and masked depth observations $\mathcal{D} = \{(D_k^p, M_k)\}_{k\in\mathcal{O}}$, where $\mathcal{Q}, \mathcal{O} \subseteq \{1, \ldots, N\}$. Camera poses use the world-to-camera convention, and a nonempty $\mathcal{Q}$ includes the reference view. The network predicts

$$f_\theta(\mathcal{I}, \mathcal{C}, \mathcal{D}) = \left\{ \hat{s}, (\hat{g}_i, \hat{D}_i, \hat{P}_i)_{i=1}^N \right\}, \quad (2)$$

together with depth and pointmap confidence maps. Here $\hat{g}_i = (\hat{\mathbf{t}}_i, \hat{\mathbf{q}}_i, \hat{\mathbf{f}}_i) \in \mathbb{R}^9$ contains translation, a rotation quaternion, and two fields of view. We retain the inherited camera model and first-view

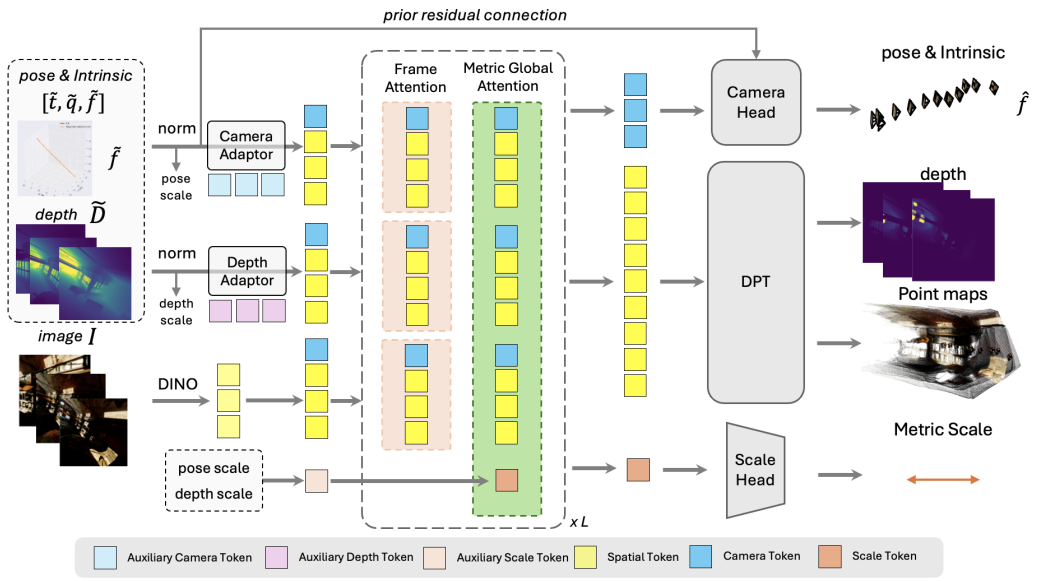


Figure 3: **Overview of VPGGT.** Images and optional camera or depth priors are encoded and processed by a shared alternating-attention backbone. *PRC* places available camera priors in the camera head's refinement state after its initial prediction. A Scale Encoder conditions a global token on available metric-scale statistics, using learned placeholders when they are absent. The camera, depth, and pointmap heads predict geometry, and a scale head supplies a shared factor that converts translations, depths, and points to metric units.

coordinate convention. The shared positive factor $\hat{s}$ produces $\hat{D}_i^{\text{met}} = \hat{s}\hat{D}_i$, $\hat{\mathbf{t}}_i^{\text{met}} = \hat{s}\hat{\mathbf{t}}_i$, and $\hat{P}_i^{\text{met}} = \hat{s}\hat{P}_i$; rotations and intrinsics are unchanged.

4.2 Prior Residual Connection

The camera head already refines its predictions through residual updates. *PRC* changes the reference for these updates from an estimated camera state to the supplied prior wherever one is available (Fig. 4). This follows the residual-learning principle of learning corrections to a reference (He et al., 2016), while retaining the inherited camera-head architecture.

Let $\mathbf{r}_i^0$ denote the initial pre-activation camera estimate for view i. We align the normalized prior to its translation scale, obtaining $\mathbf{a}_i = (\alpha\bar{\mathbf{t}}_i^p, \mathbf{q}_i^p, \mathbf{f}_i^p)$, where α is a translation norm ratio detailed in App. A.6. After recording the initial prediction, *PRC* sets

$$\mathbf{u}_i^0 = \text{sg}\left[m_i\mathbf{a}_i + (1 - m_i)\mathbf{r}_i^0\right], \tag{3}$$

where m_i indicates prior availability and sg denotes stop-gradient. The remaining iterations use the original refinement operator,

$$\mathbf{u}^k = \text{sg}(\mathbf{u}^{k-1}) + \Delta_\theta\big(\mathbf{H}, \text{sg}(\mathbf{u}^{k-1})\big), \quad k = 1, \ldots, J-1, \tag{4}$$

with camera-token features $\mathbf{H}$ and $J = 4$ in our experiments.

For a provided view, the forward prediction thus consists of the aligned prior and the accumulated residual corrections. Zero corrections retain that prior state, while nonzero corrections allow the model to refine an imperfect input. *PRC* adds no parameters and, for fixed network weights, leaves the original

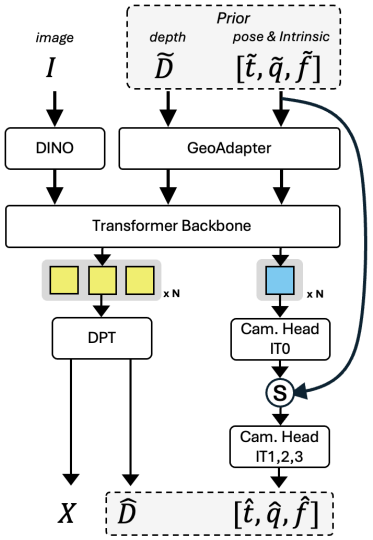


Figure 4: **Prior Residual Connection.** *PRC* retains feature-level conditioning and supplies aligned camera priors as the state from which subsequent residual updates are predicted.

Table 1: **Ablation of auxiliary information guidance.** Depth and pose accuracy with varying prior coverage. **Bold** identifies the better score between the matched-training OmniVGGT-trained and PRC-only models at identical input settings; released-model and input-prior references are not ranked. Blue parentheses report PRC-only gains over OmniVGGT-trained, in percentage points, with full camera priors and no depth input. Depth is median-aligned per frame; RTA and AUC use sign-free translation errors.

(a) TUM RGB-D

Priors			Depth		Camera		
D	C	Pose	AbsRel↓	δ_1 ↑	RRA@5↑	RTA@5↑	AUC@30↑
Input prior reference							
100	100	GT	0.000	100.00	100.00	100.00	100.00
100	100	N	0.000	100.00	100.00	66.00	81.07
OmniVGGT, released							
0	0	–	.078	92.57	99.67	38.00	61.94
100	100	GT	.018	99.20	97.22	67.78	76.66
100	100	N	.018	99.20	97.22	57.11	71.30
OmniVGGT-trained							
0	0	–	.058	**94.64**	96.33	**51.22**	69.17
0	100	GT	**.057**	**94.69**	**97.22**	58.78	71.04
0	100	N	**.057**	**94.70**	97.22	58.89	70.89
100	100	GT	.009	99.76	95.44	55.89	71.86
100	100	N	.009	99.76	95.44	55.56	71.16
VPGGT, PRC only							
0	0	–	.058	94.41	**97.22**	50.33	**69.33**
30	0	–	.031	98.04	97.22	48.56	66.66
50	0	–	.023	98.57	97.22	45.89	66.46
70	0	–	.016	99.11	97.22	49.22	68.87
100	0	–	.009	99.76	96.22	52.33	71.77
0	30	GT	.058	94.44	97.22	51.00	69.29
0	50	GT	.058	94.46	97.22	55.00	70.74
0	70	GT	.058	94.46	96.56	60.11	73.50
						73.78	**80.46**
0	100	GT	.058	94.49	96.89	(+15.00)	(+9.42)
0	30	N	.058	94.44	97.22	49.78	68.64
0	50	N	.058	94.46	97.22	51.11	69.28
0	70	N	.058	94.46	96.56	55.00	70.89
						67.00	**76.37**
0	100	N	.058	94.51	97.22	(+8.11)	(+5.48)
100	100	GT	.009	99.76	**96.56**	**68.44**	**79.08**
100	100	N	.009	99.76	**96.67**	**63.44**	**75.31**

(b) C3VD

Priors			Depth		Camera		
D	C	Pose	AbsRel↓	δ_1 ↑	RRA@5↑	RTA@5↑	AUC@30↑
Input prior reference							
100	100	GT	0.000	100.00	100.00	100.00	100.00
100	100	N	0.000	100.00	100.00	57.73	71.73
OmniVGGT, released							
0	0	–	.511	29.89	87.51	14.67	40.88
100	100	GT	.036	98.57	100.00	45.88	64.70
100	100	N	.036	98.57	100.00	39.56	57.23
OmniVGGT-trained							
0	0	–	.055	97.87	100.00	**29.19**	**61.55**
0	100	GT	.055	97.90	100.00	43.16	69.42
0	100	N	.055	97.90	100.00	42.62	67.73
100	100	GT	.004	99.98	100.00	55.51	76.36
100	100	N	.004	99.98	100.00	54.07	**74.80**
VPGGT, PRC only							
0	0	–	.055	**98.09**	100.00	27.75	61.49
30	0	–	.015	99.80	100.00	26.12	54.92
50	0	–	.010	99.87	100.00	30.72	57.29
70	0	–	.007	99.92	100.00	36.99	62.33
100	0	–	.004	99.98	100.00	54.22	74.59
0	30	GT	.055	98.10	100.00	28.74	60.89
0	50	GT	.055	98.11	100.00	35.26	64.52
0	70	GT	.055	98.11	100.00	46.81	71.01
						76.74	**88.48**
0	100	GT	.055	**98.11**	100.00	(+33.58)	(+19.06)
0	30	N	.055	98.10	100.00	26.02	58.60
0	50	N	.055	98.11	100.00	28.84	58.77
0	70	N	.055	98.11	100.00	34.77	61.50
						55.46	**72.81**
0	100	N	.055	**98.11**	100.00	(+12.84)	(+5.08)
100	100	GT	.004	99.98	100.00	**81.78**	**91.04**
100	100	N	.004	99.98	100.00	**60.40**	74.53

D/C denote provided-view percentages. GT/N specifies pose quality only; supplied depth is ground truth. Input-prior references copy ground-truth depth and score the supplied poses without a network. Angular thresholds are in degrees. Ties at the displayed precision and configurations without a matched-training counterpart are unmarked. All scores except AbsRel are percentages.

computation unchanged when camera priors are absent. It does not constrain the learned corrections to improve or preserve the input accuracy.

4.3 Metric Global Attention

Metric Global Attention retains the physical-scale information removed when normalizing geometric priors (Fig. 3). Let a_C and a_D be the normalization factors computed from the supplied camera and depth observations. For each modality $j \in \{C, D\}$, a shared encoder maps its log scale and coverage ρ_j to a feature. A learned modality-specific placeholder is used when the scale is missing or degenerate,

$$\mathbf{e}_j = \begin{cases} E([\log a_j, \rho_j]), & v_j = 1, \\ \mathbf{p}_j, & v_j = 0, \end{cases} \qquad \mathbf{t}_s^0 = \mathbf{t}_{\text{learn}} + F_0([\mathbf{e}_C, \mathbf{e}_D, v_C, v_D]). \tag{5}$$

Here v_j indicates a valid metric observation. The final projection of F_0 is zero-initialized, and all input scales are expressed in meters. Conditioning is applied once, before the attention blocks.

Following MapAnything (Keetha et al., 2025), the scale token is appended to the multi-view sequence for global attention and bypasses frame attention. Its updated state is carried through all L

global blocks. The final token predicts a positive scale factor,

$$\hat{s} = \exp\big(h(\mathrm{LN}(\mathbf{t}_s^L))\big)\,. \tag{6}$$

This factor scales the predicted depths, translations, and pointmaps without changing their heads. When metric priors are absent, the same branch predicts scale from visual and normalized geometric features.

4.4 Training Procedure

Geometry and scale supervision. We retain the camera, depth, and pointmap objectives of VGGT and OmniVGGT, with $\mathcal{L}_{\mathrm{geo}} = 5\mathcal{L}_{\mathrm{cam}} + \mathcal{L}_{\mathrm{depth}} + \mathcal{L}_{\mathrm{pmap}}$. Camera estimates receive ℓ_1 supervision at all four refinement iterations with iteration-dependent weights. Depth and pointmaps use confidence-weighted and unweighted regression, together with depth-gradient and pointmap-normal terms. These losses supervise the original geometric predictions before metric scaling.

The scale target is estimated from the current predictions and metric ground truth. For depth, we use $\ell_D = \mathrm{median}_{p\in\Omega_D}[\log D_p^* - \log \mathrm{sg}(\hat{D}_p)]$. For cameras, we use $\ell_C = \mathrm{median}_{(i,j)\in\Omega_C}[\log b_{ij}^* - \log \mathrm{sg}(\hat{b}_{ij})]$, where b_{ij} is the distance between camera centers. The valid-modality mean defines ℓ^*, giving $\mathcal{L}_{\mathrm{scale}} = |\log \hat{s} - \mathrm{sg}(\ell^*)|$. We optimize $\mathcal{L}_{\mathrm{geo}} + \mathcal{L}_{\mathrm{scale}}$, omitting the scale term when no valid metric target is available. The target is detached from the geometric predictions, while the scale loss updates the scale branch and shared backbone.

Stochastic prior availability. We follow OmniVGGT's stochastic selection of camera and depth observations. The Scale Encoder uses the same selected inputs and learned placeholders for unavailable scales. Metric ground truth supervises the scale prediction independently of input-prior availability, including image-only samples.

5 Experiments

Datasets. We train VPGGT from the released OmniVGGT checkpoint using TartanAir (Wang et al., 2020), Virtual KITTI 2 (Cabon et al., 2020), TUM RGB-D (Sturm et al., 2012), and C3VD (Bobrow et al., 2023). These datasets include simulated navigation and driving sequences, handheld RGB-D capture, and endoscopic imagery. Each epoch contains approximately 100k sampled training items. We evaluate on TartanAir-V2 and held-out sequences of the other three datasets, using ten-frame windows with stride three and up to ten non-overlapping windows per sequence. Dataset splits, exact sampling budgets, and preprocessing details are provided in the appendix.

Pose-prior corruption. We generate corrupted camera trajectories in metric units before view sampling and normalization. The drift profiles combine sequence-level biases with random motion increments, while the C3VD profile perturbs each pose independently (Table 2). Each training item uses either the clean trajectory or one of four fixed corrupted versions, consistently across its views. Intrinsics and depth receive no additional corruption. The corruption models and parameters are given in App. A.1.

Table 2: Pose-corruption profiles. Drift is applied to relative-motion increments.

Training data	Translation	Rotation
TartanAir	Time-dependent	Time-dependent
TUM RGB-D	Time-dependent	Time-dependent
Virtual KITTI 2	Step-dependent	Time-dependent
C3VD	Independent	Independent

Implementation Details. The backbone contains 24 frame/global attention pairs with token dimension 1024. We retain the original geometric adapters and prediction heads. *PRC* adds no parameters, while the scale branch adds approximately 0.447M parameters. For the prior-utilization study, we train the baseline and its *PRC* variant from the same released OmniVGGT checkpoint for two epochs under matched configurations. This study used three NVIDIA H800 GPUs over approximately seven days. We then initialize VPGGT from the trained *PRC* model and train the scale branch jointly with the network for two further epochs, using four H800 GPUs over approximately three days. Training uses AdamW, cosine learning-rate decay with warmup, bfloat16, gradient checkpointing, and gradient clipping at 1. Additional training and checkpoint settings are given in App. A.3.

Metrics. We report relative rotation accuracy (RRA), sign-free and directed translation-direction accuracy (RTA), and joint pose AUC. Joint errors are the per-pair maxima of rotation and translation errors; AUC uses the inherited one-degree histogram convention. Translation metrics exclude pairs with degenerate ground-truth baselines. Prior gaps d_M and closures κ_M are computed from the input-prior and prediction scores on identical support. Depth is evaluated both after scale alignment and in metric units; alignment settings are stated separately for each table.

5.1 Ablation Study on Auxiliary Information Guidance

We evaluate the effect of *PRC* on held-out TUM RGB-D (Sturm et al., 2012) and C3VD (Bobrow et al., 2023) sequences. The baseline and PRC variant are initialized from the same released OmniVGGT (Peng et al., 2026) checkpoint and trained with matched data, optimization settings, and budgets. We denote them as OmniVGGT-trained and VPGGT (PRC only), respectively. Evaluation uses ten-frame windows at stride three, with images resized to width 518 and heights rounded to a multiple of 14. Camera and depth priors are supplied to the first 30%, 50%, 70%, or 100% of the views.

Table 1 shows that *PRC* improves camera-prior utilization, particularly at full coverage. With exact camera priors and no depth input, it raises RTA@5° from 58.78 to 73.78 on TUM RGB-D and from 43.16 to 76.74 on C3VD, gains of 15.00 and 33.58 percentage points. The corresponding AUC@30° gains are 9.42 and 19.06 points. With noisy poses, RTA improves by 8.11 and 12.84 points, while AUC improves by 5.48 and 5.08 points. The refined output also exceeds the noisy input reference on individual metrics: 67.00 versus 66.00 RTA@5° on TUM RGB-D, and 72.81 versus 71.73 AUC@30° on C3VD. These comparisons indicate improved refinement, but not uniform superiority over the supplied priors.

Depth observations mainly improve depth prediction. At 30% depth coverage, the PRC variant reduces AbsRel from 0.058 to 0.031 on TUM RGB-D and from 0.055 to 0.015 on C3VD. Camera gains are less consistent at partial coverage. For example, at 30% noisy camera coverage, C3VD RTA@5° decreases from 27.65 to 26.02 relative to OmniVGGT-trained. With full noisy camera and depth inputs, its C3VD AUC is also slightly lower, at 74.53 versus 74.80. *PRC* therefore improves prior utilization without guaranteeing preservation under every input configuration. Complete results, stricter thresholds, directed metrics, and depth scores on unprovided views are reported in App. A.5.

5.2 Camera Pose Estimation

We compare full VPGGT with released OmniVGGT on TUM RGB-D, C3VD, and fixed Sewerage and Ocean subsets of TartanAir-V2. Released VGGT is additionally evaluated with RGB-only input on the two TartanAir-V2 subsets. All methods use the same ten-view windows with stride three: 20 windows from two TUM sequences, 45 from five C3VD sequences, and 20 from four sequences in each TartanAir-V2 subset. Sampling and scoring follow App. A.4. Full VPGGT includes PRC and MGA with additional fine-tuning; these comparisons evaluate the resulting systems, not a matched-training architectural change.

All settings include RGB. We abbreviate exact and noisy camera inputs as $C_e = (K, RT_{\text{exact}})$ and $C_n = (K, RT_{\text{noisy}})$. Available priors cover every view. Only poses are corrupted; intrinsics and supplied depth D remain ground truth. Exact-camera settings are therefore oracle-conditioned tests.

VPGGT improves translation accuracy and joint pose AUC on TUM/C3VD under all three settings, although TUM rotation accuracy decreases (Table 3). The ranking largely reverses on TartanAir-V2: OmniVGGT has higher AUC in 11 of 12 settings and lower Sim(3)-aligned ATE in all 12 (Table 4). VPGGT's sole AUC lead is on Sewerage with C_n (90.60 versus 90.24). Neither comparison establishes reliable prior improvement: on Ocean with C_n, VPGGT obtains 85.81 AUC, below the input prior's 91.66.

5.3 Depth Estimation

We evaluate camera Z-depth from the same model executions, using per-frame median alignment and shared ground-truth-valid support, then average frame-level scores. VPGGT reduces AbsRel

Table 3: **Pose and depth on TUM RGB-D and C3VD.** R, T, and A denote RRA@5°, RTA@5°, and AUC@30° in percent. AR is median-aligned depth AbsRel; δ_1 is depth accuracy at threshold 1.25, in percent. ATE uses Sim(3) alignment, in meters for TUM and millimeters for C3VD. Bold marks the unique best network within each input setting; ties at displayed precision are unmarked.

	TUM RGB-D						C3VD					
Method	$R\uparrow$	$T\uparrow$	$A\uparrow$	AR↓	$\delta_1\uparrow$	ATE↓	$R\uparrow$	$T\uparrow$	$A\uparrow$	AR↓	$\delta_1\uparrow$	ATE↓
RGB only												
OmniVGGT	**99.67**	38.00	61.94	0.0782	92.57	0.00554	87.51	14.67	40.88	0.5114	29.89	0.1962
VPGGT	96.22	**50.44**	**69.38**	**0.0607**	**94.21**	**0.00444**	**100.00**	**27.41**	**62.58**	**0.0563**	**97.86**	**0.0759**
w/ C_e												
OmniVGGT	**100.00**	70.78	78.95	0.0779	92.18	0.00243	98.62	45.23	71.44	0.4848	30.96	0.0769
VPGGT	96.89	**72.56**	**80.39**	**0.0604**	**94.29**	**0.00223**	**100.00**	**77.48**	**88.14**	**0.0563**	**97.87**	**0.0210**
w/ C_n												
OmniVGGT	**100.00**	58.89	72.21	0.0778	92.23	0.00263	98.67	36.35	61.05	0.4850	30.95	0.0873
VPGGT	97.11	**67.44**	**76.50**	**0.0604**	**94.29**	**0.00254**	**100.00**	**56.05**	**72.95**	**0.0563**	**97.87**	**0.0392**

Input-prior references: exact $T/A = 100/100$; noisy $T/A = 66.00/81.07$ on TUM and $57.73/71.73$ on C3VD. Input-prior $R = 100$ in both camera settings.

Table 4: **Pose and depth on TartanAir-V2 subsets.** Metrics and boldface follow Table 3; ATE is in meters. Sewerage uses P001/P002 and Ocean uses P000/P001 in both Easy and Hard, with five windows per sequence. These results cover the selected subsets, not the full benchmark.

	Sewerage						Ocean					
Method	$R\uparrow$	$T\uparrow$	$A\uparrow$	AR↓	$\delta_1\uparrow$	ATE↓	$R\uparrow$	$T\uparrow$	$A\uparrow$	AR↓	$\delta_1\uparrow$	ATE↓
RGB only												
VGGT	**94.67**	67.44	84.51	**0.0867**	**89.71**	0.0892	**91.22**	42.78	75.51	0.0868	92.95	0.1786
OmniVGGT	88.56	**69.11**	**84.63**	0.0940	87.92	**0.0804**	90.67	**45.00**	**78.14**	**0.0777**	94.89	**0.1591**
VPGGT	83.33	53.56	78.84	0.1009	86.96	0.1148	90.33	42.78	75.68	0.0804	**95.30**	0.2079
w/ C_e												
OmniVGGT	**100.00**	**96.67**	**95.23**	**0.0913**	**88.15**	**0.0408**	100.00	**93.11**	**93.74**	**0.0776**	94.88	**0.0557**
VPGGT	97.22	90.78	91.90	0.0977	87.25	0.0653	100.00	71.00	86.11	0.0791	**95.57**	0.1113
w/ C_n												
OmniVGGT	**100.00**	85.00	90.24	**0.0913**	**88.14**	**0.0472**	**100.00**	**77.11**	**89.37**	**0.0776**	94.88	**0.0636**
VPGGT	97.22	**87.33**	**90.60**	0.0978	87.25	0.0690	99.89	70.56	85.81	0.0791	**95.57**	0.1127
w/ D												
OmniVGGT	86.00	**56.67**	**78.42**	0.0218	99.30	**0.1147**	**92.56**	**42.11**	**77.48**	0.0396	97.26	**0.1587**
VPGGT	**88.33**	42.00	75.79	**0.0105**	**99.64**	0.1519	92.11	36.56	74.91	**0.0258**	**98.40**	0.1870
w/ C_e, D												
OmniVGGT	**97.78**	**94.22**	**95.01**	0.0210	99.30	**0.0423**	**100.00**	**93.78**	**94.36**	0.0394	97.29	**0.0519**
VPGGT	96.89	84.56	88.87	**0.0098**	**99.64**	0.0754	99.44	81.89	88.31	**0.0258**	**98.41**	0.0954
w/ C_n, D												
OmniVGGT	**97.56**	**84.22**	**89.91**	0.0210	99.30	**0.0526**	**100.00**	79.67	**89.98**	0.0394	97.29	**0.0598**
VPGGT	97.11	83.11	88.24	**0.0098**	**99.64**	0.0760	99.67	**81.33**	88.13	**0.0258**	**98.41**	0.0961

Noisy input-prior T/A: Sewerage $86.78/90.90$; Ocean $82.78/91.66$, with or without depth. Exact-prior $T/A = 100/100$; input-prior $R = 100$.

on TUM/C3VD in all three settings. Without depth inputs on TartanAir-V2, OmniVGGT instead has lower AbsRel, while VPGGT retains higher δ_1 on Ocean. With depth inputs, VPGGT improves both metrics in all six TartanAir-V2 comparisons. These full-coverage settings assess preservation of observed depth, not prediction of unavailable depth: copying the supplied ground-truth depth would give zero AbsRel and 100% δ_1 on the same valid pixels.

5.4 Metric Reconstruction and Scale

VPGGT's predicted window scale is applied once to depth and camera translations. Metric AbsRel uses no median alignment, and ATE-SE(3) fits rotation and translation only. Following the point-norm scale statistic of MapAnything (Keetha et al., 2025), we compute

$$E^X_{\text{scale}} = 100 \left| \frac{a(\hat{X}^{\text{met}})}{a(X^*)} - 1 \right|, \qquad a(X) = \frac{1}{|\Omega|} \sum_{(i,p)\in\Omega} \|X_{i\to 0}(p)\|_2. \tag{7}$$

Table 5: **Metric reconstruction of full VPGGT.** Entries are metric AbsRel/E^X_{scale} (%)/ATE-SE(3). No post-hoc scale fitting to evaluation ground truth is used. ATE is in millimeters for C3VD and meters for the other datasets. All three metrics are lower-is-better; different input settings are not ranked against one another.

Input	TUM RGB-D	C3VD
RGB	0.1003/5.07/0.02226	0.0974/7.67/0.204
C_e	0.0977/4.88/0.02058	0.0900/6.80/0.173
C_n	0.0976/4.78/0.01811	0.0899/6.72/0.176

Input	Sewerage	Ocean
RGB	0.1689/19.92/0.3626	0.1605/9.08/0.3522
C_e	0.1705/18.08/0.3482	0.1265/5.56/0.2349
C_n	0.1706/18.13/0.3498	0.1264/5.57/0.2346
D	0.1299/14.20/0.3329	0.0891/5.39/0.3589
C_e, D	0.1258/12.05/0.2726	0.0623/3.85/0.2312
C_n, D	0.1258/12.07/0.2734	0.0623/3.86/0.2305

Each reconstruction is expressed in its own first-camera frame, with common ground-truth-valid support Ω. Scale and trajectory errors are averaged over windows. Because point radii also depend on shape and camera estimates, this statistic does not isolate scale-head accuracy.

RGB-only scale error ranges from 5.07% on TUM to 19.92% on Sewerage (Table 5). Depth conditioning does not fully preserve physical scale: on Sewerage with C_e, D, aligned AbsRel is 0.0098, but metric AbsRel remains 0.1258. Accurate aligned depth therefore does not imply accurate metric reconstruction.

Limitations. The evaluation is restricted to small, fixed sequence subsets, so the reported gains on TUM/C3VD do not establish uniform cross-domain improvements. Pose corruption is synthetic, while intrinsics and supplied depth remain clean. The metric results only cover the full model, lacking a matched-training MGA ablation and an external metric baseline. A shared window scale cannot remove view-dependent distortions or enforce consistency across independently reconstructed windows. Single-frame inference, scene-level surface accuracy and completeness, clinical use, and closed-loop control are not evaluated.

6 Conclusion

We present VPGGT, a feed-forward model for geometric perception from images and optional camera and depth priors. The model extends an OmniVGGT backbone with sensor-motivated pose-prior training, a parameter-free *Prior Residual Connection*, and a prior-conditioned global scale token. *PRC* gives the camera head direct access to supplied poses during refinement, while the scale branch converts the geometric outputs to metric units. Matched-training experiments show improved translation-direction accuracy and joint pose AUC under full camera-prior coverage, including corrupted poses. The results also show that prior availability alone does not ensure prior preservation, particularly at partial coverage. VPGGT provides a common model for incorporating geometric estimates into feed-forward perception and for further adaptation to the observations available on an embodied platform.

AI Use Statement

We used AI tools to assist with manuscript drafting and editing, discussion of evaluation protocols, and the drafting and debugging of inference, visualization, and evaluation code. Reported numerical results are obtained from model executions and evaluation scripts, not generated by the language model. The authors are responsible for the experimental decisions, verification of the results and references, and the final manuscript.

Ethics Statement

This work studies geometric perception using existing datasets and synthetically corrupted camera estimates. Its experiments do not establish the safety of deploying the model for autonomous control or medical procedures. Errors in camera pose, depth, or metric scale could cause harm if used without independent validation and appropriate safeguards. The endoscopic results should therefore be interpreted as benchmark evaluations, not clinical validation. Use of reconstruction systems must

also account for privacy, dataset permissions, and the possibility of recovering sensitive information from captured environments.

Reproducibility Statement

Section 4 describes the model, and App. A.6 specifies the PRC implementation. Training and checkpoint settings are given in App. A.3. The pose-corruption process is described in App. A.1, and App. A.4 defines window sampling, input availability, validity masks, alignment, and aggregation. Complete matched-training results are reported in App. A.5. Experiment records retain checkpoint identifiers, source-window identities, input conditions, and evaluation settings.

References

Taylor L Bobrow, Mayank Golhar, Rohan Vijayan, Venkata S Akshintala, Juan R Garcia, and Nicholas J Durr. Colonoscopy 3d video dataset with paired depth from 2d-3d registration. *Medical image analysis*, 90:102956, 2023.

Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2, 2020.

Carlos Campos, Richard Elvira, Juan J Gómez Rodríguez, José MM 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.

Kathleen E Cullen. Internal models of self-motion: neural computations by the vestibular cerebellum. *Trends in neurosciences*, 46(11):986–1002, 2023.

Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 224–236, 2018.

Jonas Frey, Matías Mattamala, Hae-Won Park, Mayank Mittal, Georg Martius, Maike Osborne, Robert Sparrow, and Marco Hutter. Advances, challenges, and opportunities for legged robots. *Science Robotics*, 11(116):eaee0787, 2026.

Yasutaka Furukawa, Brian Curless, Steven M Seitz, and Richard Szeliski. Towards internet-scale multi-view stereo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pp. 1434–1441. IEEE, 2010.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Wonbong Jang, Philippe Weinzaepfel, Vincent Leroy, Lourdes Agapito, and Jerome Revaud. Pow3r: Empowering unconstrained 3d reconstruction with camera and scene priors. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 1071–1081, 2025.

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.

Dun Mao and Yong Gu. Multisensory coding of self-motion and its contribution to navigation. *Nature Reviews Neuroscience*, 26(11):715–732, 2025.

Haosong Peng, Hao Li, Yalun Dai, Yushi Lan, Yihang Luo, Tianyu Qi, Zhengshen Zhang, Yufeng Zhan, Junfei Zhang, Wenchao Xu, et al. Omnivggt: Omni-modality driven visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 36485–36497, 2026.

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.

Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4938–4947, 2020.

Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4104–4113, 2016.

Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of rgb-d slam systems. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pp. 573–580. IEEE, 2012.

Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8922–8931, 2021.

Zachary Teed and Jia Deng. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. *Advances in neural information processing systems*, 34:16558–16569, 2021.

Jianyuan Wang, Nikita Karaev, Christian Rupprecht, and David Novotny. Vggsfm: Visual geometry grounded deep structure from motion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 21686–21697, 2024a.

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

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, 2024b.

Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. Tartanair: A dataset to push the limits of visual slam. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4909–4916. IEEE, 2020.

Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 767–783, 2018.

Shibo Zhao, Sifan Zhou, Yuchen Zhang, Ji Zhang, Chen Wang, Wenshan Wang, and Sebastian Scherer. Resilient odometry via hierarchical adaptation. *Science Robotics*, 10(109):eadv1818, 2025.

A Appendix

A.1 Pose-Prior Corruption

We corrupt complete metric camera-to-world trajectories before selecting training views or normalizing geometric inputs. The profiles model errors in supplied camera estimates rather than raw sensor measurements. For each cached training realization, a sequence-level multiplier $g \sim \mathrm{LogU}(0.6, 1.8)$ scales the noise parameters. Training selects either the clean trajectory or one of four cached corrupted versions, using the same source for all views in a sample.

Accumulated drift. Let $T_k = [R_k, \mathbf{c}_k]$ denote a camera-to-world pose and $\Delta T_k = T_{k-1}^{-1} T_k$ its relative motion. Both drift profiles start from $T_0^p = T_0$ and accumulate perturbed increments,

$$T_k^p = T_{k-1}^p \begin{bmatrix} \delta R_k \Delta R_k & \Delta \mathbf{t}_k + \delta \mathbf{t}_k \\ \mathbf{0}^\top & 1 \end{bmatrix}. \tag{8}$$

Only the first pose of the complete trajectory is fixed. The first pose of a later evaluation window need not be uncorrupted.

Table 6: Base pose-corruption parameters before severity scaling. Time-scaled white increments use a reference interval of 0.1 s.

Parameter	Unit	TartanAir	TUM	VKITTI2	C3VD
Heading-rate bias std.	rad/s	4.37×10^{-3}	$\pi/10800$	$\pi/36000$	–
Heading white std.	deg	0.1100	0.08	0.03	–
Other-axis white std.	deg	0.02	0.03	0.02	–
Velocity-bias std.	m/s	0.076	0.0045	–	–
Translation white std.	m	0.001	0.0009	–	–
Translation scale-bias std.	–	–	–	0.01	–
Step-noise coefficient	–	–	–	0.01	–
Conditional translation floor	mm	–	–	0.5	–
Independent position std.	mm/axis	–	–	–	0.03
Independent Euler-angle std.	deg/axis	–	–	–	0.05

A sequence-constant heading-rate bias contributes $b_\psi \Delta t_k$. Angular white increments are scaled by $q_k = \sqrt{\Delta t_k / t_0}$, where $t_0 = 0.1\,\text{s}$,

$$\delta\psi_k = b_\psi \Delta t_k + \sigma_\psi q_k \epsilon_{\psi,k}, \qquad \begin{bmatrix}\delta\theta_k \\ \delta\phi_k\end{bmatrix} = \sigma_{rp} q_k \begin{bmatrix}\epsilon_{\theta,k} \\ \epsilon_{\phi,k}\end{bmatrix}. \tag{9}$$

The random variables are standard normal. All rotational components accumulate through trajectory composition; no gravity-aiding correction bounds the other-axis errors.

The time-dependent translation profile uses a sequence-constant velocity bias and white increments,

$$\delta\mathbf{t}_k = \mathbf{b}_v \Delta t_k + \sigma_t q_k \boldsymbol{\epsilon}_k. \tag{10}$$

The step-dependent profile instead uses

$$\delta\mathbf{t}_k = b_s \Delta\mathbf{t}_k + (\sigma_f d_k + \sigma_0 \mathbf{1}[d_k > 5\sigma_0])\,\boldsymbol{\epsilon}_k, \qquad d_k = \|\Delta\mathbf{t}_k\|_2. \tag{11}$$

Here b_s is sampled once per trajectory. The translation perturbation vanishes at an exactly stationary step, but angular drift can continue. The final trajectory error is not generally linear in time or distance because rotational errors also affect subsequent translations.

Rotation axes. When gravity metadata is unavailable, the implementation estimates a heading axis from the mean camera-up direction. This direction is used when the norm of the mean exceeds 0.5; otherwise, perturbations use a camera-axis Euler construction. The procedure is a coordinate heuristic rather than a model of gravity observability.

Independent perturbations. The tracker-style profile independently perturbs each absolute pose,

$$R_k^p = R_{\mathrm{ZYX}}(\boldsymbol{\epsilon}_{R,k}) R_k, \qquad \mathbf{c}_k^p = \mathbf{c}_k + \boldsymbol{\epsilon}_{p,k}. \tag{12}$$

The angle and position components are Gaussian with the standard deviations in Table 6. The first pose is also perturbed. The perturbations are independent conditional on the shared sequence-level severity. This is an idealized operating regime, not a calibration of a particular tracking device.

Evaluation corruption. For the pose-corruption conditions, evaluation generates a complete corrupted trajectory for each scene and then extracts the selected windows. It does not load a fifth cached training realization. The random generator is initialized from the evaluation seed, a CRC32 key of the canonical scene path, and a fixed stream identifier. The main evaluation uses a profile multiplier of 1, without additional random severity sampling. Time steps are derived from the declared nominal frame rate. Camera intrinsics and depth receive no additional corruption.

Reproducibility. Corruption profiles, source-code versions, random seeds, and cached training realizations are recorded separately. Parameters describing the current implementation must match those used to generate the training inputs. Existing caches are not regenerated as part of result reporting.

Algorithm 1 Accumulated pose-prior corruption

Require: Camera-to-world poses $\{T_k\}$, time steps Δt_k, profile parameters, severity g
Scale the profile parameters by g
Sample $b_\psi \sim \mathcal{N}(0, \sigma_{b\psi}^2)$
if time-dependent translation **then**
    Sample $\mathbf{b}_v \sim \mathcal{N}(\mathbf{0}, \sigma_{bv}^2 I_3)$
else
    Sample $b_s \sim \mathcal{N}(0, \sigma_{bs}^2)$
end if
$T_0^p \leftarrow T_0$
for $k = 1, \ldots, N-1$ **do**
    $\Delta T_k \leftarrow T_{k-1}^{-1} T_k$
    Sample angular increments using Eq. 9
    Construct δR_k using the profile's rotation axes
    if time-dependent translation **then**
        Sample $\delta \mathbf{t}_k$ using Eq. 10
    else
        Sample $\delta \mathbf{t}_k$ using Eq. 11
    end if
    Compose T_k^p using Eq. 8
end for
return $\{T_k^p\}$

Algorithm 2 Independent pose-prior corruption

Require: Camera-to-world poses $\{T_k\}$, profile parameters, severity g
$\sigma_R \leftarrow g\sigma_R$; $\sigma_p \leftarrow g\sigma_p$
for $k = 0, \ldots, N-1$ **do**
    Sample $\boldsymbol{\epsilon}_{R,k} \sim \mathcal{N}(\mathbf{0}, \sigma_R^2 I_3)$
    Sample $\boldsymbol{\epsilon}_{p,k} \sim \mathcal{N}(\mathbf{0}, \sigma_p^2 I_3)$
    Compute T_k^p using Eq. 12
end for
return $\{T_k^p\}$

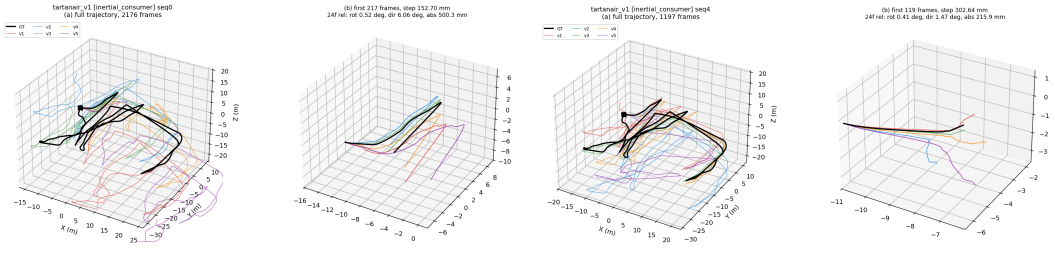
A.2 Visualization of Pose-Prior Corruptions

Figure 5 presents two training-trajectory examples from each of the four datasets. Within each panel, the left plot shows the complete trajectory, and the right plot shows a selected segment from the same example. The segment is extracted from the displayed corruption realizations, rather than generated using new noise samples. The global and local views illustrate the spatial structure of accumulated drift and independent pose perturbations.

A.3 Training and Checkpoint Details

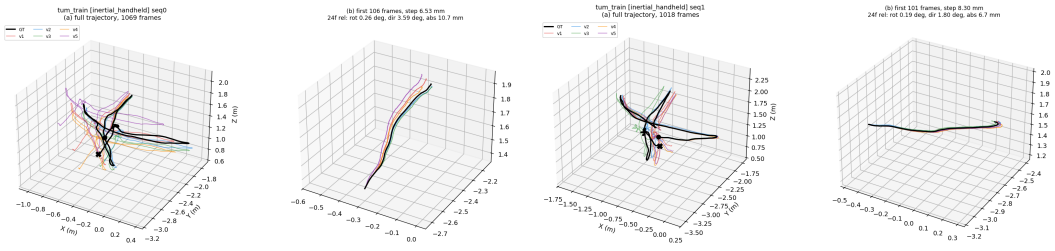
The matched prior-utilization comparison trains the released OmniVGGT model with and without PRC for two epochs. The final metric model is initialized from the trained PRC variant and receives two additional epochs. It therefore has a larger fine-tuning budget than either model in the PRC ablation.

The scale stage uses four GPUs, 24 nominal image slots per GPU, and three gradient-accumulation steps. Learning rates are 10^{-5} for the backbone and geometric heads, 5×10^{-6} for the patch embedding, and 2×10^{-5} for the scale branch. Training uses 500 warmup steps and cosine decay to 0.1 times the initial learning rate. The scale-loss weight is 1, and depth-based scale supervision requires at least 100 valid pixels.



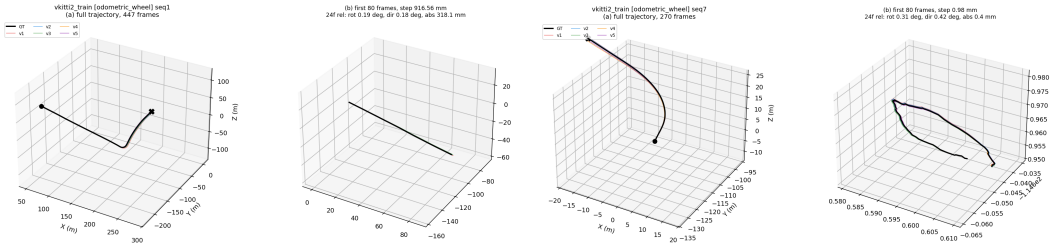
(a) TartanAir-V1: example 1

(b) TartanAir-V1: example 2



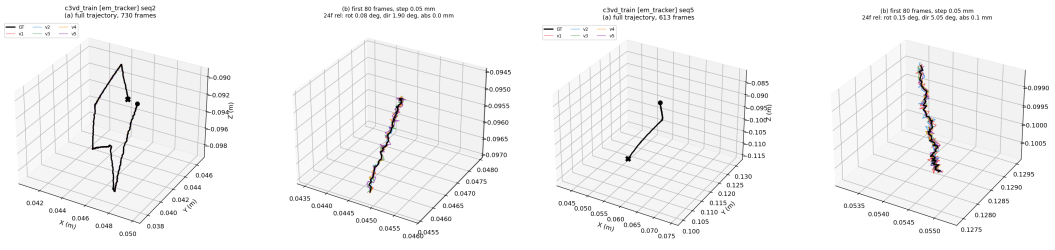
(c) TUM RGB-D: example 1

(d) TUM RGB-D: example 2



(e) Virtual KITTI 2: example 1

(f) Virtual KITTI 2: example 2



(g) C3VD: example 1

(h) C3VD: example 2

Figure 5: **Pose-prior corruption in the training data.** Two examples are shown for each dataset. Within each panel, the left plot displays the complete trajectory, while the right plot displays a selected segment of the same ground-truth and corrupted trajectories. Colors identify the ground truth and corruption realizations, as indicated by the plot legends. Ground-truth and corrupted poses share a coordinate system, without independently fitted trajectory or scale alignment. Spatial units and segment information are provided in the plots.

The data mixture contains 95,004 sampled training items per epoch. This is a sampling budget, not a count of unique images. Split manifests, checkpoint hashes, effective run configurations, and evaluation-window identities are retained with the experiment records.

A.4 Evaluation Protocol

Windows and input selection. Evaluation uses ten-frame windows with a frame stride of three. Each window spans 28 frames of the source sequence. Up to ten windows are distributed across each eligible sequence, without overlap between their temporal spans. Sequences shorter than the required span are excluded according to ground-truth availability. Images are resized to width 518, with heights rounded to a multiple of 14; intrinsics are adjusted accordingly. For a coverage ratio ρ, priors are supplied to the first $\mathrm{round}(10\rho)$ views. Camera and depth coverage are controlled separately. The camera input consists of intrinsics and poses together.

Relative-pose metrics. We evaluate every unordered image pair within a window. Let e_R denote relative-rotation error and e_T^{dir} the angle between the predicted and ground-truth relative translation directions. The sign-free error is $e_T^{\mathrm{sf}} = \min(e_T^{\mathrm{dir}}, 180^\circ - e_T^{\mathrm{dir}})$. We report RRA and RTA at 2°, 5°, and 15°, using strict threshold comparisons. Joint pose error is $e_{\mathrm{pose}} = \max(e_R, e_T)$. AUC@$15^\circ$ and AUC@30° use the cumulative mean of the one-degree error histogram, matching the evaluator. Unless marked "dir", RTA and AUC use sign-free translation errors.

Ground-truth baselines shorter than 10^{-6} in the processed metric coordinates are excluded from translation and joint-pose metrics, but remain eligible for rotation metrics. An exactly zero predicted baseline against a valid ground-truth baseline receives a translation error of 90°. Non-finite pose predictions cause evaluation failure rather than being removed from the scoring support. Pose scores pool eligible pairs across evaluation windows.

Depth metrics. Depth validity is determined from ground truth. Pixels must have finite depth greater than 10^{-4} and satisfy the dataset-specific upper-depth limit, when one is configured. Frames with fewer than 100 valid pixels are excluded. For the scale-aligned results, each predicted depth map is multiplied by the ratio of the ground-truth and predicted depth medians on these pixels. We then compute AbsRel and δ_1, using the threshold 1.25. The final scores average eligible frame-level metrics. Predicted invalid depths on valid ground-truth pixels cause a failure rather than a reduction in evaluation support. We also report scores restricted to views without supplied depth.

Trajectory alignment. The ATE values accompanying the prior-utilization experiments use Sim(3) alignment and measure trajectory shape up to scale. They are not evidence of metric-scale accuracy. Metric evaluation instead uses rigid SE(3) alignment, without fitting a scale to test ground truth.

Metric reconstruction. Metric depth is evaluated on the same ground-truth-valid support as aligned depth, but no median alignment is applied. ATE-SE(3) is the root-mean-square camera-center error after fitting rotation and translation only. The reported value averages window-level errors. C3VD trajectory errors are converted from meters to millimeters only for display.

For the point-norm scale statistic, each ground-truth and predicted reconstruction is expressed in its respective first-camera frame. We average point distances over the common valid support and compute their absolute relative difference. VPGGT points are obtained from predicted Z-depth, predicted intrinsics, and predicted poses. MapAnything uses its predicted local geometry and camera poses. This evaluation adopts the scale statistic, but not the datasets or view-sampling protocol, of the original MapAnything benchmark.

Supplied-prior scale baseline. For the PRC-only scale baseline, let $\hat{b}_{ij}$ be a predicted camera-center distance and b_{ij}^p its supplied-prior counterpart. A shared positive scale is obtained as

$$s_p = \exp\left(\mathrm{median}_{(i,j)\in\Omega_p}[\log b_{ij}^p - \log \hat{b}_{ij}]\right). \tag{13}$$

The support uses nondegenerate supplied-camera baselines. The fitted scale is applied to all translations and depth maps in the window. Evaluation ground truth is not used in this fit.

Table 7: Dataset-specific evaluation settings. Depth limits apply to ground-truth scoring support. The PNG divisor is used only for encoded PNG depth.

Dataset	Nominal FPS	Maximum depth (m)	PNG divisor
TartanAir-V2	10	500	–
Virtual KITTI 2	10	655	100
TUM RGB-D	30	10	–
C3VD	30	0.15	–

Implementation and aggregation. The external comparison reuses the data loading, window selection, pose corruption, angular metrics, and aligned-depth functions of the prior-utilization evaluator. Additional metric scores are stored separately. Pose metrics pool image-pair errors; depth metrics average eligible frame scores; metric scale and trajectory metrics average window scores. Prediction confidence does not alter scoring support. Incomplete executions are not treated as completed benchmark rows.

Prior-retention support. AA, AU, and UU denote pairs with two, one, and zero supplied camera priors, respectively. The input prior is scored only on AA pairs. At partial coverage, its scores are compared with the model's AA scores, not with all-pair prediction scores. The supplied prior is a reference, not an upper bound, when camera poses are corrupted.

Gap and closure. For a higher-is-better score M, the signed prior gap is $d_M = M_{\text{prior}} - M_{\text{out}}$. A positive value indicates prior dilution. When $M_{\text{prior}} > M_0$, the closure is $\kappa_M = (M_{\text{out}} - M_0)/(M_{\text{prior}} - M_0)$, where M_0 uses the same model, windows, depth inputs, and evaluation pairs without the camera bundle. Because the camera bundle contains both intrinsics and poses, this reference removes both. Closure is undefined when the prior does not outperform the reference. For partial coverage, closure additionally requires reaggregating the no-camera predictions on the same AA pairs.

A.5 Complete Matched-Training Results

C1 denotes OmniVGGT-trained and C2 denotes VPGGT (PRC only). Both are initialized from the same released checkpoint and trained for two epochs with matched data, optimization settings, and training budgets. Their architectural difference is PRC. Neither model contains the metric-scale branch. Matched configurations do not imply identical realized stochastic training samples at every optimization step.

At full camera-prior coverage without depth input, C2 improves RTA@5° and AUC@30° over C1 on all four datasets for both exact and corrupted poses. For exact priors, the RTA gains are 3.33, 0.10, 15.00, and 33.58 percentage points on TartanAir-V2, Virtual KITTI 2, TUM RGB-D, and C3VD, respectively. The corresponding noisy-prior gains are 2.65, 0.13, 8.11, and 12.84 points.

The improvements depend on input coverage and the evaluated metric. At 30% exact camera coverage on TartanAir-V2, RTA decreases from 46.28 to 43.81. At 30% noisy camera coverage on C3VD, it decreases from 27.65 to 26.02. Rotation accuracy does not improve uniformly. These exceptions are retained in the complete tables.

Input-prior scores further distinguish improvement over a learned baseline from improvement over the supplied observations. With full noisy camera priors and no depth input, C2 exceeds the TUM RGB-D prior in RTA@5° but not in AUC@30°. On C3VD, it exceeds the prior in AUC but not in RTA@5°. Consequently, prior retention and refinement are reported separately for each metric.

A.6 PRC Implementation Details

PRC operates on the camera head's pre-activation nine-dimensional state. For availability indicators m_i, the translation norm ratio is

$$\alpha = \frac{\sum_i m_i \|\mathbf{r}^0_{i,t}\|_2}{\sum_i m_i \|\bar{\mathbf{t}}^p_i\|_2}. \tag{14}$$

Table 8: **Complete matched-training results on TartanAir-V2.** C1: OmniVGGT-trained; C2: VPGGT (PRC only). Bold marks the better displayed score within each matched pair. Angular thresholds are in degrees; angular scores and δ_1 are percentages.

(a) Sign-free pose metrics

D	C	Pose	Model	R@2↑	R@5↑	R@15↑	T@2↑	T@5↑	T@15↑	A@15↑	A@30↑
0	0	–	C1	58.52	**88.69**	**96.46**	10.49	**46.48**	**88.31**	**58.07**	**75.01**
0	0	–	C2	**58.97**	88.37	96.12	**10.69**	46.35	87.78	57.77	74.63
30	0	–	C1	**60.92**	**88.91**	**96.28**	9.30	**42.18**	**86.72**	**55.78**	**73.39**
30	0	–	C2	60.56	88.48	96.24	**9.32**	41.98	86.39	55.50	73.12
50	0	–	C1	**62.41**	**89.00**	**96.03**	**9.04**	**41.86**	**86.26**	**55.49**	**73.01**
50	0	–	C2	62.32	88.78	95.96	8.99	41.66	85.77	55.12	72.77
70	0	–	C1	**64.55**	**89.48**	96.00	**8.87**	**41.28**	**86.35**	**55.42**	**72.90**
70	0	–	C2	64.30	89.34	**96.11**	8.48	40.72	85.92	55.07	72.67
100	0	–	C1	67.02	90.17	**96.12**	**8.33**	**40.87**	86.62	**55.53**	73.01
100	0	–	C2	**67.26**	**90.32**	96.03	8.14	40.40	**86.64**	55.34	**73.04**
0	30	GT	C1	**62.22**	**89.37**	**96.46**	**12.13**	**46.28**	**84.95**	**55.86**	**72.91**
0	30	GT	C2	62.05	89.13	96.28	11.46	43.81	82.52	53.62	70.92
0	50	GT	C1	**64.34**	**89.81**	**96.61**	**13.86**	**49.60**	**85.73**	**57.51**	**74.03**
0	50	GT	C2	64.31	89.48	96.59	13.72	46.96	83.26	55.14	72.00
0	70	GT	C1	69.40	**91.25**	**97.14**	17.68	**57.03**	**87.90**	**61.93**	**76.87**
0	70	GT	C2	**70.03**	91.03	97.02	**18.79**	55.55	86.12	60.54	75.65
0	100	GT	C1	85.50	95.83	98.36	25.89	73.32	94.21	73.09	84.67
0	100	GT	C2	**88.55**	**95.85**	**98.49**	**31.41**	**76.65**	**94.34**	**75.33**	**85.85**
0	30	N	C1	**62.22**	**89.37**	**96.46**	**12.11**	**46.31**	**84.94**	**55.83**	**72.90**
0	30	N	C2	62.10	89.11	96.29	11.49	43.68	82.54	53.57	70.89
0	50	N	C1	**64.27**	**89.81**	**96.61**	**13.87**	**49.63**	**85.71**	**57.49**	**74.02**
0	50	N	C2	64.18	89.47	96.59	13.55	46.83	83.24	55.04	71.94
0	70	N	C1	69.21	**91.26**	**97.14**	17.65	**57.08**	**87.91**	**61.91**	**76.86**
0	70	N	C2	**69.82**	91.02	97.01	**18.19**	55.28	86.03	60.32	75.52
0	100	N	C1	85.37	**95.82**	98.34	25.94	73.12	94.20	73.06	84.65
0	100	N	C2	**87.89**	95.81	**98.47**	**29.32**	**75.77**	**94.24**	**74.65**	**85.50**
100	100	GT	C1	87.33	95.87	98.13	24.08	74.66	94.06	73.36	84.76
100	100	GT	C2	**90.59**	**96.42**	**98.68**	**27.59**	**76.87**	**94.60**	**74.99**	**85.80**
100	100	N	C1	87.06	95.86	98.14	23.86	74.65	94.05	73.31	84.72
100	100	N	C2	**89.89**	**96.39**	**98.70**	**25.99**	**76.05**	**94.56**	**74.44**	**85.51**

(b) Directed translation and joint pose

D	C	Pose	Model	T@2↑	T@5↑	T@15↑	A@15↑	A@30↑
0	0	–	C1	10.48	**46.43**	**87.99**	**57.97**	**74.81**
0	0	–	C2	**10.69**	46.30	87.46	57.68	74.41
30	0	–	C1	9.29	**42.11**	**86.41**	**55.68**	**73.15**
30	0	–	C2	**9.31**	41.92	86.08	55.39	72.88
50	0	–	C1	**9.03**	**41.77**	**85.92**	**55.38**	**72.78**
50	0	–	C2	8.98	41.56	85.44	55.01	72.49
70	0	–	C1	**8.84**	**41.17**	**85.88**	**55.29**	**72.60**
70	0	–	C2	8.47	40.65	85.55	54.93	72.37
100	0	–	C1	**8.32**	**40.77**	86.17	**55.39**	72.73
100	0	–	C2	8.13	40.33	**86.29**	55.21	**72.77**
0	30	GT	C1	**12.13**	**46.23**	**84.64**	**55.75**	**72.71**
0	30	GT	C2	11.45	43.76	82.20	53.51	70.68
0	50	GT	C1	**13.85**	**49.54**	**85.42**	**57.41**	**73.84**
0	50	GT	C2	13.72	46.92	82.97	55.03	71.78
0	70	GT	C1	17.68	**57.00**	**87.62**	**61.84**	**76.69**
0	70	GT	C2	**18.78**	55.51	85.86	60.46	75.46
0	100	GT	C1	25.89	73.32	94.18	73.08	84.65
0	100	GT	C2	**31.41**	**76.65**	**94.32**	**75.32**	**85.82**
0	30	N	C1	**12.11**	**46.26**	**84.65**	**55.73**	**72.70**
0	30	N	C2	11.48	43.63	82.22	53.47	70.65
0	50	N	C1	**13.86**	**49.57**	**85.40**	**57.39**	**73.82**
0	50	N	C2	13.55	46.79	82.94	54.93	71.72
0	70	N	C1	17.65	**57.05**	**87.63**	**61.82**	**76.67**
0	70	N	C2	**18.18**	55.25	85.76	60.24	75.33
0	100	N	C1	25.94	73.12	94.17	73.06	84.64
0	100	N	C2	**29.32**	**75.76**	**94.21**	**74.64**	**85.47**
100	100	GT	C1	24.08	74.66	93.99	73.33	84.70
100	100	GT	C2	**27.59**	**76.87**	**94.58**	**74.98**	**85.76**
100	100	N	C1	23.86	74.65	93.98	73.29	84.66
100	100	N	C2	**25.99**	**76.05**	**94.53**	**74.43**	**85.47**

(c) Per-frame median-aligned depth

D	C	Pose	Model	AbsRel↓	δ_1 ↑	AbsRel-U↓	$\delta_{1,U}$ ↑
0	0	–	C1	0.109	**89.62**	0.109	**89.62**
0	0	–	C2	0.109	89.59	0.109	89.59
30	0	–	C1	0.065	**94.00**	**0.084**	**91.86**
30	0	–	C2	0.065	93.95	0.085	91.78
50	0	–	C1	0.050	**95.58**	**0.080**	**92.15**
50	0	–	C2	0.050	95.56	0.081	92.10
70	0	–	C1	0.037	**97.14**	0.075	**92.77**
70	0	–	C2	0.037	97.12	0.075	92.69
100	0	–	C1	0.021	99.00	–	–
100	0	–	C2	0.021	**99.01**	–	–
0	30	GT	C1	**0.108**	**89.70**	**0.108**	**89.70**
0	30	GT	C2	0.109	89.65	0.109	89.65
0	50	GT	C1	0.108	**89.75**	0.108	**89.75**
0	50	GT	C2	0.108	89.70	0.108	89.70
0	70	GT	C1	**0.107**	**89.79**	**0.107**	**89.79**
0	70	GT	C2	0.108	89.75	0.108	89.75
0	100	GT	C1	0.107	**89.88**	0.107	**89.88**
0	100	GT	C2	0.107	89.83	0.107	89.83
0	30	N	C1	**0.108**	**89.70**	**0.108**	**89.70**
0	30	N	C2	0.109	89.65	0.109	89.65
0	50	N	C1	0.108	**89.75**	0.108	**89.75**
0	50	N	C2	0.108	89.71	0.108	89.71
0	70	N	C1	**0.107**	**89.79**	**0.107**	**89.79**
0	70	N	C2	0.108	89.75	0.108	89.75
0	100	N	C1	0.107	**89.88**	0.107	**89.88**
0	100	N	C2	0.107	89.83	0.107	89.83
100	100	GT	C1	0.020	99.01	–	–
100	100	GT	C2	0.020	99.01	–	–
100	100	N	C1	0.020	99.01	–	–
100	100	N	C2	0.020	99.01	–	–

R/T/A denote RRA/RTA/AUC; panel (b) uses directed errors. U denotes unprovided depth views. D/C are provided-view percentages.

Table 9: **Complete matched-training results on Virtual KITTI 2.** C1: OmniVGGT-trained; C2: VPGGT (PRC only). Bold marks the better displayed score within each matched pair. Angular thresholds are in degrees; angular scores and δ_1 are percentages.

(a) Sign-free pose metrics

D	C	Pose	Model	R@2↑	R@5↑	R@15↑	T@2↑	T@5↑	T@15↑	A@15↑	A@30↑
0	0	–	C1	100.00	100.00	100.00	**97.87**	99.65	99.81	**98.54**	**99.21**
0	0	–	C2	100.00	100.00	100.00	96.98	**99.68**	**99.84**	98.43	99.15
30	0	–	C1	100.00	100.00	100.00	**96.92**	99.65	99.84	**98.40**	**99.14**
30	0	–	C2	100.00	100.00	100.00	95.46	**99.68**	99.84	98.17	99.01
50	0	–	C1	**99.43**	100.00	100.00	**95.75**	**99.40**	**99.75**	**97.97**	**98.89**
50	0	–	C2	98.89	100.00	100.00	93.97	99.14	99.65	97.46	98.57
70	0	–	C1	99.30	100.00	100.00	**95.84**	**99.08**	**99.65**	**97.81**	**98.78**
70	0	–	C2	**99.33**	100.00	100.00	94.29	98.98	99.62	97.51	98.63
100	0	–	C1	**98.95**	100.00	100.00	**96.16**	98.54	99.24	**97.38**	98.44
100	0	–	C2	98.89	100.00	100.00	94.57	**98.79**	**99.43**	97.35	**98.47**
0	30	GT	C1	100.00	100.00	100.00	**99.05**	99.65	**99.81**	**99.02**	**99.45**
0	30	GT	C2	100.00	100.00	100.00	98.10	**99.68**	99.78	98.86	99.33
0	50	GT	C1	100.00	100.00	100.00	**98.95**	**99.68**	**99.81**	**98.97**	**99.43**
0	50	GT	C2	100.00	100.00	100.00	97.75	99.56	99.78	98.74	99.27
0	70	GT	C1	100.00	100.00	100.00	**97.68**	**99.33**	**99.84**	**98.75**	**99.32**
0	70	GT	C2	100.00	100.00	100.00	97.14	99.02	99.56	98.35	99.05
0	100	GT	C1	100.00	100.00	100.00	99.27	99.71	99.84	99.40	99.66
0	100	GT	C2	100.00	100.00	100.00	**99.59**	**99.81**	**99.94**	**99.76**	**99.87**
0	30	N	C1	100.00	100.00	100.00	**99.08**	99.65	**99.81**	**99.02**	**99.45**
0	30	N	C2	100.00	100.00	100.00	98.06	**99.68**	99.78	98.85	99.33
0	50	N	C1	100.00	100.00	100.00	**98.92**	**99.68**	**99.81**	**98.98**	**99.43**
0	50	N	C2	100.00	100.00	100.00	97.78	99.56	99.78	98.74	99.27
0	70	N	C1	100.00	100.00	100.00	**97.59**	**99.27**	**99.84**	**98.74**	**99.31**
0	70	N	C2	100.00	100.00	100.00	97.21	99.02	99.56	98.35	99.05
0	100	N	C1	100.00	100.00	100.00	99.24	99.68	99.84	99.39	99.66
0	100	N	C2	100.00	100.00	100.00	**99.59**	**99.81**	**99.94**	**99.72**	**99.85**
100	100	GT	C1	100.00	100.00	100.00	98.73	99.75	99.87	99.17	99.54
100	100	GT	C2	100.00	100.00	100.00	**99.27**	**99.81**	**99.94**	**99.64**	**99.79**
100	100	N	C1	100.00	100.00	100.00	98.63	99.75	99.87	99.13	99.52
100	100	N	C2	100.00	100.00	100.00	**99.27**	**99.78**	**99.94**	**99.60**	**99.77**

(b) Directed translation and joint pose

D	C	Pose	Model	T@2↑	T@5↑	T@15↑	A@15↑	A@30↑
0	0	–	C1	**97.87**	99.65	99.81	**98.54**	**99.21**
0	0	–	C2	96.98	**99.68**	**99.84**	98.43	99.15
30	0	–	C1	**96.92**	99.65	99.84	**98.40**	**99.14**
30	0	–	C2	95.46	**99.68**	99.84	98.17	99.01
50	0	–	C1	**95.75**	**99.40**	**99.75**	**97.97**	**98.88**
50	0	–	C2	93.97	99.14	99.59	97.42	98.52
70	0	–	C1	**95.84**	**99.08**	**99.62**	**97.80**	**98.75**
70	0	–	C2	94.29	98.98	99.59	97.50	98.61
100	0	–	C1	**95.75**	**97.90**	**98.54**	**96.83**	**97.75**
100	0	–	C2	94.22	97.71	98.25	96.39	97.41
0	30	GT	C1	**99.05**	99.65	**99.81**	**99.02**	**99.45**
0	30	GT	C2	98.10	**99.68**	99.78	98.86	99.33
0	50	GT	C1	**98.95**	**99.68**	**99.81**	**98.97**	**99.43**
0	50	GT	C2	97.75	99.56	99.78	98.74	99.27
0	70	GT	C1	**97.68**	**99.33**	**99.84**	**98.75**	**99.32**
0	70	GT	C2	97.14	99.02	99.56	98.35	99.05
0	100	GT	C1	99.27	99.71	99.84	99.40	99.66
0	100	GT	C2	**99.59**	**99.81**	**99.94**	**99.76**	**99.87**
0	30	N	C1	**99.08**	99.65	**99.81**	**99.02**	**99.45**
0	30	N	C2	98.06	**99.68**	99.78	98.85	99.33
0	50	N	C1	**98.92**	**99.68**	**99.81**	**98.98**	**99.43**
0	50	N	C2	97.78	99.56	99.78	98.74	99.27
0	70	N	C1	**97.59**	**99.27**	**99.84**	**98.74**	**99.31**
0	70	N	C2	97.21	99.02	99.56	98.35	99.05
0	100	N	C1	99.24	99.68	99.84	99.39	99.66
0	100	N	C2	**99.59**	**99.81**	**99.94**	**99.72**	**99.85**
100	100	GT	C1	98.73	99.75	99.87	99.17	99.54
100	100	GT	C2	**99.27**	**99.81**	**99.94**	**99.64**	**99.79**
100	100	N	C1	98.63	99.75	99.87	99.13	99.52
100	100	N	C2	**99.27**	**99.78**	**99.94**	**99.60**	**99.77**

(c) Per-frame median-aligned depth

D	C	Pose	Model	AbsRel↓	δ_1 ↑	AbsRel-U↓	$\delta_{1,U}$ ↑
0	0	–	C1	0.080	**92.15**	0.080	**92.15**
0	0	–	C2	0.080	92.09	0.080	92.09
30	0	–	C1	0.071	**93.37**	0.077	**92.59**
30	0	–	C2	0.071	93.33	0.077	92.53
50	0	–	C1	0.067	**93.95**	0.077	**92.66**
50	0	–	C2	**0.066**	93.93	0.077	92.61
70	0	–	C1	0.062	**94.51**	0.077	**92.64**
70	0	–	C2	0.062	94.49	**0.076**	92.60
100	0	–	C1	0.056	95.39	–	–
100	0	–	C2	**0.055**	95.39	–	–
0	30	GT	C1	0.080	**92.22**	0.080	**92.22**
0	30	GT	C2	**0.079**	92.16	**0.079**	92.16
0	50	GT	C1	0.080	**92.22**	0.080	**92.22**
0	50	GT	C2	**0.079**	92.17	**0.079**	92.17
0	70	GT	C1	0.080	**92.23**	0.080	**92.23**
0	70	GT	C2	**0.079**	92.17	**0.079**	92.17
0	100	GT	C1	0.080	**92.23**	0.080	**92.23**
0	100	GT	C2	**0.079**	92.17	**0.079**	92.17
0	30	N	C1	0.080	**92.22**	0.080	**92.22**
0	30	N	C2	**0.079**	92.16	**0.079**	92.16
0	50	N	C1	0.080	**92.22**	0.080	**92.22**
0	50	N	C2	**0.079**	92.17	**0.079**	92.17
0	70	N	C1	0.080	**92.23**	0.080	**92.23**
0	70	N	C2	**0.079**	92.17	**0.079**	92.17
0	100	N	C1	0.080	**92.23**	0.080	**92.23**
0	100	N	C2	**0.079**	92.17	**0.079**	92.17
100	100	GT	C1	0.056	95.40	–	–
100	100	GT	C2	**0.055**	95.40	–	–
100	100	N	C1	0.056	95.40	–	–
100	100	N	C2	**0.055**	95.40	–	–

R/T/A denote RRA/RTA/AUC; panel (b) uses directed errors. U denotes unprovided depth views. D/C are provided-view percentages.

Table 10: **Complete matched-training results on TUM RGB-D.** C1: OmniVGGT-trained; C2: VPGGT (PRC only). Bold marks the better displayed score within each matched pair. Angular thresholds are in degrees; angular scores and δ_1 are percentages.

(a) Sign-free pose metrics

D	C	Pose	Model	R@2↑	R@5↑	R@15↑	T@2↑	T@5↑	T@15↑	A@15↑	A@30↑
0	0	–	C1	94.78	96.33	97.22	**22.22**	**51.22**	75.78	**56.78**	69.17
0	0	–	C2	**95.44**	**97.22**	97.22	20.22	50.33	**76.56**	56.53	**69.33**
30	0	–	C1	94.11	97.00	97.22	**17.44**	47.33	73.00	52.87	66.46
30	0	–	C2	**96.56**	**97.22**	97.22	17.11	**48.56**	**73.44**	**52.90**	**66.66**
50	0	–	C1	94.78	95.56	97.22	**17.11**	45.78	**73.56**	**52.44**	66.21
50	0	–	C2	**96.89**	**97.22**	97.22	16.56	**45.89**	73.33	52.10	**66.46**
70	0	–	C1	94.33	95.33	97.00	19.11	47.89	75.56	54.80	68.22
70	0	–	C2	**95.44**	**97.22**	**97.22**	19.11	**49.22**	**76.78**	**55.41**	**68.87**
100	0	–	C1	93.56	95.44	95.44	**27.22**	51.89	78.11	58.50	70.71
100	0	–	C2	**95.89**	**96.22**	**97.22**	26.00	**52.33**	**79.44**	**59.19**	**71.77**
0	30	GT	C1	95.00	96.44	97.22	19.11	49.00	75.33	55.01	67.92
0	30	GT	C2	**95.67**	**97.22**	97.22	**20.33**	**51.00**	**76.33**	**56.64**	**69.29**
0	50	GT	C1	95.11	96.44	97.22	16.89	49.67	74.44	54.41	67.19
0	50	GT	C2	**95.67**	**97.22**	97.22	**24.00**	**55.00**	**77.89**	**58.95**	**70.74**
0	70	GT	C1	95.11	96.11	96.56	19.22	51.56	75.56	56.17	68.26
0	70	GT	C2	**95.44**	**96.56**	**97.67**	**31.56**	**60.11**	**80.22**	**62.79**	**73.50**
0	100	GT	C1	96.22	**97.22**	97.22	25.11	58.78	77.56	60.35	71.04
0	100	GT	C2	**96.44**	96.89	**98.89**	**52.00**	**73.78**	**84.00**	**72.22**	**80.46**
0	30	N	C1	95.00	96.44	97.22	**18.67**	49.22	74.67	54.95	67.85
0	30	N	C2	**95.56**	**97.22**	97.22	18.33	**49.78**	**75.67**	**55.70**	**68.64**
0	50	N	C1	95.11	96.44	97.22	16.22	49.11	74.56	54.08	67.00
0	50	N	C2	**95.67**	**97.22**	97.22	**20.11**	**51.11**	**77.00**	**56.81**	**69.28**
0	70	N	C1	95.11	96.11	96.56	19.67	51.44	75.56	56.23	68.25
0	70	N	C2	**95.44**	**96.56**	**97.67**	**22.56**	**55.00**	**78.11**	**58.86**	**70.89**
0	100	N	C1	96.11	97.22	97.22	27.22	58.89	76.11	60.44	70.89
0	100	N	C2	**96.78**	97.22	**98.89**	**35.33**	**67.00**	**81.44**	**66.07**	**76.37**
100	100	GT	C1	94.89	95.44	95.44	27.89	55.89	78.00	59.99	71.86
100	100	GT	C2	**95.33**	**96.56**	**97.22**	**41.78**	**68.44**	**84.78**	**68.91**	**79.08**
100	100	N	C1	94.78	95.44	95.44	27.33	55.56	78.56	59.01	71.16
100	100	N	C2	**95.22**	**96.67**	**97.22**	**27.78**	**63.44**	**82.11**	**63.78**	**75.31**

(b) Directed translation and joint pose

D	C	Pose	Model	T@2↑	T@5↑	T@15↑	A@15↑	A@30↑
0	0	–	C1	**22.22**	**51.22**	75.67	**56.73**	69.09
0	0	–	C2	20.22	50.33	**76.44**	56.50	**69.17**
30	0	–	C1	**17.44**	47.33	72.78	52.84	66.23
30	0	–	C2	17.11	**48.56**	**73.44**	**52.90**	**66.56**
50	0	–	C1	**17.11**	45.78	**73.44**	**52.42**	66.00
50	0	–	C2	16.56	**45.89**	73.22	52.10	**66.39**
70	0	–	C1	19.11	47.89	75.33	54.75	67.91
70	0	–	C2	19.11	**49.22**	**76.67**	**55.39**	**68.66**
100	0	–	C1	**27.22**	51.89	78.00	58.50	70.58
100	0	–	C2	26.00	**52.33**	**79.33**	**59.19**	**71.65**
0	30	GT	C1	19.11	49.00	75.33	55.01	67.88
0	30	GT	C2	**20.33**	**51.00**	**76.33**	**56.64**	**69.21**
0	50	GT	C1	16.89	49.67	74.44	54.41	67.17
0	50	GT	C2	**24.00**	**55.00**	**77.67**	**58.95**	**70.63**
0	70	GT	C1	19.22	51.56	75.44	56.14	68.19
0	70	GT	C2	**31.56**	**60.11**	**80.22**	**62.79**	**73.49**
0	100	GT	C1	25.11	58.78	77.56	60.35	71.04
0	100	GT	C2	**52.00**	**73.78**	**84.00**	**72.22**	**80.46**
0	30	N	C1	**18.67**	49.22	74.67	54.95	67.81
0	30	N	C2	18.33	**49.78**	**75.67**	**55.70**	**68.53**
0	50	N	C1	16.22	49.11	74.56	54.08	66.98
0	50	N	C2	**20.11**	**51.11**	**76.78**	**56.81**	**69.21**
0	70	N	C1	19.67	51.44	75.22	56.22	68.19
0	70	N	C2	**22.56**	**55.00**	**78.11**	**58.86**	**70.87**
0	100	N	C1	27.22	58.89	76.11	60.44	70.89
0	100	N	C2	**35.33**	**67.00**	**81.44**	**66.07**	**76.37**
100	100	GT	C1	27.89	55.89	78.00	59.99	71.81
100	100	GT	C2	**41.78**	**68.44**	**84.78**	**68.91**	**79.08**
100	100	N	C1	27.33	55.56	78.56	59.01	71.13
100	100	N	C2	**27.78**	**63.44**	**82.11**	**63.78**	**75.31**

(c) Per-frame median-aligned depth

D	C	Pose	Model	AbsRel↓	δ_1 ↑	AbsRel-U↓	$\delta_{1,U}$ ↑
0	0	–	C1	0.058	**94.64**	0.058	**94.64**
0	0	–	C2	0.058	94.41	0.058	94.41
30	0	–	C1	0.031	**98.05**	0.040	**97.32**
30	0	–	C2	0.031	98.04	0.040	97.31
50	0	–	C1	0.024	**98.58**	0.038	**97.40**
50	0	–	C2	**0.023**	98.57	**0.037**	97.39
70	0	–	C1	0.017	99.11	0.034	97.62
70	0	–	C2	**0.016**	99.11	**0.033**	97.62
100	0	–	C1	0.009	99.76	–	–
100	0	–	C2	0.009	99.76	–	–
0	30	GT	C1	**0.057**	**94.66**	**0.057**	**94.66**
0	30	GT	C2	0.058	94.44	0.058	94.44
0	50	GT	C1	**0.057**	**94.68**	**0.057**	**94.68**
0	50	GT	C2	0.058	94.46	0.058	94.46
0	70	GT	C1	**0.057**	**94.69**	**0.057**	**94.69**
0	70	GT	C2	0.058	94.46	0.058	94.46
0	100	GT	C1	**0.057**	**94.69**	**0.057**	**94.69**
0	100	GT	C2	0.058	94.49	0.058	94.49
0	30	N	C1	**0.057**	**94.66**	**0.057**	**94.66**
0	30	N	C2	0.058	94.44	0.058	94.44
0	50	N	C1	**0.057**	**94.68**	**0.057**	**94.68**
0	50	N	C2	0.058	94.46	0.058	94.46
0	70	N	C1	**0.057**	**94.69**	**0.057**	**94.69**
0	70	N	C2	0.058	94.46	0.058	94.46
0	100	N	C1	**0.057**	**94.70**	**0.057**	**94.70**
0	100	N	C2	0.058	94.51	0.058	94.51
100	100	GT	C1	0.009	99.76	–	–
100	100	GT	C2	0.009	99.76	–	–
100	100	N	C1	0.009	99.76	–	–
100	100	N	C2	0.009	99.76	–	–

R/T/A denote RRA/RTA/AUC; panel (b) uses directed errors. U denotes unprovided depth views. D/C are provided-view percentages.

Table 11: **Complete matched-training results on C3VD.** C1: OmniVGGT-trained; C2: VPGGT (PRC only). Bold marks the better displayed score within each matched pair. Angular thresholds are in degrees; angular scores and δ_1 are percentages.

(a) Sign-free pose metrics

D	C	Pose	Model	R@2↑	R@5↑	R@15↑	T@2↑	T@5↑	T@15↑	A@15↑	A@30↑
0	0	–	C1	100.00	100.00	100.00	**8.64**	**29.19**	71.85	**43.97**	**61.55**
0	0	–	C2	100.00	100.00	100.00	7.65	27.75	**72.30**	42.73	61.49
30	0	–	C1	99.95	100.00	100.00	**6.32**	25.33	60.30	36.48	54.68
30	0	–	C2	**100.00**	100.00	100.00	6.17	**26.12**	**61.38**	**36.58**	**54.92**
50	0	–	C1	100.00	100.00	100.00	7.75	30.52	63.41	39.78	57.00
50	0	–	C2	100.00	100.00	100.00	**8.40**	**30.72**	**64.69**	**40.18**	**57.29**
70	0	–	C1	100.00	100.00	100.00	11.31	**37.48**	69.88	46.17	62.24
70	0	–	C2	100.00	100.00	100.00	11.31	36.99	**70.12**	**46.45**	**62.33**
100	0	–	C1	100.00	100.00	100.00	16.74	53.23	83.51	59.73	74.26
100	0	–	C2	100.00	100.00	100.00	**17.04**	**54.22**	**84.30**	**60.13**	**74.59**
0	30	GT	C1	100.00	100.00	100.00	8.00	27.95	70.81	42.07	60.43
0	30	GT	C2	100.00	100.00	100.00	**9.68**	**28.74**	**71.01**	**42.23**	**60.89**
0	50	GT	C1	100.00	100.00	100.00	8.69	28.89	71.80	43.31	61.84
0	50	GT	C2	100.00	100.00	100.00	**15.01**	**35.26**	**73.19**	**46.99**	**64.52**
0	70	GT	C1	100.00	100.00	100.00	9.23	32.30	73.98	46.34	64.50
0	70	GT	C2	100.00	100.00	100.00	**25.28**	**46.81**	**78.67**	**56.20**	**71.01**
0	100	GT	C1	100.00	100.00	100.00	10.37	43.16	78.37	52.76	69.42
0	100	GT	C2	100.00	100.00	100.00	**47.90**	**76.74**	**95.41**	**78.63**	**88.48**
0	30	N	C1	100.00	100.00	100.00	**7.90**	**27.65**	**69.78**	**41.53**	**59.62**
0	30	N	C2	100.00	100.00	100.00	7.26	26.02	68.44	39.77	58.60
0	50	N	C1	100.00	100.00	100.00	8.30	28.00	**70.62**	**42.30**	**60.38**
0	50	N	C2	100.00	100.00	100.00	**10.22**	**28.84**	67.75	40.87	58.77
0	70	N	C1	100.00	100.00	100.00	9.19	31.85	**72.49**	**45.31**	**62.94**
0	70	N	C2	100.00	100.00	100.00	**14.81**	**34.77**	68.69	45.06	61.50
0	100	N	C1	100.00	100.00	100.00	9.48	42.62	77.38	51.77	67.73
0	100	N	C2	100.00	100.00	100.00	**29.28**	**55.46**	**79.46**	**60.70**	**72.81**
100	100	GT	C1	100.00	100.00	100.00	20.74	55.51	85.19	62.26	76.36
100	100	GT	C2	100.00	100.00	100.00	**56.49**	**81.78**	**97.19**	**82.95**	**91.04**
100	100	N	C1	100.00	100.00	100.00	19.85	54.07	**83.16**	60.86	**74.80**
100	100	N	C2	100.00	100.00	100.00	**35.16**	**60.40**	80.15	**63.58**	74.53

(b) Directed translation and joint pose

D	C	Pose	Model	T@2↑	T@5↑	T@15↑	A@15↑	A@30↑
0	0	–	C1	**8.64**	**29.19**	71.70	**43.95**	**61.39**
0	0	–	C2	7.65	27.75	**72.20**	42.70	61.27
30	0	–	C1	**6.32**	25.28	59.80	36.30	54.17
30	0	–	C2	6.17	**26.07**	**60.99**	**36.45**	**54.39**
50	0	–	C1	7.70	30.47	63.01	39.63	56.60
50	0	–	C2	**8.40**	**30.67**	**64.30**	**40.06**	**56.78**
70	0	–	C1	11.31	**37.43**	69.58	46.08	61.89
70	0	–	C2	11.31	36.99	**69.83**	**46.35**	61.89
100	0	–	C1	16.74	53.23	83.51	59.73	74.26
100	0	–	C2	**17.04**	**54.22**	**84.30**	**60.13**	**74.58**
0	30	GT	C1	8.00	27.95	70.72	42.05	60.29
0	30	GT	C2	**9.68**	**28.74**	**70.91**	**42.19**	**60.70**
0	50	GT	C1	8.69	28.89	71.70	43.29	61.68
0	50	GT	C2	**15.01**	**35.26**	**73.14**	**46.99**	**64.41**
0	70	GT	C1	9.23	32.25	73.88	46.28	64.36
0	70	GT	C2	**25.28**	**46.81**	**78.62**	**56.17**	**70.91**
0	100	GT	C1	10.37	43.16	78.37	52.76	69.39
0	100	GT	C2	**47.90**	**76.74**	**95.41**	**78.63**	**88.48**
0	30	N	C1	**7.85**	**27.60**	**69.63**	**41.47**	**59.40**
0	30	N	C2	7.26	26.02	68.25	39.70	58.35
0	50	N	C1	8.30	27.95	**70.52**	**42.26**	**60.22**
0	50	N	C2	**10.22**	**28.84**	67.65	40.85	58.59
0	70	N	C1	9.14	31.80	**72.40**	**45.23**	**62.75**
0	70	N	C2	**14.81**	**34.77**	68.64	45.05	61.35
0	100	N	C1	9.48	42.57	77.33	51.72	67.66
0	100	N	C2	**29.28**	**55.46**	**79.36**	**60.67**	**72.75**
100	100	GT	C1	20.74	55.51	85.14	62.26	76.34
100	100	GT	C2	**56.49**	**81.78**	**97.19**	**82.95**	**91.04**
100	100	N	C1	19.85	54.07	**83.16**	60.86	**74.78**
100	100	N	C2	**35.16**	**60.40**	80.05	**63.56**	74.47

(c) Per-frame median-aligned depth

D	C	Pose	Model	AbsRel↓	δ_1 ↑	AbsRel-U↓	$\delta_{1,U}$ ↑
0	0	–	C1	0.055	97.87	0.055	97.87
0	0	–	C2	0.055	**98.09**	0.055	**98.09**
30	0	–	C1	0.015	99.78	0.020	99.70
30	0	–	C2	0.015	**99.80**	**0.019**	**99.72**
50	0	–	C1	0.011	99.86	0.017	99.75
50	0	–	C2	**0.010**	**99.87**	0.017	99.75
70	0	–	C1	0.007	99.92	0.015	99.77
70	0	–	C2	0.007	99.92	**0.014**	**99.78**
100	0	–	C1	0.004	99.98	–	–
100	0	–	C2	0.004	99.98	–	–
0	30	GT	C1	0.055	97.88	0.055	97.88
0	30	GT	C2	0.055	**98.10**	0.055	**98.10**
0	50	GT	C1	0.055	97.89	0.055	97.89
0	50	GT	C2	0.055	**98.11**	0.055	**98.11**
0	70	GT	C1	0.055	97.89	0.055	97.89
0	70	GT	C2	0.055	**98.11**	0.055	**98.11**
0	100	GT	C1	0.055	97.90	0.055	97.90
0	100	GT	C2	0.055	**98.11**	0.055	**98.11**
0	30	N	C1	0.055	97.88	0.055	97.88
0	30	N	C2	0.055	**98.10**	0.055	**98.10**
0	50	N	C1	0.055	97.89	0.055	97.89
0	50	N	C2	0.055	**98.11**	0.055	**98.11**
0	70	N	C1	0.055	97.89	0.055	97.89
0	70	N	C2	0.055	**98.11**	0.055	**98.11**
0	100	N	C1	0.055	97.90	0.055	97.90
0	100	N	C2	0.055	**98.11**	0.055	**98.11**
100	100	GT	C1	0.004	99.98	–	–
100	100	GT	C2	0.004	99.98	–	–
100	100	N	C1	0.004	99.98	–	–
100	100	N	C2	0.004	99.98	–	–

R/T/A denote RRA/RTA/AUC; panel (b) uses directed errors. U denotes unprovided depth views. D/C are provided-view percentages.

The implementation applies this ratio only when the prior translation-norm sum exceeds 10^{-4}, with a denominator floor of 10^{-6}. Otherwise, it uses $\alpha = 1$. The aligned prior is $\mathbf{a}_i = (\alpha \bar{\mathbf{t}}_i^p, \mathbf{q}_i^p, \mathbf{f}_i^p)$.

The IT0 prediction is recorded before state replacement. Provided views then use $\mathbf{a}_i$, while unprovided views retain their IT0 states. The running state is detached before further refinement, following the inherited camera-head computation. Subsequent updates are additive in the camera encoding rather than compositions on SE(3). The existing activation and camera-decoding functions are retained.

PRC does not add a new refinement network or introduce the original residual-update mechanism. It changes the state to which that mechanism is applied. For a provided view, the numerical forward computation can be written as $\mathbf{u}_i^{J-1} = \mathbf{a}_i + \sum_{k=1}^{J-1} \Delta_i^k$. This expression describes forward values, not an uninterrupted gradient path through the iterations. The residuals are unconstrained, so zero corrections preserve the aligned prior state but are not enforced.

A.7 Further Directions

Raw depth priors. This work uses dataset depth references without additional corruption. Extending the model to raw depth observations requires studying sensor-specific noise, missing values, and systematic artifacts. A dedicated depth-processing front end is one possible extension, but is not evaluated here.

Additional sensing modalities. Other sensors may contribute through upstream estimators that produce camera or depth observations. Such integration requires appropriate calibration, coordinate conversion, and treatment of uncertainty. The current geometric interface does not by itself establish robustness to arbitrary sensor error distributions.

Prior reliability. Detecting and rejecting unreliable priors is an important direction. The present model does not include an explicit reliability gate, and its results do not establish safe fallback behavior under severe prior corruption.

Vision dropout. The current architecture requires images. Training and evaluating missing or degraded visual inputs would require additional design and supervision. Sensor-only inference and fallback to odometry are not demonstrated in this work.

Streaming reconstruction. Extending window-level prior conditioning to streaming reconstruction may enable longer-term geometric constraints. Maintaining a consistent state across windows and incorporating loop-closure information remain open integration problems.