GAUGE: GROUP-WISE VIEW-INCONSISTENCY RECTIFICATION FOR FEED-FORWARD 4D TRACKING

Anonymous authors
Paper under double-blind review

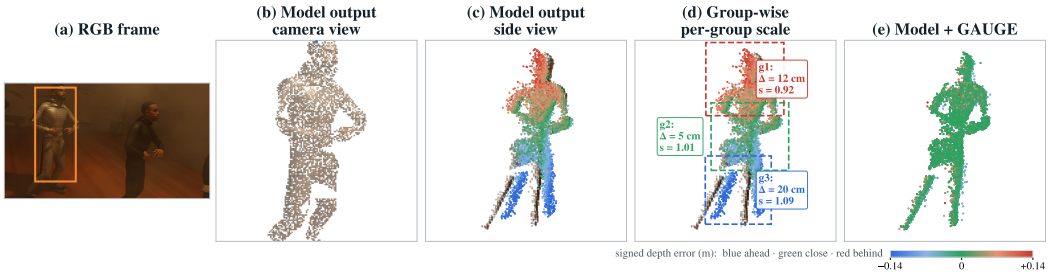


Figure 1: Per-group radial scale ambiguity. (a) RGB frame. (b) The prediction appears accurate in the camera view, but (c) seen from the side, it exposes a systematic depth bias along the view direction; (d) grouping recovers this bias as per-group radial scales, and (e) GAUGE removes it.

ABSTRACT

Feed-forward models regress dense 3D point trajectories directly from monocular video, yet the residual after global alignment is substantial and lacks a structural explanation. Measured on dynamic query points across models and datasets, the error concentrates along the view direction, while the scale correction each motion group requires differs. The predicted displacement direction nevertheless supports reliable grouping, with a median angle far below the 90° random baseline. The systematic part of the residual is therefore a family of radial degrees of freedom per motion group, along directions 2D observations cannot constrain. We call it group-wise view inconsistency. We present GAUGE (*Group-wise Adaptive Unsupervised Gauge Estimation*), a training-free and model-agnostic post-hoc module. It recovers motion groups from direction consistency and spatial connectivity, then estimates a per-frame radial scale and group-level translation from 1% to 5% metric anchors, four degrees of freedom per group and frame. On dynamic query points of eight trackers, including D4RT, 4RC and SM4RT, our correction lowers endpoint error by 15.1% to 62.6% over the uncorrected predictions, while spending the same anchors on gradient fine-tuning improves the same models by only -1.1% to 15.2%.

1 INTRODUCTION

Recovering dense 3D motion from monocular video is a basic capability for applications such as embodied perception and augmented reality. A monocular observation determines scene geometry and motion only up to a global similarity transform, so physical scale is unrecoverable and downstream tasks that need metric information cannot benefit directly. Feed-forward models have changed the efficiency landscape. D4RT, 4RC, V-DPM and SM4RT (Zhang et al., 2026; Luo et al., 2026; Sucar et al., 2026; Lin et al., 2026) regress long dense 3D trajectories in a single forward pass, but their accuracy lags far behind their speed.

Existing work absorbs scale priors during training, adds test-time optimization or bundle adjustment, takes depth as an extra input, or injects structured motion priors. Their costs are priors bound to the training distribution, test-time overhead, an external depth source and larger models. Section 2 surveys each line. A more basic question is whether the error itself has a usable structure.

We therefore characterize the error across the eight trackers with three measurements. First, on the five feed-forward models the predicted direction is closer to ground truth than the roughly 90° random baseline, and the residual concentrates along the view direction, with 61.6% to 77.7% of the residual energy on dynamic query points (points on moving objects) falling along it. Second, the residual is not a single global scale. A single global factor makes things worse (Section 3.3), and a seven-degree-of-freedom Sim(3) fit under the same grouping and the same anchors is even weaker

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.

than no correction on some models. Third, the bias is not independent per point but aligned across motion groups. Unsupervised grouping explains 15.1% to 76.7% of the per-point optimal radial scale. All three properties are visible in the point cloud before and after grouping (Figure 1).

The above three properties share one source. For every frame and group, the radial distance to the camera center is never constrained by 2D correspondences. Together these distances form a family of degrees of freedom per motion group, which we call *group-wise view inconsistency*.

To remove this inconsistency, we present GAUGE (***G**roup-wise **A**daptive **U**nsupervised **G**auge **E**stimation*). It treats direction and magnitude separately, grouping trajectories by flow-direction consistency and spatial proximity alone. It then estimates a per-frame radial scale and a group-level translation per group, four degrees of freedom per group and frame, from true scale observations covering only 1% to 5% of the query points, which need not be precise.

We validate this design on eight trackers, including D4RT, 4RC and SM4RT, and two benchmarks. Endpoint error on dynamic query points drops by 15.1% to 62.6% over the uncorrected trackers. On most models, four degrees of freedom beat a seven-degree-of-freedom Sim(3) fit with the same grouping and anchors. If the same number of ground-truth observations is spent on gradient fine-tuning, the gain in endpoint error on the same models is only -1.1% to 15.2%. The recoverable part of this error is therefore limited by its structure rather than by anchor amount or model capacity, as Section 5.7 quantifies with an upper bound.

Our contributions are threefold.

- Three measurements on eight trackers show that the residual error of feed-forward 4D trackers is structured. It concentrates along the view direction, the radial scale a point requires is shared within its motion group, and a per-group family can remove what a single global factor cannot.
- We recover these motion groups from what the tracker already outputs, linking points whose flow directions agree on shared frames and that are close in space, so the correction acts group by group without extra supervision.
- We build GAUGE on the groups recovered above. Each group receives a per-frame radial scale and a group-level translation, four degrees of freedom in total, estimated from a few metric anchors, and the correction lowers endpoint error across trackers and benchmarks.

2 Related Work

2.1 Feed-Forward 3D Motion Prediction

Feed-forward reconstruction methods such as DUSt3R, MASt3R and VGGT (Wang et al., 2024; Leroy et al., 2024; Wang et al., 2025) regress scene geometry directly from unposed images, replacing the multi-stage optimization pipeline. MonST3R (Zhang et al., 2025b) extends this representation to dynamic scenes. D4RT, 4RC and V-DPM (Zhang et al., 2026; Luo et al., 2026; Sucar et al., 2026) regress long dense 3D trajectories, SM4RT (Lin et al., 2026) additionally characterizes scene dynamics with explicit SE(3) motion bases, while TAPIP3D, SpatialTrackerV2, DELTA and CoTracker3 (Zhang et al., 2025a; Xiao et al., 2025; Ngo et al., 2025; Karaev et al., 2025) approach the same task through depth input, explicit camera motion, dense long-range tracking and a depth-lifting pipeline, respectively. St4RTrack (Feng et al., 2025) places reconstruction and tracking in the world coordinate frame and provides a corresponding benchmark. Metric depth methods such as ZoeDepth, Metric3D and Depth Pro (Bhat et al., 2023; Yin et al., 2023; Bochkovskiy et al., 2025) output absolute scale directly from priors, but on single-frame depth rather than on temporal trajectories. GAUGE is orthogonal to all of these methods, acting directly on the 3D trajectories they output.

Parallel to the feed-forward line is test-time optimization. One family optimizes explicit geometric objectives. DROID-SLAM (Teed & Deng, 2022) updates pose and depth iteratively via differentiable bundle adjustment, and MegaSaM (Li et al., 2025) extends to unconstrained dynamic video with depth priors and motion probability maps. Another updates only a small subset of weights, typically the affine parameters of normalization layers or the output head, with TENT, T3A and MEMO (Wang et al., 2021; Iwasawa & Matsuo, 2021; Zhang et al., 2022) as representatives, and the fine-tuning comparison in Section 5.3 follows this protocol. Both families add computation at test time, whereas GAUGE enters no optimization loop, produces no gradient and solves four degrees of freedom per group after inference.

2.2 Unsupervised Motion Grouping

Decomposing a scene into several jointly moving rigid bodies is a long-standing modeling assumption. In scene flow estimation, methods recover moving groups without dense flow supervision, from weak supervision (Gojcic et al., 2021), self-supervision (Baur et al., 2021), or explicit dynamic classification (Zhang et al., 2024a). On 3D representations, multi-body SE(3)-equivariant methods (Zhong et al., 2023) recover rigid segmentation and motion jointly from point clouds, PD^{2}GS (Wang et al., 2026) clusters Gaussian primitives by normalized displacement direction, and AiM (Ai et al., 2026) clusters motion primitives into rigid bodies with sequential RANSAC and no prior on the number of parts. What these methods share is treating grouping as a standalone segmentation task. Grouping in GAUGE requires only that points inside a group agree in flow direction, so that four degrees of freedom per group suffice to describe them.

2.3 Sparse Metric Anchoring

It is classical that absolute scale is unobservable in monocular vision. Systems such as ORB-SLAM and DSO (Campos et al., 2021; Engel et al., 2018) recover trajectories only up to a global similarity transform, and the field has long handled scale drift with Sim(3) pose-graph optimization and global alignment. With learned depth priors, estimating a per-frame scale and offset and folding them into bundle adjustment has become standard practice, as in HI-SLAM, S3PO-GS and MASt3R-SLAM (Zhang et al., 2024b; Cheng et al., 2025; Murai et al., 2025). The standard treatment of unobservable directions in bundle adjustment goes back to the work of Triggs et al. (2000) on gauge freedom. Appendix A.2 uses the same term but restricts it to the invisible directions defined by the projection observations.

Calibrating predictions directly with sparse true scale observations is another line, found mainly in monocular depth estimation. Li & Althoff (2026) fit a linear combination of low-dimensional basis maps to sparse anchors, MRAC and VI-Depth (Roy et al., 2026; Wofk et al., 2023) suppress outlier anchors with robust statistics, and PRISM-SLAM (Im et al., 2026) instead makes metric scale Fisher-identifiable in the factor graph with a Plücker ray distance factor.

GAUGE is likewise training-free. It handles temporal 3D trajectories with per-frame drift rather than single-frame depth, and expands the radial degrees of freedom per motion group rather than fitting a global low-dimensional basis map. This matches the error structure measured in Section 3, and therefore needs only 1% to 5% sparse anchors.

3 Error Structure of Feed-Forward 4D Trackers

3.1 Diagnosis Setup

All measurements in this section are performed after global alignment, that is, on predicted trajectories and ground truth in the same reference frame. We evaluate the eight trackers of Section 5.1. The diagnosis uses the test split of the synthetic dataset PointOdyssey (Zheng et al., 2023). The cross-dataset replication uses WorldTrack (Feng et al., 2025), and the reasons for these choices are given in Section 5.1. Of the two query protocols, DQS (dynamic query selection) keeps only dynamic query points and the Full protocol uses all query points visible in the first frame. Unless stated otherwise, this section reports DQS. The three structural quantities are point-level statistics, with definitions, thresholds and aggregation in Appendix A.3.

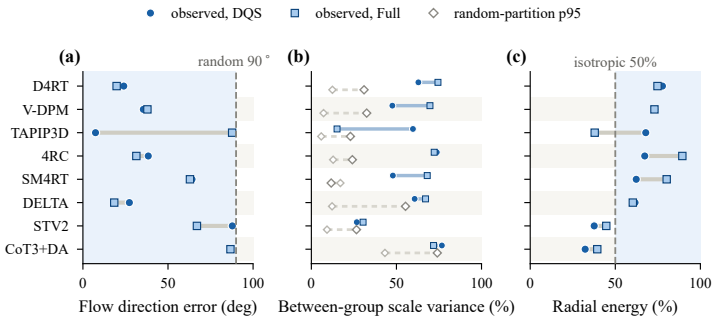


Figure 2: Error structure. Values in Tables A3 and A13.

3.2 Direction Accuracy, Group Structure and Radial Concentration

Motion direction is reliable, whereas motion magnitude carries a systematic bias. Across the eight models, the median angle between the predicted flow direction and ground truth is 7.4° to 87.8°, with the five feed-forward models between 24° and 64°, all better than the roughly 90°

random baseline (Figure 2a). Reliable direction with biased magnitude rules out isotropic noise, which would corrupt direction as well.

The bias aligns across motion groups rather than being independent per point. We compute for each point the radial scale that would bring it into agreement with ground truth, and measure how much of it is explained by the point's motion group, the between-group scale variance (Figure 2b). Across the eight models this fraction is 15.1% to 76.7%. Finer grouping raises the ratio itself, so we compare against a size-preserving label-permutation null. The observed values exceed its 95th percentile in most settings, by 18 to 61 percentage points, which rules out independent per-point noise. The procedure and the per-setting values are in Appendix C.5.

The residual concentrates along the view direction. On dynamic query points, 61.6% to 77.7% of the residual energy lies along the view direction (Figure 2c), far above the ratio of 0.5 that isotropic error would give. That correction can act on the radial direction alone rests on a separate property, that radial correction is compatible with the observed 2D trajectory while tangential correction is not (Box 1). Radial energy and flow direction error characterize where the bias concentrates rather than the recoverable amount, as Section 5.7 makes quantitative.

3.3 Failure of Global Scale Correction

If the residual were entirely caused by a single global factor, no residual requiring separate correction should remain after global alignment, and this is not what we measure. Fitting a single scale factor over the whole sequence without grouping leaves D4RT slightly worse than no correction (Table 1), and on 4RC it moves the error in the right direction but far short of the grouped result (Appendix Table A10). A single global factor is therefore not a harmless approximation under a misspecified model.

Table 1: Parameterization comparison under the same grouping and anchors (D4RT).

Parameterization	**DOF** (T frames)	**EPE** $\downarrow$
No correction	0	0.1812 / 0.1930
Single global scale	1	0.1835 / 0.1927
Per-group Sim(3)	7	0.1528 / 0.1983
GAUGE (ours)	$T+3$	**0.1269 / 0.1597**

A fit with more degrees of freedom under the same anchors fails as well. A full seven-degree-of-freedom Sim(3) similarity transform under the same grouping and the same anchors is worse than no correction on some models, and on dynamic query points 4RC and SM4RT degrade most (last column of Table 2). The extra rotation and anisotropic scale of Sim(3) fall along directions that 2D correspondences already constrain (Box 1).

Box 1. Structure of the residual error

1. **Only one dimension is strictly unobservable.** Moving a point along the view direction leaves its projection unchanged, so each frame and group has one radial scale that no 2D observation constrains.
2. **The other three translational degrees of freedom are only weakly constrained.** Translation changes the projection, so it is observable, but with sensitivity proportional to Z/f. Fitting them jointly injects error into directions the observations already constrain.
3. **Degrees of freedom are counted per group, not per point.** A rigid group contributes a single radial scale regardless of how many points it contains, so a sequence with G groups has G scales rather than one global factor.
4. **Correction can therefore only be partial.** Metrically pretrained models constrain the view direction weakly rather than leaving it free, so the repairable fraction is bounded.

Formal statements and proofs are in Appendix A.2.

4 GAUGE

Section 3 measured three properties of the residual. GAUGE answers each with one component, grouping for the alignment of the bias across motion groups, a sparse anchor budget for the missing metric scale, and a per-group radial correction for the concentration along the view direction.

4.1 Minimal Tracker Interface

Given a video of T frames and N query points, a feed-forward tracker outputs 3D trajectories $\mathbf{P} \in \mathbb{R}^{T\times N\times 3}$ and visibility $\mathbf{V} \in \{0,1\}^{T\times N}$. We modify no weight of the tracker and design only a post-hoc geometric head that maps $\mathbf{P}$ to corrected trajectories $\hat{\mathbf{P}}$. The inputs are the globally aligned predicted trajectories, visibility, the per-frame camera center C_t from the datasets' ground-truth extrinsics, and the ground-truth 3D positions $\mathbf{Q}$ of a very small number of query points used as

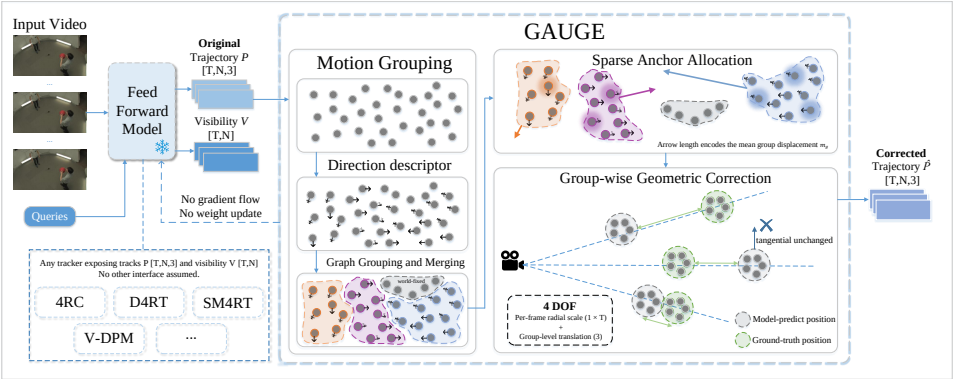


Figure 3: Pipeline of GAUGE. Any tracker that outputs trajectories and visibility can plug in without retraining. The three blocks in GAUGE show main steps of correction.

anchors. The pipeline has three steps: first partition the query points into approximately rigid groups by motion-direction consistency (4.2), then allocate and sample sparse anchors per group (4.3), and finally estimate a low-degree-of-freedom geometric transform per group and apply it to all points in the group (4.4). Pseudocode and implementation details for all steps are given in Appendix A.1.

The method is evaluated on all query points under both protocols of Section 3.1, with static points on the world-fixed branch and ungrouped points keeping the baseline prediction. Figure 3 shows the pipeline and the interfaces between the steps.

4.2 Unsupervised Motion Grouping

The goal of grouping is to make the points within each group undergo approximately the same rigid motion, because it is along such groups that the scale bias aligns (Section 3.2).

Static point detection. A point whose coordinates have standard deviation $\sigma_i < \tau_{\text{static}}$ over its visible frames is static and joins the world-fixed class, which covers the background and barely moving objects.

Motion direction descriptor. For each non-static point, collect the velocities $v_i^t = \mathbf{P}_i^{t+1} - \mathbf{P}_i^t$ over consecutive co-visible frame pairs, keep the unit directions $\hat{\mathbf{v}}_i^t$ on the frame set $\mathcal{F}^i$, and resample to at most K_d reliable directions, forming the descriptor $\mathbf{D}_i$. Direction is unreliable on points with small motion, so a spatial branch is needed as a second filter.

Grouping starts from a graph on the dynamic query points, whose edges are

$$(i,j) \in E \iff |\mathcal{F}_{ij}| \geq \tau_{\text{cov}} \ \wedge \ \frac{1}{|\mathcal{F}_{ij}|} \sum_{t \in \mathcal{F}_{ij}} \hat{\mathbf{v}}_i^t \cdot \hat{\mathbf{v}}_j^t \geq \tau_{\text{dir}}, \tag{1}$$

where $\mathcal{F}_{ij} = \mathcal{F}^i \cap \mathcal{F}^j$ is the set of shared frames and the test is evaluated only inside a spatial kNN neighborhood. The connected components of E form the initial co-moving groups. Distant points may be different objects moving in the same direction, so each group is split once more by spatial kNN components at its 3D median positions, and groups above the size threshold are merged by centroid distance and direction consistency, which repairs fragments cut by occlusion or brief tracking failure. Remaining small components are assigned to the independent-dynamic class.

The result is three group types, world-fixed, co-moving and independent-dynamic. Independent-dynamic points take no part in correction.

4.3 Sparse Anchor Allocation

The total budget is $K = \lfloor \rho N \rfloor$ with $\rho = 0.05$. Each group is scored by $s_g = n_g(1 + m_g / \max_{g'} m_{g'})$, where n_g is its size and m_g its mean 3D displacement, and receives $k_g = \lfloor K s_g / \sum_{g'} s_{g'} \rfloor$ anchors, largest-remainder rounded and capped at n_g. Against ungrouped uniform allocation this changes endpoint error by 24.3 and 30.2 percentage points. Other controls are in Appendix C.1.

4.4 Group-Wise Geometric Correction

Basis for radial correction. The residual error lies mostly along the view direction (Section 3.2), so correction is restricted to the camera ray direction, which leaves the projection unchanged and preserves the tangential component, since an anisotropic scale would break local rigidity.

Choice of scaling center. Radial direction is defined with respect to the camera, so dynamic groups are scaled about the camera center C_t. The alternative form that takes an in-group anchor as origin is clearly worse under the Full protocol and no worse under DQS (Section 5.4), and is reserved for the two special group types in Section 4.5.

For ordinary co-moving groups the correction is a radial rescale about the camera center followed by one group-level translation, and both scales involved have the same form. For a center C and a set $\mathcal{S}$ of anchor point-frame pairs, the scale is the median ratio of the ground-truth to the predicted distance from that center,

$$\alpha(C, \mathcal{S}) = \underset{(i,t)\in\mathcal{S}}{\text{median}} \frac{\|\mathbf{Q}_i^t - C\|}{\|\mathbf{P}_i^t - C\|}. \tag{2}$$

With $C = C_{t_0}$, the camera center of the frame of the first anchor, and $\mathcal{S} = \mathcal{A}_g$, it gives the group-level scale α_{global}, which rescales the group about C_{t_0}. With $C = C_t$ and $\mathcal{S}$ restricted to the anchors that are visible at frame t, it gives the per-frame scale α_t, which is estimated from the raw prediction rather than from the group-level residual and carries the per-frame drift. Frames with too few valid anchors are interpolated and filled with α_{global}. A constant translation, the median residual of the scaled anchors over all frames, $\boldsymbol{\delta} = \text{median}_{(i,t)\in\mathcal{A}_g}(\mathbf{Q}_i^t - \mathbf{P}_i^t)$, is then added to every point of the group and absorbs the rigid translational bias that scaling leaves.

4.5 Special Groups and Degenerate Cases

Both special types use the same estimator centered on the first in-group anchor instead of the camera center. The reason is numerical solvability rather than geometric structure, because the large depth spread of static points and the near-collinearity of low-parallax points make radial scaling about the camera center ill-posed. On failure the correction falls back to the Umeyama Sim(3) (Umeyama, 1991) or a pure translation, and a group that fails both is left uncorrected, as detailed in Appendix A.1.

5 Experiments

5.1 Experimental Setup

Datasets. We evaluate on two benchmarks. PointOdyssey (Zheng et al., 2023) provides complete depth, occlusion and pose ground truth, so the diagnosis and all ablations are carried out on it. WorldTrack (Feng et al., 2025) evaluates tracking in the world coordinate frame on four test sets, Aria Digital Twin (Pan et al., 2023), Panoptic Studio (Joo et al., 2015), PointOdyssey and Dynamic Replica (Karaev et al., 2023). PO appears both under its own benchmark protocol and as a WorldTrack subset under the unified world-frame protocol. Dataset construction is described in Appendix C.6.

Baseline models. We evaluate eight trackers or pipelines covering the feed-forward, depth-assisted and camera-motion-aware paradigms. The feed-forward models are D4RT, 4RC, V-DPM and SM4RT (Zhang et al., 2026; Luo et al., 2026; Sucar et al., 2026; Lin et al., 2026). TAPIP3D (Zhang et al., 2025a) requires RGB-D input, SpatialTrackerV2 (Xiao et al., 2025) estimates camera motion explicitly, DELTA (Ngo et al., 2025) tracks densely on estimated depth, and CoTracker3 (Karaev et al., 2025) with DepthAnything v2 (Yang et al., 2024) forms a 2D-tracking-plus-depth pipeline. D4RT is not open-sourced and is evaluated through the OpenD4RT (Li & RHOS Team, 2026) reimplementation.

Evaluation protocol. Following TAPVid-3D (Koppula et al., 2024), the metrics are end-point error (EPE, ↓), Average Percent of Points within Delta (APD, ↑) and 3D Average Jaccard (AJ, ↑), whose computation is given in Appendix A.4. DQS (Section 3.1) evaluates only dynamic query points, while the Full protocol uses all query points visible in the first frame, and these include a large amount of static background. Improvement is the relative reduction in endpoint error with respect to applying no correction. Values in each table are computed from the absolute values in that table. Table cells report DQS / Full. Table 3 instead reports four-subset averages under DQS. DQS screens query points for non-static motion, model-reported confidence and multi-frame visibility. The three filters share one threshold set across all models and benchmarks (values and exact definitions in Appendix A.5). A control replacing the predicted screening quantity with ground-truth motion leaves the gain almost unchanged (Appendix C.4). The anchor budget defaults to 5%. The budget curve and the sampling dispersion are analyzed in Section 5.4 and Appendix B.2.

Table 2: Main results on PointOdyssey. The Sim(3) column is the seven-degree-of-freedom fit under the same grouping and anchors. Cell values are DQS / Full; improvements (%) are in parentheses.

Model	Method	PointOdyssey			EPE–Sim(3)
		EPE↓	APD↑	AJ↑	Correction
D4RT	Base	0.1812 / 0.1930	0.6877 / 0.6654	0.1038 / 0.3358	—
	+GAUGE	0.1269 / 0.1597 (**+29.9 / +17.3**)	0.7648 / 0.7015 (**+11.2 / +5.4**)	0.1214 / 0.3320 (**+17.0** / -1.1)	+15.7 / -2.8
4RC	Base	0.1421 / 0.2471	0.7334 / 0.6127	0.0711 / 0.2603	—
	+GAUGE	0.1035 / 0.1470 (**+27.2 / +40.5**)	0.7832 / 0.7332 (**+6.8 / +19.7**)	0.0892 / 0.3172 (**+25.5 / +21.9**)	-29.6 / +12.8
V-DPM	Base	0.1451 / 0.0778	0.7322 / 0.8792	0.1164 / 0.4478	—
	+GAUGE	0.1086 / 0.0619 (**+25.2 / +20.4**)	0.8091 / 0.9077 (**+10.5 / +3.2**)	0.1164 / 0.4634 (**+0.0 / +3.5**)	+10.8 / +19.9
SM4RT	Base	0.1351 / 0.1538	0.7709 / 0.7478	0.0791 / 0.3594	—
	+GAUGE	0.1000 / 0.0939 (**+26.0 / +38.9**)	0.8010 / 0.8372 (**+3.9 / +12.0**)	0.0851 / 0.3864 (**+7.6 / +7.5**)	-24.9 / -13.5
TAPIP3D	Base	0.1548 / 0.4080	0.7600 / 0.4854	0.1560 / 0.0348	—
	+GAUGE	0.0966 / 0.4023 (**+37.6 / +1.4**)	0.8365 / 0.4978 (**+10.1 / +2.6**)	0.1576 / 0.0415 (**+1.0 / +19.3**)	+14.5 / +2.2
SpatialTrackerV2	Base	2.0526 / 1.3295	0.0306 / 0.0863	0.0411 / 0.1230	—
	+GAUGE	1.7428 / 1.3674 (**+15.1** / -2.9)	0.1164 / 0.1397 (**+280.4 / +61.9**)	0.0521 / 0.1165 (**+26.8** / -5.3)	-1.9 / +6.4
DELTA	Base	0.7013 / 0.5493	0.2867 / 0.3607	0.0700 / 0.1525	—
	+GAUGE	0.2974 / 0.2798 (**+57.6 / +49.1**)	0.5274 / 0.4828 (**+84.0 / +33.9**)	0.1057 / 0.2414 (**+51.0 / +58.3**)	+58.4 / +34.2
CoTracker3+DA	Base	1.6050 / 1.1562	0.0496 / 0.1412	0.0344 / 0.1088	—
	+GAUGE	0.5997 / 0.6016 (**+62.6 / +48.0**)	0.3646 / 0.3436 (**+635.1 / +143.3**)	0.0756 / 0.1665 (**+119.8 / +53.0**)	+62.3 / +42.8

Table 3: WorldTrack results averaged over the four subsets under DQS. Averages are unweighted over the evaluated subsets; per-subset values are in Table A1.

	EPE↓			APD↑			AJ↑		
Model	Base	+Ours	Impr. (%)	Base	+Ours	Impr. (%)	Base	+Ours	Impr. (%)
D4RT	0.1345	0.0838	**+37.7**	0.7783	0.8650	**+11.1**	0.0180	0.0207	**+15.0**
4RC	0.1531	0.1154	**+24.6**	0.7359	0.7970	**+8.3**	0.0128	0.0144	**+12.5**
V-DPM	0.1087	0.0758	**+30.3**	0.8082	0.8491	**+5.1**	0.0587	0.0646	**+10.1**
SM4RT	0.1256	0.0930	**+26.0**	0.7838	0.8331	**+6.3**	0.0154	0.0160	**+3.9**
TAPIP3D	0.0915	0.0623	**+31.9**	0.8351	0.8523	**+2.1**	0.0403	0.0452	**+12.2**
SpatialTrackerV2	1.8069	1.4086	**+22.0**	0.0483	0.2149	**+344.9**	0.0455	0.0592	**+30.1**
DELTA	0.4563	0.1974	**+56.7**	0.5404	0.6966	**+28.9**	0.0232	0.0293	**+26.3**
CoTracker3+DA	0.9559	0.3781	**+60.4**	0.2690	0.5463	**+103.1**	0.0220	0.0294	**+33.6**

5.2 Main Results

Under DQS, endpoint error drops on all five test splits for all eight models (Table 2 and Table 3), by 15.1% to 62.6% on PointOdyssey and by 4.9% to 76.3% on the four WorldTrack subsets. Negative gains occur only under the Full protocol, where the static background dominates and the world-fixed branch of Section 4.5 is misspecified for static points. The worst is -15.6% (Table A1).

The seven-degree-of-freedom Sim(3) comparison is the most direct conclusion. On most models, four degrees of freedom beat Sim(3) under the same grouping and the same anchors, and on four models Sim(3) is worse than applying no correction at all (last column of Table 2). Sim(3) wins in three settings, exactly those where the direction signal is weakest or the grouping loosest, consistent with Sections 5.5 and 5.7.

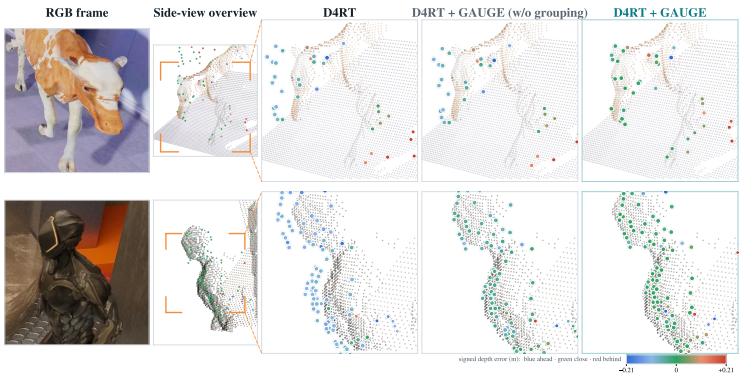


Figure 4: Qualitative comparison on two scenes. Query points are colored by signed depth error. Without grouping the errors scatter ahead of and behind the surface, while D4RT + GAUGE concentrates them better.

Trajectory accuracy improves as well, with magnitudes set by the baseline. Models with a high baseline APD improve moderately, while the three starting from near-zero grow several-fold and remain below the first set after correction. Thresholded metrics can drop through boundary-crossing even where continuous error falls, as in the Full-protocol AJ cells of Table 2. A qualitative comparison is shown in Figure 4.

5.3 Comparison with Test-Time Optimization

One direct option is to fine-tune the base model on the same small number of ground-truth observations. Full fine-tuning of models with up to a billion parameters is impractical for a single test

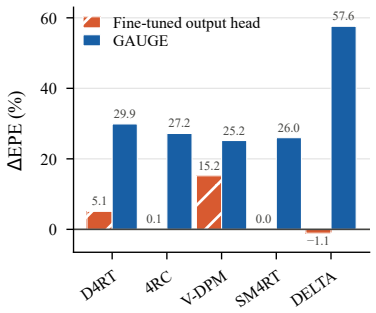


Figure 5: Comparison with TTO.

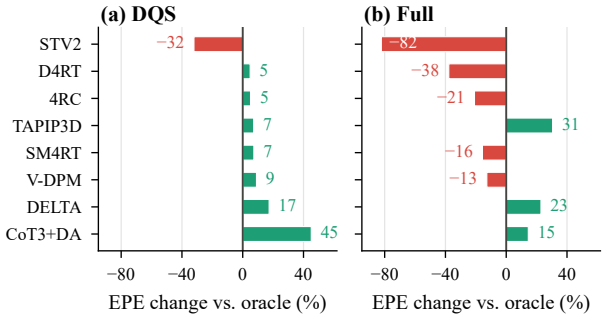


Figure 6: Unsupervised grouping versus oracle instance grouping.

scene, so we freeze the backbone and fine-tune the output head, sweeping the learning rate and reporting the best result per model. The comparison covers the five models whose pipelines support this protocol (Figure 5).

Across five architecturally very different base models, fine-tuning brings a gain of -1.1% to 15.2%, whereas GAUGE achieves 25.2% to 57.6% with exactly the same amount of information. SM4RT already models motion groups explicitly during training, yet fine-tuning leaves its position error unchanged and only visibility-related metrics shift slightly, indicating that written-in motion structure does not remove this family of per-group degrees of freedom.

Fine-tuning gains more under the Full protocol than under DQS, consistent with its improving mainly the majority static background, while GAUGE's gain concentrates on dynamic query points. Values for the Full protocol are given in Appendix Table A2.

5.4 Ablation Studies

Table 4: Ablation studies (D4RT) under the camera-center parameterization; the 4RC version is in Appendix Table A10.

(a) Grouping strategy

Grouping	EPE ↓	APD ↑	AJ ↑
Oracle instance	0.1334 / 0.1159	0.7518 / 0.7717	0.1260 / 0.3913
Direction KNN only	0.1274 / 0.1606	0.7692 / 0.7006	0.1210 / 0.3328
Spatial merge only	0.1271 / 0.1602	0.7693 / 0.6992	0.1213 / 0.3322
Full (Ours)	**0.1269 / 0.1597**	**0.7648 / 0.7015**	**0.1214 / 0.3320**

(b) Degrees of freedom of the correction

Configuration	EPE ↓
Scale only (1-DOF)	0.1792 / 0.1868
Scale + translation	0.1624 / 0.1810
Scale about anchor (1-DOF)	0.1706 / 0.2068
Scale about anchor + translation	0.1535 / 0.1958
Single form on all groups	0.1568 / 0.1663
Ours (per-frame scale + translation)	**0.1269 / 0.1597**

Grouping strategies. Introducing grouping lowers the error substantially, while differences between grouping strategies are much smaller. The complete scheme combining direction clustering and spatial merging is best.

Degrees of freedom of the correction. Under the camera-center parameterization the error decreases monotonically with the degrees of freedom (Table 4b). The decisive step is per-frame scale, which shows that the time-varying drift it captures cannot be covered by a group-level constant. Scaling about an in-group anchor and a single form shared by all groups are both inferior to the complete scheme.

Anchor budget. Endpoint error decreases as the budget grows, with the optimal budget anywhere from 2% to 100% by model (Appendix Table A5 gives eight models, both protocols, seven levels). Across all model-protocol combinations, the median difference between the 100% and 50% levels is -0.0000, with a single significant setting (SM4RT under the Full protocol), so the high-budget range saturates rather than degrades. The choice of 5% rests not on accuracy but on per-anchor efficiency. The average accumulated per-anchor efficiency falls from 22.8 at 1% to 6.0 at 5% and 0.7 at 50%, after which the cost is no longer proportional (Appendix Figure A3).

5.5 Unsupervised Grouping versus Oracle Instance Grouping

Under DQS, unsupervised grouping gives lower endpoint error on seven of the eight models, with a median relative change of +7.3%. Motion clustering produces finer groups than semantic instances, giving more consistent within-group scale bias. Extending to the Full protocol divides the models. On five of the eight, oracle instance grouping is better, with a median relative change of -14.2%, because oracle segmentation is correct for the static background that dominates Full. On the other

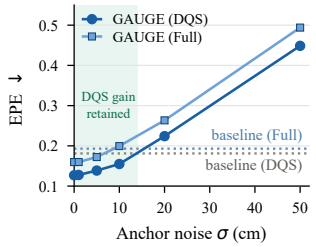


Figure 7: Anchor noise robustness.

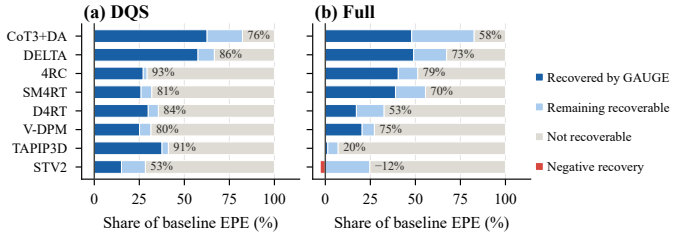


Figure 8: Error budget. Values in Table A14.

three, unsupervised grouping is better under both protocols. A per-model comparison is shown in Figure 6, and the complete set of values is given in Appendix Table A6.

5.6 Robustness to Imperfect and Non-Ground-Truth Anchors

We apply increasing Gaussian noise to the ground-truth anchor positions. At 10 cm the corrected endpoint error is still clearly better than the uncorrected baseline. At 20 cm the gain disappears. This sets a clear bound for practical use, and the depth accuracy of sparse LiDAR returns, visual SLAM keyframes and consumer RGB-D all fall within this range (Figure 7, values in Table A4).

The method needs an external source of true scale, so we also evaluate it with anchors drawn from observable sources instead of ground truth. Table 5 reports the four sources.

Table 5: Effect of anchor source on the gain (D4RT).

Anchor source	Median error (cm)	Radial fraction	ΔEPE
Ground truth	0	—	**29.9%**
Visual SLAM keyframes	8.57	0.49	**20.6%**
Single-frame depth	8.50	1	**10.1%**
Sparse LiDAR returns	0.81	1	**29.2%**

All three sources are usable, with gains not explained by error magnitude. Visual SLAM keyframes give an anchor error of 8.57 cm, a radial fraction of 0.49 and a gain of 20.6%. Single-frame depth gives almost the same anchor error of 8.50 cm but a radial fraction of 1.00 and a gain of only 10.1%, because its errors fall entirely on the estimated dimension and contaminate the radial scale, whereas nearly half of the SLAM error is tangential and absorbed by the median and by anchor aggregation. The structure of anchor error matters more than its magnitude.

5.7 Error Budget and Failure Prediction

The upper bound is fit on all query points with the same parameterization as GAUGE and takes the lower of two minima, making the ratio conservative. The family characterized in Section 3 explains at most 34.1% of the total error. GAUGE recovers 28.5 percentage points of it, a recovery equal to 75.6% of that upper bound. All three are medians, with ranges given in Appendix C.7. The remainder is non-rigid deformation and sensor noise, outside four degrees of freedom, consistent with Box 1, and the three fractions are shown in Figure 8.

Correlating the three structural metrics with the gain over all available settings, only between-group scale variance has predictive power ($r = 0.61$, $n = 58$, permutation test $p < 0.0001$) and can identify in advance the settings with the weakest or negative gain (Appendix C.6). The other two metrics predict no such gain, but they are indispensable to the construction of the method, quantifying where the error concentrates and how reliable motion direction is (Section 3).

Cost. On a single CPU core the correction takes a median of 0.49 s per sequence under DQS and 2.60 s under the Full protocol, and adds 0.74 MiB and 29.7 MiB to a peak resident memory of 594 and 770 MiB, and requires no GPU, backpropagation or optimizer (Appendix Table A9).

6 Conclusion and Future Work

We measure the residual error across eight trackers and benchmarks and find it structured rather than random. The bias is shared within motion groups, and most of the residual energy lies along the view direction, where projection alone cannot recover it. Our correction module, GAUGE, corrects this bias with four degrees of freedom per group and frame, without training, and endpoint error on dynamic query points drops by 15.1% to 62.6%.

Three boundaries remain, and each suggests a concrete direction. The method requires true-scale anchors and the gain disappears beyond about 20 cm of anchor noise. Groups are assumed to stay rigid and estimates are solved per frame, so non-rigid deformation, sensor noise and long-sequence drift are out of scope. Future work can focus on addressing these limitations.

AI USE STATEMENT

We used generative AI tools to assist with translation, polishing and LaTeX typesetting. The research question, the method, the experiments and the analysis are the authors' own. All AI-assisted text was verified sentence by sentence against the Chinese manuscript, and we take full responsibility for the final content.

ETHICS STATEMENT

This paper does not raise ethical concerns.

REPRODUCIBILITY STATEMENT

Full hyperparameters, pseudocode of the complete pipeline and all numerical tables are given in the appendix, and the experimental setup is described in Section 5. The complete implementation of the correction head is included in the supplementary material. It depends only on numpy and scipy, runs on a single CPU core, and includes a synthetic smoke test.

REFERENCES

Hao Ai, Wenjie Chang, Jianbo Jiao, Aleš Leonardis, and Ofek Eyal. Articulation in motion: Prior-free part mobility analysis for articulated objects by dynamic-static disentanglement, 2026. arXiv:2603.02910.

Stefan Andreas Baur, David Josef Emmerichs, Frank Moosmann, Peter Pinggera, Björn Ommer, and Andreas Geiger. SLIM: Self-supervised LiDAR scene flow and motion segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13126–13136, 2021.

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 (ICLR)*, pp. 75602–75637, 2025.

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.

Chong Cheng, Sicheng Yu, Zijian Wang, Yifan Zhou, and Hao Wang. Outdoor monocular SLAM with global scale-consistent 3D gaussian pointmaps. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 26035–26044, 2025.

Jakob Engel, Vladlen Koltun, and Daniel Cremers. Direct sparse odometry. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(3):611–625, 2018.

Haiwen Feng, Junyi Zhang, Qianqian Wang, Yufei Ye, Pengcheng Yu, Michael J. Black, Trevor Darrell, and Angjoo Kanazawa. St4RTrack: Simultaneous 4D reconstruction and tracking in the world. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8503–8513, 2025.

Zan Gojcic, Or Litany, Andreas Wieser, Leonidas J. Guibas, and Tolga Birdal. Weakly supervised learning of rigid 3D scene flow. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5692–5703, 2021.

Eunsoo Im, Gyeonggwan Lee, Seunghwan Hong, and Junghun Suh. PRISM-SLAM: Probabilistic ray-grounded inference for scale-aware metric SLAM, 2026. arXiv:2605.19257.

Yusuke Iwasawa and Yutaka Matsuo. Test-time classifier adjustment module for model-agnostic domain generalization. In *Advances in Neural Information Processing Systems*, volume 34, pp. 2427–2440. Curran Associates, Inc., 2021.

Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3334–3342, 2015.

Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. DynamicStereo: Consistent dynamic depth from stereo videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13229–13239, 2023.

Nikita Karaev, Yuri Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. CoTracker3: Simpler and better point tracking by pseudo-labelling real videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6013–6022, 2025.

Skanda Koppula, Ignacio Rocco, Yi Yang, Joe Heyward, João Carreira, Andrew Zisserman, Gabriel Brostow, and Carl Doersch. TAPVid-3D: A benchmark for tracking any point in 3D. In *Advances in Neural Information Processing Systems*, volume 37, pp. 82149–82165. Curran Associates, Inc., 2024.

Vincent Leroy, Yohann Cabon, and Jerome Revaud. Grounding image matching in 3D with MASt3R. In *Computer Vision – ECCV 2024*, pp. 71–91. Springer Nature Switzerland, 2024.

Jiaxin Li and RHOS Team. OpenD4RT: An unofficial PyTorch/GPU implementation of D4RT for 4D reconstruction and tracking. `https://github.com/Lijiaxin0111/Open-d4rt`, 2026. Accessed: 2026-09-15.

Yuanyang Li and Matthias Althoff. Learning image-adaptive scale fields for metric depth recovery, 2026. arXiv:2605.07418.

Zhengqi Li, Richard Tucker, Forrester Cole, Qianqian Wang, Linyi Jin, Vickie Ye, Angjoo Kanazawa, Aleksander Holynski, and Noah Snavely. MegaSaM: Accurate, fast and robust structure and motion from casual dynamic videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10486–10496, 2025.

Shing Ho J. Lin, Wenzhao Zheng, Dong Zhuo, Yuqi Wu, Jie Zhou, and Jiwen Lu. SM4RT: Learning structured motion geometry for 4D reconstruction, 2026. arXiv:2607.22534.

Yihang Luo, Shangchen Zhou, Yushi Lan, Xingang Pan, and Chen Change Loy. 4RC: 4D reconstruction via conditional querying anytime and anywhere, 2026. arXiv:2602.10094.

Riku Murai, Eric Dexheimer, and Andrew J. Davison. MASt3R-SLAM: Real-time dense SLAM with 3D reconstruction priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16695–16705, 2025.

Tuan Duc Ngo, Peiye Zhuang, Chuang Gan, Evangelos Kalogerakis, Sergey Tulyakov, Hsin-Ying Lee, and Chaoyang Wang. DELTA: Dense efficient long-range 3D tracking for any video, 2025. arXiv:2410.24211.

Xiaqing Pan, Nicholas Charron, Yongqian Yang, Scott Peters, Thomas Whelan, Chen Kong, Omkar Parkhi, Richard Newcombe, and Yuheng (Carl) Ren. Aria digital twin: A new benchmark dataset for egocentric 3D machine perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20133–20143, 2023.

Sohag Roy, Rajesh Misra, Swami Shastravidyananda, and Tamal Maharaj. The multipath blind spot: k-agnostic robust calibration for sparse-anchor metric depth from frozen foundations, 2026. arXiv:2607.04101.

Edgar Sucar, Eldar Insafutdinov, Zihang Lai, and Andrea Vedaldi. V-DPM: 4D video reconstruction with dynamic point maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14502–14511, 2026.

Zachary Teed and Jia Deng. DROID-SLAM: Deep visual SLAM for monocular, stereo, and RGB-D cameras, 2022. arXiv:2108.10869.

Bill Triggs, Philip F. McLauchlan, Richard I. Hartley, and Andrew W. Fitzgibbon. Bundle Adjustment — A Modern Synthesis. In *Vision Algorithms: Theory and Practice*, pp. 298–372, Berlin, Heidelberg, 2000. Springer Berlin Heidelberg.

Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. volume 13, pp. 376–380, 1991.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization, 2021. arXiv:2006.10726.

Haowen Wang, Xiaoping Yuan, Zhao Jin, Zhen Zhao, Zhengping Che, Yousong Xue, Jin Tian, Yakun Huang, and Jian Tang. PD^{2}GS: Part-level decoupling and continuous deformation of articulated objects via gaussian splatting, 2026. arXiv:2506.09663.

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, 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 (CVPR)*, pp. 20697–20709, 2024.

Diana Wofk, René Ranftl, Matthias Müller, and Vladlen Koltun. Monocular visual-inertial depth estimation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6095–6101, 2023.

Yuxi Xiao, Jianyuan Wang, Nan Xue, Nikita Karaev, Yuri Makarov, Bingyi Kang, Xing Zhu, Hujun Bao, Yujun Shen, and Xiaowei Zhou. SpatialTrackerV2: Advancing 3D point tracking with explicit camera motion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6726–6737, 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*, volume 37, pp. 21875–21911. Curran Associates, Inc., 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 (ICCV)*, pp. 9043–9053, 2023.

Bowei Zhang, Lei Ke, Adam W. Harley, and Katerina Fragkiadaki. TAPIP3D: Tracking any point in persistent 3D geometry. *arXiv preprint arXiv:2504.14717*, 2025a.

Chuhan Zhang, Guillaume Le Moing, Skanda Koppula, Ignacio Rocco, Liliane Momeni, Junyu Xie, Shuyang Sun, Rahul Sukthankar, Joëlle K. Barral, Raia Hadsell, Zoubin Ghahramani, Andrew Zisserman, Junlin Zhang, and Mehdi S. M. Sajjadi. Efficiently reconstructing dynamic scenes one D4RT at a time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7382–7392, 2026.

Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. MonST3R: A simple approach for estimating geometry in the presence of motion, 2025b. arXiv:2410.03825.

Marvin Zhang, Sergey Levine, and Chelsea Finn. MEMO: Test time robustness via adaptation and augmentation. In *Advances in Neural Information Processing Systems*, volume 35, pp. 38629–38642. Curran Associates, Inc., 2022.

Qingwen Zhang, Yi Yang, Peizheng Li, Olov Andersson, and Patric Jensfelt. SeFlow: A self-supervised scene flow method in autonomous driving. In *Computer Vision – ECCV 2024*, pp. 353–369. Springer Nature Switzerland, 2024a.

Wei Zhang, Tiecheng Sun, Sen Wang, Qing Cheng, and Norbert Haala. HI-SLAM: Monocular real-time dense mapping with hybrid implicit fields. *IEEE Robotics and Automation Letters*, 9(2): 1548–1555, 2024b.

Yang Zheng, Adam W. Harley, Bokui Shen, Gordon Wetzstein, and Leonidas J. Guibas. PointOdyssey: A large-scale synthetic dataset for long-term point tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19855–19865, 2023.

Jia-Xing Zhong, Ta-Ying Cheng, Yuhang He, Kai Lu, Kaichen Zhou, Andrew Markham, and Niki Trigoni. Multi-body SE(3) equivariance for unsupervised rigid segmentation and motion estimation. In *Advances in Neural Information Processing Systems*, volume 36, pp. 76085–76097. Curran Associates, Inc., 2023.

APPENDIX

A METHOD DETAILS AND FORMAL STATEMENTS

A.1 COMPLETE PIPELINE

Algorithm 1 gives the complete pipeline, calling in order the grouping of Algorithm 2, the anchor allocation and sampling of Algorithm 3, and the two forms of per-group correction in Algorithm 4 and Algorithm 5. The inputs are the globally aligned trajectories $\mathbf{P}$, the visibility mask $\mathbf{V}$, the per-frame camera centers $\{\mathbf{C}_t\}$, and the ground-truth trajectories $\mathbf{Q}$ of the anchors. The output is the corrected trajectory $\hat{\mathbf{P}}$.

Algorithm 1: GAUGE: overall pipeline.

Require: tracks $\mathbf{P} \in \mathbb{R}^{T\times N\times 3}$, visibility $\mathbf{V} \in \{0,1\}^{T\times N}$, camera centers $\{\mathbf{C}_t\}_{t=1}^{T}$,
Require: ground-truth trajectories $\mathbf{Q}$ of the anchor points, anchor budget ρ
Ensure: Corrected tracks $\hat{\mathbf{P}}$

$(\mathcal{G}, \mathcal{U}) \leftarrow$ MOTIONGROUPING$(\mathbf{P}, \mathbf{V})$ {Algorithm 2; $\mathcal{U}$: ungrouped points}
$(\{k_g\}, \{\mathcal{A}_g\}) \leftarrow$ ALLOCATEANDSAMPLEANCHORS$(\mathcal{G}, \rho)$ {Algorithm 3}
$\hat{\mathbf{P}} \leftarrow \mathbf{P}$ {ungrouped points and skipped groups keep the baseline}
for each group $g \in \mathcal{G}$ **do**
  if g is `independent_dynamic` **or** $\mathcal{A}_g = \emptyset$ **then**
    continue
  end if
  if g is `world_fixed` **then**
    $\hat{\mathbf{P}}[:, g] \leftarrow$ CORRECTABOUTANCHOR$(\mathbf{P}[:, g], \mathcal{A}_g)$ {Algorithm 5}
  else
    $\hat{\mathbf{P}}[:, g] \leftarrow$ CORRECTRADIALPERFRAME$(\mathbf{P}[:, g], \mathcal{A}_g, \{\mathbf{C}_t\})$ {Algorithm 4}
  end if
end for
return $\hat{\mathbf{P}}$

The cost of grouping is $O(Nk_{\text{knn}}K_d)$, because direction similarity is computed only within a spatial kNN neighborhood. Correction is $O(TN)$. The whole pipeline contains no backpropagation and no learnable parameters, and the per-group fitting uses only medians and ratios.

Algorithm 2 corresponds to the five steps of Section 4.2. The default hyperparameters are listed in Section A.5.

Algorithm 2: Unsupervised motion grouping.

$\bar{\mathbf{p}}^i \leftarrow \text{median}_t\{\mathbf{P}_i^t \mid V_i^t = 1\}$ {representative position}
$\mathcal{W} \leftarrow \{i \mid \text{std}_t(\mathbf{P}_i^t \mid V_i^t = 1) < \tau_{\text{static}},\ \sum_t V_i^t \geq 5\}$ {world-fixed}
for each $i \notin \mathcal{W}$ **do**
  $\mathcal{M}^i \leftarrow \{\mathbf{P}_i^{t+1} - \mathbf{P}_i^t \mid V_i^t = V_i^{t+1} = 1,\ \|\mathbf{P}_i^{t+1} - \mathbf{P}_i^t\|_2 > \tau_{\text{motion}}\}$
  if $|\mathcal{M}^i| \geq 3$ **then**
    $\mathcal{D}^i \leftarrow$ at most K_d representative directions drawn from $\mathcal{M}^i$,
      carried on a frame set $\mathcal{F}^i$ {direction descriptor}
  else
    mark i ungrouped {no descriptor; left uncorrected}
  end if
end for
build a kNN graph over the points carrying a descriptor ($k = 50$)
connect $i \sim j$ if $|\mathcal{F}_{ij}| \geq \tau_{\text{cov}}$ and $\frac{1}{|\mathcal{F}_{ij}|}\sum_{t\in\mathcal{F}_{ij}} \hat{\mathbf{v}}_i^t \cdot \hat{\mathbf{v}}_j^t \geq \tau_{\text{dir}}$
  where $\mathcal{F}_{ij} \leftarrow \mathcal{F}^i \cap \mathcal{F}^j$ {frames on which both descriptors are defined}
take connected components; components of size at least 3 become co-move groups, the rest are `independent_dynamic`
for each co-move group **do**
  build a spatial kNN graph over $\{\bar{\mathbf{p}}^i\}$ ($k = 10$); keep spatially connected components of size at least 3, demote the rest
end for
for each pair of co-move groups of size at least 5 **do**
  merge the pair if the centroid distance is below $3.0\times$ the local scale
    and the descriptors of the two groups satisfy the same test with $\tau_{\text{dir}}^{\text{merge}}$
end for

split $\mathcal{W}$ into spatially connected static subgroups; merge subgroups below 5 points into the nearest static subgroup
merge dynamic fragments below 5 points into the nearest co-move group within the distance threshold; fragments without such a target become `independent_dynamic`
return $\mathcal{G}$, and the points left ungrouped

Algorithm 3 corresponds to Section 4.3. Allocation uses the largest-remainder method to guarantee that the total is exactly K and that $k_g \leq n_g$. Sampling runs over candidates for which predicted and ground-truth visibility both hold, and prefers co-frame anchors, because co-frame anchors share the same camera center and object pose.

Algorithm 3: Anchor allocation and sampling.

$K \leftarrow \lfloor \rho N \rfloor$
for each group g **do**
  $n_g \leftarrow$ number of queries of group g valid in at least one frame; $m_g \leftarrow$ mean 3D displacement over the group
  $s_g \leftarrow n_g\big(1 + m_g / \max_{g'} m_{g'}\big)$ {motion-weighted score}
end for
$k_g \leftarrow \lfloor K s_g / \sum_{g'} s_{g'} \rfloor$, capped at n_g for every g
  distribute the remainder by decreasing fractional part until $\sum_g k_g = \min(K, \sum_g n_g)$
for each group g **do**
  if there exists a frame at which at least k_g queries of group g are valid **then**
    draw such a frame $t^\star$ uniformly; draw k_g queries of group g at $t^\star$ without replacement; $\mathcal{A}_g \leftarrow \{(i, t^\star) : i \in$ the drawn queries$\}$
  else
    draw k_g queries of group g without replacement, and for each one a valid frame {no single frame carries all k_g anchors}
  end if
end for
return $\{k_g\}, \{\mathcal{A}_g\}$

Algorithm 4 corresponds to the three steps of Section 4.4. The α_t of the second step is re-estimated from the raw prediction $\mathbf{P}$ rather than from the residual of the first step, so the two steps are estimated independently and then applied in cascade. Missing frames are first linearly interpolated and then smoothed with a Gaussian kernel at $\sigma = 2$, and positions still missing fall back to α_{global}. The last line gives the fallback chain, which runs from the form with fewer degrees of freedom that fits the error structure to the form with more degrees of freedom that is generic and finally to a pure translation, so the degrees of freedom rise at the first step and fall at the last.

Algorithm 4: Group-wise geometric correction for co-move groups.

$t_0 \leftarrow$ frame of the first anchor; $\mathbf{C} \leftarrow \mathbf{C}_{t_0}$
$\alpha_{\text{global}} \leftarrow \text{median}_{(i,t)\in\mathcal{A}_g} \|\mathbf{Q}_i^t - \mathbf{C}\|_2 \,/\, \|\mathbf{P}_i^t - \mathbf{C}\|_2$ {global radial scale about the reference camera center}
$\tilde{\mathbf{P}}_i^t \leftarrow \mathbf{C} + \alpha_{\text{global}}\big(\mathbf{P}_i^t - \mathbf{C}\big)$
for $t = 1, \ldots, T$ **do**
  if $\exists\, i \in \mathcal{I}^g_{\text{anchor}}$ valid at frame t **then**
    $\alpha_t \leftarrow \text{median}_{i\in\mathcal{I}^g_{\text{anchor}},\ V_i^t = G_i^t = 1} \|\mathbf{Q}_i^t - \mathbf{C}_t\|_2 \,/\, \|\mathbf{P}_i^t - \mathbf{C}_t\|_2$
      computed per frame from the raw prediction
  else
    $\alpha_t \leftarrow$ missing
  end if
end for
interpolate the missing α_t, smooth with a Gaussian kernel ($\sigma = 2$), and fill any remaining gap with α_{global}
$\bar{\mathbf{P}}_i^t \leftarrow \mathbf{C}_t + \alpha_t\big(\tilde{\mathbf{P}}_i^t - \mathbf{C}_t\big)$ {per-frame radial scale}
$\boldsymbol{\delta} \leftarrow \text{median}_{(i,t)\in\mathcal{A}_g}\big(\mathbf{Q}_i^t - \bar{\mathbf{P}}_i^t\big)$; $\hat{\mathbf{P}} \leftarrow \bar{\mathbf{P}} + \boldsymbol{\delta}$ {group-wise translation}
fallback: fit a Sim(3) transform by Umeyama when at least three non-collinear anchors are available,
  then a pure translation when at least one remains, and otherwise leave the group uncorrected
return $\hat{\mathbf{P}}$

Algorithm 5 is the special-group form of Section 4.5, used for world-fixed static groups and, in the low-parallax fallback, for groups with too little parallax. It takes the first in-group anchor as the

local origin, so the scale is no longer strictly along the view direction, in exchange for numerical stability. If s cannot be estimated it falls back to a pure translation.

Algorithm 5: Correction about an anchor for world-fixed and low-parallax groups.

1: $a \leftarrow$ first anchor of $\mathcal{A}_g$
2: $s \leftarrow \mathrm{median}_{(i,t)\in\mathcal{A}_g\setminus\{a\}} \|\mathbf{Q}_i^t - \mathbf{Q}^{(0)}\|_2 \,/\, \|\mathbf{P}_i^t - \mathbf{P}^{(0)}\|_2$
3: $\mathbf{P}^{(0)}, \mathbf{Q}^{(0)}$: prediction and ground truth of a
4: $\bar{\mathbf{P}}_i^t \leftarrow \mathbf{Q}^{(0)} + s\big(\mathbf{P}_i^t - \mathbf{P}^{(0)}\big)$ {the anchor is matched exactly}
5: $\boldsymbol{\delta} \leftarrow \mathrm{median}_{(i,t)\in\mathcal{A}_g}\big(\mathbf{Q}_i^t - \bar{\mathbf{P}}_i^t\big)$; $\hat{\mathbf{P}} \leftarrow \bar{\mathbf{P}} + \boldsymbol{\delta}$ {same translation step as Algorithm 4, step 11}
6: **fallback:** degenerate to a pure translation if s is not estimable, and leave the group uncorrected if no translation is available either
7: **return** $\hat{\mathbf{P}}$

A.2 Formal Statements

Let C_t be the camera center of frame t, K_t and R_t the intrinsics and rotation, and $u_t(X) \sim K_t R_t(X - C_t)$ the projection of a world point X, where $\sim$ denotes equality up to a positive scale factor. The remaining symbols are listed below.

Symbol	Meaning
T, N, G	Number of frames, query points, and recovered motion groups
t, i, g	Indices of frame, query point, and motion group
$\mathbf{P}_i^t$	Predicted 3D position of query point i at frame t; all predictions are written $\mathbf{P} \in \mathbb{R}^{T\times N\times 3}$
$\mathbf{Q}_i^t$	Ground-truth 3D position of the same point, used only for anchor selection and evaluation
$\bar{\mathbf{P}}_i^t$, $\hat{\mathbf{P}}_i^t$	The intermediate trajectory before the group-level translation, and the final corrected trajectory
V_i^t	Visibility predicted by the model
G_i^t	Ground-truth visibility
$\mathcal{I}^g_{\text{anchor}}$	Set of query points of group g selected as anchors
$\mathcal{A}_g$	Point-frame pairs of group g where predicted and ground-truth visibility both hold
α_{global}	Group-level scale about the camera center C_{t_0} of the reference frame
α_t	Radial scale estimate at frame t, interpolated and smoothed where anchors are missing
$\boldsymbol{\delta}$	Group-level constant translation
ρ	Fraction of query points used as anchors, 0.05 by default
τ_m, τ_{static}, τ_{dir}	Measurement lower bound of the structural quantities, the static-point criterion, the direction-consistency threshold, and τ_{cov}; values in Appendix A.5

Proposition 1 (radial scale is unobservable). For any frame t, any point X and any scalar $\alpha > 0$, let $X' = C_t + \alpha(X - C_t)$; then $u_t(X') = u_t(X)$. Substituting gives $X' - C_t = \alpha(X - C_t)$, and the extra factor α cancels when homogeneous coordinates are normalized. A single-frame observation therefore imposes no constraint on the radial distance from a point to the camera center of that frame. The only thing it constrains is the angular relation between the point and the view direction.

Proposition 2 (structure of the per-group family). Let g be a group of points moving rigidly over time. Let Φ map X_i^t to $s_{g,t}(X_i^t - C_t) + C_t + b_g$, with $G(T+3)$ parameters, where G is the number of groups. Then (i) each element is a similarity transform within the group and preserves the shape of the rigid group. (ii) The subfamily with $b_g = 0$ preserves all projections exactly by Proposition 1 and is the exact null space of the projection observations. (iii) The translational part does change the projection, with sensitivity $\|\partial u/\partial b\|$ proportional to f/Z_g, so distant and low-parallax groups can be constrained only to an accuracy of $O(Z_g\sigma_{\text{px}}/f)$, that is, they are observable but only weakly constrained.

For rigidity, expanding gives $\|\bar{X}_i^t - \bar{X}_j^t\|^2 = \alpha^2 \|X_i^t - X_j^t\|^2$. If the original trajectory is rigid, this quantity is independent of t and remains so after the transform. Each rigid motion group therefore carries one free positive scalar.

Corollary. The number of free parameters of this family equals the number G of rigid motion groups contained in the sequence rather than a single global scalar, and taking values per group is the main difference between this method and the conventional understanding. If predictions are treated as independent per frame, the degrees of freedom given by Proposition 1 are one scalar per frame and per group, $G \cdot T$ in total. Actual feed-forward trackers lie between the two, since temporal priors let them recover cross-frame consistency partially but not completely, which also explains why the correction is implemented as a per-frame scale plus smoothing rather than by estimating a single group constant directly. Classical monocular global scale ambiguity is the special case $G = 1$.

Role of the anchors. An anchor gives a true radial distance $\|\mathbf{Q}_i^t - C_t\|$. Comparing it with the predicted $\|\mathbf{P}_i^t - C_t\|$ yields the α_t that this point needs, so each anchor pins down one degree of freedom of that group at that frame. What an anchor provides is a scalar rather than a complete 3D position, which is why it can be effective at a very low budget. In terms of unobservability, the degrees of freedom of each group are just one scalar along the view direction. The three translational degrees of freedom do not belong to this family. They absorb the residual systematic bias beyond the scalar, and the two are different in kind.

What we do not claim. We do not claim that these models are unbiased in the tangential direction, nor do we claim the full gauge freedom. These models are trained on data with metric supervision, so scale is already partly pinned down by the prior, though only weakly. The theory therefore predicts partial rather than complete correction, and Section 5.7 reports a quantitative measurement of this upper bound.

A.3 Definitions of the Three Structural Metrics

Let $\mathbf{P}_i^t$ be the predicted position of query point i at frame t, $\mathbf{Q}_i^t$ the ground truth, $e_i^t = \mathbf{P}_i^t - \mathbf{Q}_i^t$ the error vector, and $r_i^t = (\mathbf{P}_i^t - C_t)/\|\mathbf{P}_i^t - C_t\|$ the unit vector of the view direction.

Radial energy is the ratio of the error energy projected onto the view direction to the total error energy, aggregated over all valid dynamic query points. A value of 1 means the error is entirely along the view direction, and 0.5 roughly corresponds to isotropic error, so 0.5 is the natural reference line for judging radial dominance.

Flow direction error is the median angle between the predicted and ground-truth velocity directions, with velocities computed only on consecutive co-visible frames. The expected value of a random baseline is about $90°$, so this quantity also characterizes the signal-to-noise ratio of the direction signal.

Between-group scale variance first assigns a group label $g(i)$ to each point from the grouping result of Section 4.2, then computes the optimal per-point per-frame radial scale $\alpha_i^{t\star}$, and finally performs a one-way ANOVA decomposition of $\log \alpha$ by group label, reporting the fraction of the between-group sum of squares in the total sum of squares. This quantity measures how much of the scale correction a point needs can be explained by knowing which motion group it belongs to, not the absolute size of the scale bias. The measurement is restricted to dynamic query points whose ground-truth displacement is at least τ_m, with $\tau_m = 0.05$ by default, because static background points carry no scale information.

Threshold sensitivity. Repeating the measurement at $\tau_m = 0.02$ and 0.10, the largest changes in the absolute values of the three metrics are 22.9 percentage points, $34.5°$ and 25.5 percentage points. The ranking among the eight models is not preserved for any of the three metrics, so we make no claim about rank stability. The qualitative thresholds on which the conclusions rest, radial energy above 0.5 and direction error below $90°$, hold under all combinations. Between-group scale variance across the eight models and two protocols ranges from 15.1% to 76.7%. Recomputing under all combinations of τ_{dir} from 0.80 to 0.95 and min_group_size from 3 to 10, endpoint error fluctuates by -10.1% to +14.7% relative to the default configuration.

A.4 Computation of the Evaluation Metrics

We give the exact computation of the three metrics introduced in Section 5.1, on the query-point set selected by the protocol of that section. In every case a sequence-level value is computed first, and

each table entry is the median of these values over the evaluation records, one record per seed and sequence.

End-point error. EPE is the median of the error magnitude $\|e_i^t\|_2$ over the points and frames where the ground truth is valid, in the world coordinate frame and in meters.

Average Percent of Points within Delta. At each threshold, APD is the fraction of ground-truth-visible query points whose prediction lies within the threshold distance of the ground truth, regardless of the predicted visibility. Following TAPVid-3D, the threshold is depth-adaptive and back-projects a pixel threshold δ_{2D} at the ground-truth depth of the point, $\delta_{3D}(i,t) = Z_i^t\,\delta_{2D}/f$, where Z_i^t is the depth of $\mathbf{Q}_i^t$ and f is the geometric mean of the two focal lengths. The pixel thresholds are $\delta_{2D} \in \{1, 2, 4, 8, 16\}$, and APD is the average of the five resulting fractions.

3D Average Jaccard. At the same five thresholds, a point is a true positive when the prediction lies within the threshold and both the prediction and the ground truth are visible, a false positive when the prediction is visible but the ground truth is not or the prediction lies outside the threshold, and a false negative when the ground truth is visible but the prediction is occluded or outside the threshold. AJ is the average over the five thresholds of $\mathrm{TP}/(\mathrm{TP} + \mathrm{FP} + \mathrm{FN})$, which jointly scores positional accuracy and visibility prediction.

Scale convention. Predictions enter the computation at their native metric scale. The global scale alignment of the TAPVid-3D reference implementation is not applied, because the evaluated models output world-frame trajectories in meters.

A.5 Hyperparameters

$\tau_{\text{static}} = 0.05$, $\tau_{\text{motion}} = 0.01$, $K_d = 3$, $\tau_{\text{dir}} = 0.90$, $\tau_{\text{dir}}^{\text{merge}} = 0.85$, 50 for the direction kNN, 10 for the spatial kNN, a minimum group size of 3, a fragment threshold of 5, a size lower bound of 5 for motion-consistency merging, $\tau_{\text{cov}} = 3$ (the lower bound on the frames common to two descriptors), a merge distance factor of 3.0, and static detection additionally requires at least 5 visible frames. The anchor budget is $\rho = 0.05$, the temporal smoothing kernel $\sigma = 2.0$, and the minimum number of anchors per frame is 1. The correction sets no lower bound on fitting size. The per-frame radial scale requires only that the frame have at least one anchor, and the Sim(3) and pure-translation fallbacks are enabled in turn only in degenerate cases. For the evaluation protocol, the three DQS filters are motion_min_norm = 0.05, conf_min_threshold = 0.3 and min_visible_frames = 5 in the code. They do not change with model or benchmark and are not overridden in any experiment script. Motion magnitude is the sum of the L2 norms of per-frame 2D displacement over the whole sequence, in normalized image coordinates. Visible frames are counted by predicted visibility. Confidence is taken from each model's own confidence field and is not recomputed by the evaluation script, and the form of that field differs across models. This motion threshold and the 3D coordinate spread threshold for static detection in Section 4.2 are two independent quantities, and so are the two lower bounds on visible frames.

For the interface, the correction head depends only on the 3D trajectories `coords`, the visibility `visible`, the per-frame camera center C_t taken from the ground-truth extrinsics provided by the datasets, and the ground-truth positions `gt_coords` and `gt_visible` used only for anchor sampling and fitting. The confidence `conf` enters only the query-point screening of the DQS protocol, and the normalized query coordinates `query_uv` only the reading of the instance labels for the Oracle row of Table A6. Neither enters the correction head. No depth map, optical flow, or intermediate feature of the base model is needed.

B Complete Numerical Results

This section gives the complete numerical results behind the main text.

B.1 Main Results and Structural Metrics

Table A1 gives the per-subset WorldTrack results under DQS for all eight models. Figure 5 in the main text plots only the relative-DQS-improvement column of Table A2, whose complete values under both protocols are given here. Table A3 reports the three structural metrics of Section 3 for every setting.

Table A1: Per-subset WorldTrack results under the DQS and Full protocols, in the world coordinate frame. ADT and PStudio provide no depth ground truth, so the three depth-dependent models are marked with a dash. Improvements are computed from the ratio of the absolute values in this table. Table 3 in the main text reports the averages over the subsets under DQS.

Model	Method	ADT			PO			PStudio			DS		
		EPE↓	APD↑	AJ↑	EPE↓	APD↑	AJ↑	EPE↓	APD↑	AJ↑	EPE↓	APD↑	AJ↑
DQS protocol													
D4RT	Base	0.1516	0.7462	0.0398	0.1770	0.7191	0.0187	0.1092	0.8141	0.0094	0.1000	0.8338	0.0039
	+Ours	0.1142	0.8025	0.0450	0.1001	0.8152	0.0216	0.0526	0.9482	0.0120	0.0681	0.8941	0.0043
	Impr. (%)	**+24.7**	**+7.5**	**+13.1**	**+43.4**	**+13.4**	**+15.5**	**+51.8**	**+16.5**	**+27.7**	**+31.9**	**+7.2**	**+10.3**
4RC	Base	0.2350	0.6687	0.0218	0.1587	0.6944	0.0161	0.1269	0.7955	0.0089	0.0917	0.7850	0.0043
	+Ours	0.2235	0.6630	0.0227	0.0988	0.7818	0.0196	0.0683	0.9066	0.0105	0.0709	0.8364	0.0047
	Impr. (%)	**+4.9**	-0.9	**+4.1**	**+37.7**	**+12.6**	**+21.7**	**+46.2**	**+14.0**	**+18.0**	**+22.7**	**+6.5**	**+9.3**
V-DPM	Base	0.0675	0.8966	0.0329	0.1372	0.7406	0.0396	0.1202	0.8191	0.1322	0.1100	0.7765	0.0299
	+Ours	0.0576	0.8997	0.0335	0.1023	0.7812	0.0402	0.0709	0.9107	0.1525	0.0723	0.8048	0.0321
	Impr. (%)	**+14.7**	**+0.3**	**+1.8**	**+25.4**	**+5.5**	**+1.5**	**+41.0**	**+11.2**	**+15.4**	**+34.3**	**+3.6**	**+7.4**
SM4RT	Base	0.0724	0.8970	0.0342	0.1663	0.6943	0.0154	0.1541	0.7557	0.0086	0.1097	0.7882	0.0035
	+Ours	0.0596	0.9013	0.0346	0.1181	0.7712	0.0154	0.0914	0.8802	0.0101	0.1027	0.7795	0.0038
	Impr. (%)	**+17.7**	**+0.5**	**+1.2**	**+29.0**	**+11.1**	**+0.0**	**+40.7**	**+16.5**	**+17.4**	**+6.4**	-1.1	**+8.6**
TAPIP3D	Base	—	—	—	0.1332	0.7660	0.0433	—	—	—	0.0498	0.9042	0.0373
	+Ours	—	—	—	0.0909	0.8102	0.0454	—	—	—	0.0336	0.8944	0.0450
	Impr. (%)	—	—	—	**+31.8**	**+5.8**	**+4.8**	—	—	—	**+32.5**	-1.1	**+20.6**
SpatialTrackerV2	Base	—	—	—	2.0970	0.0300	0.0299	—	—	—	1.5167	0.0665	0.0611
	+Ours	—	—	—	1.9514	0.1281	0.0375	—	—	—	0.8657	0.3016	0.0809
	Impr. (%)	—	—	—	**+6.9**	**+327.0**	**+25.4**	—	—	—	**+42.9**	**+353.5**	**+32.4**
DELTA	Base	—	—	—	0.8519	0.1949	0.0221	—	—	—	0.0607	0.8859	0.0242
	+Ours	—	—	—	0.3512	0.5062	0.0314	—	—	—	0.0435	0.8869	0.0272
	Impr. (%)	—	—	—	**+58.8**	**+159.7**	**+42.1**	—	—	—	**+28.3**	**+0.1**	**+12.4**
CoTracker3+DA	Base	0.8059	0.2437	0.0244	1.6076	0.0727	0.0121	0.2695	0.6270	0.0414	1.1407	0.1326	0.0102
	+Ours	0.1907	0.6643	0.0342	0.5136	0.4034	0.0208	0.1996	0.7060	0.0482	0.6085	0.4115	0.0145
	Impr. (%)	**+76.3**	**+172.6**	**+40.2**	**+68.1**	**+454.9**	**+71.9**	**+25.9**	**+12.6**	**+16.4**	**+46.7**	**+210.3**	**+42.2**
Full protocol													
D4RT	Base	0.1512	0.7705	0.4519	0.1865	0.7118	0.2965	0.1054	0.8375	0.4637	0.1230	0.7804	0.4140
	+Ours	0.1066	0.8174	0.4790	0.1526	0.7387	0.3176	0.0784	0.8947	0.5048	0.1332	0.7735	0.4103
	Impr. (%)	**+29.5**	**+6.1**	**+6.0**	**+18.2**	**+3.8**	**+7.1**	**+25.6**	**+6.8**	**+8.9**	-8.3	-0.9	-0.9
4RC	Base	0.1959	0.7060	0.3522	0.1630	0.7173	0.2723	0.0776	0.9037	0.4836	0.0807	0.8357	0.4340
	+Ours	0.2208	0.6749	0.3531	0.1197	0.7701	0.3087	0.0722	0.9122	0.5029	0.0758	0.8321	0.4498
	Impr. (%)	-12.7	-4.4	**+0.3**	**+26.6**	**+7.4**	**+13.4**	**+7.0**	**+0.9**	**+4.0**	**+6.1**	-0.4	**+3.6**
V-DPM	Base	0.0442	0.9496	0.5412	0.0765	0.8967	0.3649	0.0589	0.9468	0.5294	0.1094	0.7991	0.4162
	+Ours	0.0402	0.9409	0.5557	0.0789	0.8852	0.3709	0.0539	0.9520	0.5391	0.1087	0.7878	0.4083
	Impr. (%)	**+9.0**	-0.9	**+2.7**	-3.1	-1.3	**+1.6**	**+8.5**	**+0.5**	**+1.8**	**+0.6**	-1.4	-1.9
SM4RT	Base	0.0440	0.9438	0.5480	0.1010	0.8250	0.3290	0.0877	0.8794	0.4763	0.0739	0.8481	0.4370
	+Ours	0.0425	0.9398	0.5490	0.0900	0.8457	0.3423	0.0900	0.8802	0.4647	0.0699	0.8432	0.4534
	Impr. (%)	**+3.4**	-0.4	**+0.2**	**+10.9**	**+2.5**	**+4.0**	-2.6	**+0.1**	-2.4	**+5.4**	-0.6	**+3.8**
TAPIP3D	Base	—	—	—	0.0118	0.9896	0.5746	—	—	—	0.0076	0.9959	0.6851
	+Ours	—	—	—	0.0128	0.9849	0.5721	—	—	—	0.0079	0.9959	0.6891
	Impr. (%)	—	—	—	-8.5	-0.5	-0.4	—	—	—	-3.9	**+0.0**	**+0.6**
SpatialTrackerV2	Base	—	—	—	1.9093	0.0704	0.0875	—	—	—	1.6139	0.0662	0.1419
	+Ours	—	—	—	2.2072	0.0517	0.0660	—	—	—	1.5686	0.0984	0.1272
	Impr. (%)	—	—	—	-15.6	-26.6	-24.6	—	—	—	**+2.8**	**+48.6**	-10.4
DELTA	Base	—	—	—	0.5589	0.3341	0.1643	—	—	—	0.0524	0.8910	0.4596
	+Ours	—	—	—	0.3270	0.5126	0.2031	—	—	—	0.0436	0.8931	0.4638
	Impr. (%)	—	—	—	**+41.5**	**+53.4**	**+23.6**	—	—	—	**+16.8**	**+0.2**	**+0.9**
CoTracker3+DA	Base	0.5854	0.3235	0.2613	1.2544	0.1080	0.1134	0.2112	0.7006	0.3450	1.3998	0.0728	0.1444
	+Ours	0.4753	0.4184	0.2707	0.9002	0.2770	0.1280	0.2022	0.7113	0.3638	1.2969	0.1670	0.1528
	Impr. (%)	**+18.8**	**+29.3**	**+3.6**	**+28.2**	**+156.5**	**+12.9**	**+4.3**	**+1.5**	**+5.4**	**+7.4**	**+129.4**	**+5.8**

Table A2: Comparison with test-time optimization. Three metrics and the relative change in endpoint error for five models under three settings, base, +TTO and +Ours.

Model	Status	EPE↓	APD↑	AJ↑	ΔEPE (%)
D4RT	base	0.1812 / 0.1930	0.6877 / 0.6654	0.1038 / 0.3358	—
	+TTO	0.1720 / 0.1685	0.7090 / 0.6933	0.1057 / 0.3380	5.1 / 12.7
	+Ours	0.1269 / 0.1597	0.7648 / 0.7015	0.1214 / 0.3320	**29.9 / 17.3**
4RC	base	0.1421 / 0.2471	0.7334 / 0.6127	0.0711 / 0.2603	—
	+TTO	0.1419 / 0.2436	0.7358 / 0.6153	0.0863 / 0.2611	0.1 / 1.4
	+Ours	0.1035 / 0.1470	0.7832 / 0.7332	0.0892 / 0.3172	**27.2 / 40.5**
V-DPM	base	0.1451 / 0.0778	0.7322 / 0.8792	0.1164 / 0.4478	—
	+TTO	0.1230 / 0.0760	0.7845 / 0.8842	0.1417 / 0.4477	15.2 / 2.3
	+Ours	0.1086 / 0.0619	0.8091 / 0.9077	0.1164 / 0.4634	**25.2 / 20.4**
SM4RT	base	0.1351 / 0.1538	0.7709 / 0.7478	0.0791 / 0.3594	—
	+TTO	0.1351 / 0.1538	0.7710 / 0.7478	0.0890 / 0.3590	0.0 / 0.0
	+Ours	0.1000 / 0.0939	0.8010 / 0.8372	0.0851 / 0.3864	**26.0 / 38.9**
DELTA	base	0.7013 / 0.5493	0.2867 / 0.3607	0.0700 / 0.1525	—
	+TTO	0.7090 / 0.5554	0.2871 / 0.3697	0.0699 / 0.1525	-1.1 / -1.1
	+Ours	0.2974 / 0.2798	0.5274 / 0.4828	0.1057 / 0.2414	**57.6 / 49.1**

Table A3: Error structure of the eight trackers. The isotropic baseline of radial energy is 50% and the random baseline of flow direction error is about 90°.

Model	Radial energy (%)	Flow direction error ($^\circ$)	Between-group scale variance (%)
D4RT	77.7 / 74.8	24.0 / 19.8	62.9 / 74.4
4RC	67.3 / 89.4	38.4 / 31.4	73.7 / 72.2
V-DPM	73.2 / 73.0	35.5 / 37.9	47.6 / 69.7
SM4RT	62.3 / 80.2	64.1 / 62.9	47.9 / 68.1
TAPIP3D	67.9 / 37.9	7.4 / 87.7	59.7 / 15.1
SpatialTrackerV2	37.6 / 44.7	87.8 / 67.0	26.7 / 30.4
DELTA	61.6 / 60.3	27.3 / 18.4	60.7 / 67.1
CoTracker3+DA	32.3 / 39.4	86.7 / 86.7	76.7 / 71.8

B.2 Ablations and Anchor Sensitivity

GAUGE involves no training process, and given the input and the anchor set the whole pipeline is deterministic except for anchor sampling, that is, which points are drawn from the candidates satisfying the visibility conditions. The complete sweeps behind Sections 5.6, 5.4 and 5.5 are tabulated here. Table A4 gives the anchor-noise sweep, Table A5 the seven-level anchor-budget sweep and Table A6 the unsupervised-versus-oracle grouping comparison. The two kinds of dispersion that substitute for multi-seed reporting, anchor resampling and the sequence-level paired distribution, are given in Table A7 and Table A8.

Table A4: Robustness to anchor noise (D4RT). The correction problem is re-solved after applying isotropic Gaussian noise of standard deviation σ to the ground-truth anchor positions; the base row is the uncorrected baseline and $\sigma = 0$ is the noiseless anchor.

Noise σ (cm)	EPE↓	APD↑	AJ↑	ΔEPE vs. clean (%)
base	0.1812 / 0.1930	0.6877 / 0.6654	0.1038 / 0.3358	—
0 (clean)	0.1269 / 0.1597	0.7648 / 0.7015	0.1214 / 0.3320	0.0000 / 0.0000
1	0.1277 / 0.1598	0.7645 / 0.7040	0.1216 / 0.3313	-0.6524 / -0.1072
5	0.1386 / 0.1722	0.7502 / 0.6944	0.1210 / 0.3234	-9.2242 / -7.8694
10	0.1549 / 0.1994	0.7248 / 0.6634	0.1198 / 0.3122	-22.0386 / -24.9011
20	0.2241 / 0.2634	0.6565 / 0.6033	0.1107 / 0.2872	-76.5662 / -64.9686
50	0.4486 / 0.4940	0.4748 / 0.4262	0.0925 / 0.2378	-253.4478 / -209.3561

Table A5: Anchor budget. Endpoint error at each budget level; Appendix Figure A3 plots this curve for D4RT.

Model	Protocol	1%	2%	5%	10%	20%	50%	100%
D4RT	DQS	0.1399	0.1371	**0.1269**	0.1312	0.1216	0.1167	0.1289
	Full	0.1704	0.1702	**0.1597**	0.1495	0.1326	0.1300	0.1363
4RC	DQS	0.1048	0.1124	**0.1035**	0.1004	0.1177	0.1052	0.1095
	Full	0.1591	0.1510	**0.1470**	0.1347	0.1467	0.1253	0.1257
V-DPM	DQS	0.1217	0.0981	**0.1086**	0.1000	0.1014	0.1115	0.1150
	Full	0.0669	0.0641	**0.0619**	0.0626	0.0595	0.0611	0.0566
SM4RT	DQS	0.0930	0.0931	**0.1000**	0.0919	0.0929	0.0980	0.0977
	Full	0.1078	0.1110	**0.0939**	0.1017	0.0918	0.1072	0.0801
TAPIP3D	DQS	0.0962	0.0936	**0.0966**	0.1031	0.0923	0.0948	0.0911
	Full	0.4033	0.4022	**0.4023**	0.4018	0.3996	0.3793	0.3789
SpatialTrackerV2	DQS	2.1103	2.0447	**1.7428**	1.7104	1.8521	1.6782	1.7588
	Full	1.4069	1.3943	**1.3674**	1.3898	1.3203	1.2624	1.3297
DELTA	DQS	0.4067	0.3392	**0.2974**	0.2337	0.2467	0.2587	0.2492
	Full	0.3479	0.3151	**0.2798**	0.2270	0.1958	0.1823	0.1863
CoTracker3+DA	DQS	1.3798	1.1582	**0.5997**	0.4258	0.3365	0.3267	0.3229
	Full	1.0389	0.9125	**0.6016**	0.3318	0.2677	0.2562	0.2554

Table A6: Unsupervised grouping versus oracle instance grouping. On PointOdyssey, the three metrics of the eight trackers when the correction switches from oracle instances to the unsupervised motion grouping of this paper; the Unsupervised row is the result of applying GAUGE in Table 2 of the main text. Instance labels are used only in the Oracle instance row.

Model	Grouping	EPE↓	APD↑	AJ↑
D4RT	Oracle instance	0.1334 / 0.1159	0.7518 / 0.7717	0.1260 / 0.3913
	Unsupervised (Ours)	0.1269 / 0.1597	0.7648 / 0.7015	0.1214 / 0.3320
4RC	Oracle instance	0.1091 / 0.1216	0.7655 / 0.7435	0.0777 / 0.3213
	Unsupervised (Ours)	0.1035 / 0.1470	0.7832 / 0.7332	0.0892 / 0.3172
V-DPM	Oracle instance	0.1194 / 0.0549	0.7871 / 0.9071	0.1043 / 0.4674
	Unsupervised (Ours)	0.1086 / 0.0619	0.8091 / 0.9077	0.1164 / 0.4634
SM4RT	Oracle instance	0.1079 / 0.0812	0.7952 / 0.8167	0.0779 / 0.3733
	Unsupervised (Ours)	0.1000 / 0.0939	0.8010 / 0.8372	0.0851 / 0.3864
TAPIP3D	Oracle instance	0.1041 / 0.5789	0.8260 / 0.3971	0.1339 / 0.0419
	Unsupervised (Ours)	0.0966 / 0.4023	0.8365 / 0.4978	0.1576 / 0.0415
SpatialTrackerV2	Oracle instance	1.3194 / 0.7506	0.1499 / 0.3045	0.0580 / 0.1695
	Unsupervised (Ours)	1.7428 / 1.3674	0.1164 / 0.1397	0.0521 / 0.1165
DELTA	Oracle instance	0.3599 / 0.3627	0.4917 / 0.5190	0.0900 / 0.2180
	Unsupervised (Ours)	0.2974 / 0.2798	0.5274 / 0.4828	0.1057 / 0.2414
CoTracker3+DA	Oracle instance	1.0923 / 0.7038	0.1882 / 0.3171	0.0678 / 0.1822
	Unsupervised (Ours)	0.5997 / 0.6016	0.3646 / 0.3436	0.0756 / 0.1665

Table A7: Resampling dispersion. These values use a different estimator from the improvement in Table 2 of the main text and are not directly comparable.

Model	ΔEPE (%)	Resampling std (%)	Paired bootstrap 95% CI (%)
D4RT	**28.70**	1.0	[18.2, 39.3]
4RC	**19.7**	3.8	[16.8, 40.2]
V-DPM	**18.6**	2.4	[8.0, 28.8]
SM4RT	**20.4**	8.3	[5.4, 36.2]
TAPIP3D	**31.2**	1.2	[17.4, 45.8]
SpatialTrackerV2	**9.4**	3.2	[-3.1, 20.8]
DELTA	**52.4**	1.2	[41.7, 62.6]
CoTracker3+DA	**59.0**	1.0	[52.3, 66.2]

Table A8: Distribution of per-sequence endpoint error improvement. The quantiles use a different estimator from the improvement in Table 2 of the main text and are not directly comparable.

Model	P25 (%)	P50 (%)	P75 (%)	Seq. with gain (%)
D4RT	19.3	**31.0**	41.7	**92.3**
4RC	-16.9	**36.6**	50.8	**66.7**
V-DPM	-3.6	**18.4**	37.5	**53.8**
SM4RT	-3.9	**22.7**	38.4	**66.7**
DELTA	43.8	**60.2**	65.2	**100.0**
TAPIP3D	8.7	**28.6**	65.0	**91.7**
SpatialTrackerV2	-5.0	**16.3**	28.3	**61.5**
CoTracker3+DA	47.1	**60.5**	66.6	**100.0**

B.3 Efficiency

Table A9 details the post-processing cost summarized in Section 5.7.

Table A9: Post-processing cost. Timings in seconds per sequence, from a single core whose frequency scales dynamically between 1.50 and 3.71 GHz; each cell gives DQS / Full. No GPU, autograd or optimizer is used.

Median wall (s)	P95 wall (s)	Median CPU (s)	Peak RSS (MiB)	Extra RSS (MiB)
0.490 / 2.601	3.181 / 35.183	0.490 / 2.600	594.109 / 769.656	0.742 / 29.688

C Additional Ablations and Controls

C.1 Design Choices

- The rotational degrees of freedom of the full Sim(3) are unnecessary to introduce. Both rotation and tangential scale change the projection, and nothing in the data supports such a change. The Sim(3) column of Table 2 in Section 5.2 gives the measured comparison under the same grouping and the same anchors.
- Per-frame scale and per-group translation remove different kinds of bias. The former corresponds to time-varying drift and the latter to a time-independent residual. Both are estimated from the same anchors and their degrees of freedom do not overlap. The step-by-step results in Section 4.4 support this division of labor.
- Estimates take the median rather than the mean, because the distribution of depth ratios has extreme values at points close to the camera center and where tracking is briefly lost. The median is not dominated by such points, at the cost of slightly lower statistical efficiency. This choice is what lets the 10 cm noise setting of Section 5.6 still beat the baseline.
- Three controls on anchor allocation. Motion weighting differs from allocation proportional to group size by 0.0 and 0.7 percentage points, and from completely uniform and ungrouped allocation by 24.3 and 30.2 percentage points. Preferring co-frame anchors differs from cross-frame sampling by 1.6 and 0.2 percentage points. The first two are given in Section 4.3. The third shows that a common frame is not a critical design choice.
- We also tried and abandoned several parameterizations with higher capacity, including depth-dependent radial scale, a three-degree-of-freedom anisotropic scale, correction that targets the trajectory accuracy metrics directly, and iterative correction. They either made the error worse outright, or fit the validation set and then degraded on the test set.

C.2 Ablation Consistency on a Second Model (4RC)

Table A10: Ablations repeated on 4RC. Configurations are taken from Table 4 in the main text. The two forms that scale about an in-group anchor were not repeated on 4RC, and the single-form-on-all-groups and pure-translation rows report endpoint error only. In (c), AE is the average per-anchor efficiency accumulated to that budget, defined as in Table A5.

(a) Grouping strategy

Grouping Method	EPE↓	APD↑	AJ↑
No grouping (global)	0.1370 / 0.2427	0.7511 / 0.6217	0.0716 / 0.2651
Oracle instance	0.1091 / 0.1216	0.7655 / 0.7435	0.0777 / 0.3213
Full (Ours)	**0.1035 / 0.1470**	**0.7832 / 0.7332**	**0.0892 / 0.3172**

(b) Degrees of freedom of the correction

Configuration	EPE ↓
Scale only (1-DOF)	0.1403 / 0.1953
Translation only (3-DOF)	0.1354 / 0.1815
Scale + translation	0.1377 / 0.1823
Single form on all groups	0.1432 / 0.1667
Ours (per-frame scale + translation)	**0.1035 / 0.1470**

(c) Anchor budget

Budget (%)	EPE↓	APD↑	AJ↑	AE↑
1%	0.1048 / 0.1591	0.7881 / 0.7112	0.0858 / 0.3142	26.2 / 35.6
5% (default)	**0.1035 / 0.1470**	**0.7832 / 0.7332**	**0.0892 / 0.3172**	**5.4 / 8.1**
20%	0.1177 / 0.1467	0.7652 / 0.7334	0.0872 / 0.3118	0.9 / 2.0
100%	0.1095 / 0.1257	0.7785 / 0.7572	0.0880 / 0.3132	0.2 / 0.5

On the two dimensions of grouping and correction degrees of freedom, the qualitative conclusions on 4RC agree with those on D4RT, namely that grouping is clearly better than no grouping. Two details differ from D4RT. First, removing grouping gives 0.1370 on 4RC, better than the baseline 0.1421, rather than being slightly worse than the baseline as on D4RT, which is consistent with the report in Section 3.3. Second, the combination of group-level scale and translation gives 0.1377, slightly worse than the 0.1354 of translation alone, showing that the contribution of group-level scale on this model is close to zero and the gain comes mainly from the per-frame scale, consistent with the conclusion of the degree-of-freedom breakdown in Section 4.4.

C.3 Depth Input versus Post-Hoc Correction

Table A11: Depth input versus post-hoc correction. Three metrics of TAPIP3D before and after correction under two input conditions, ground-truth depth and predicted depth; DQS only.

Configuration	EPE↓	APD↑	AJ↑
TAPIP3D (GT depth input)	0.1548	0.7600	0.1560
TAPIP3D (GT depth input) + Ours	**0.0966**	**0.8365**	**0.1576**
TAPIP3D (predicted depth input)	1.6705	0.0565	0.0319
TAPIP3D (predicted depth) + Ours	**0.6949**	**0.3639**	**0.0651**

Table A11 reports the comparison. The depth input determines the accuracy ceiling of this line. Endpoint error with ground-truth depth is 0.1548. Switching to the predicted depth available in deployment changes it to 1.6705, a degradation of more than tenfold. Correction is effective for both inputs but does not change this ceiling, because the error introduced by depth noise is something that four degrees of freedom cannot characterize. GAUGE needs no change to the depth input of the model and works equally well on imprecise input, but it cannot replace a reliable depth source.

C.4 Ground-Truth Motion Control

To check whether DQS inflates the gain by using predicted motion magnitude to screen points, we redo the screening with ground-truth motion magnitude in place of the predicted value.

Table A12: Screening control with ground-truth motion magnitude. Gain in endpoint error after redoing the dynamic query screening with ground-truth motion magnitude in place of the predicted value; Δ is the predicted-motion gain minus the ground-truth one and Jaccard is the overlap of the two selected point sets. These gains are computed on motion-screened point sets and are not comparable with the DQS improvement in Table 2 of the main text.

Model	Predicted-motion gain %	GT-motion gain %	Δ (pp)	Jaccard
D4RT	**30.5**	**30.5**	-0.01	0.7
4RC	**16.5**	**37.6**	-21.11	0.59
V-DPM	**23.6**	**29.4**	-5.88	0.71
SM4RT	**83.4**	**24.6**	58.82	0.16
SpatialTrackerV2	**30.4**	**36.1**	-5.69	0.35
TAPIP3D	**34.5**	**38.3**	-3.80	0.61
DELTA	**70.2**	**67.5**	2.75	0.52
CoTracker3+DA	**51.2**	**48.9**	2.29	0.62

Table A12 gives the per-model comparison. The median of the differences across the eight models is -1.91 percentage points, so screening by predicted motion magnitude does not systematically inflate the gain. The largest difference is on SM4RT (58.82 percentage points), whose two selected point sets overlap by only 0.16, so the difference comes from which points are selected rather than from the method itself. The values of the three thresholds are given in Section 5.1. They were fixed before the experiments and were not tuned for this method.

C.5 Null Test Details

Table A13: Null hypothesis test. The tested statistic is the between-group scale variance, that is, the between-group difference in per-point optimal radial scale that unsupervised grouping can explain; the definition is given in Appendix A.3. The null model keeps the size of each group unchanged and randomly permutes the point-to-group labels, one hundred times per sequence.

Model	Protocol	Observed	Null mean	Null p95	Excess over p95
D4RT	DQS	**62.94**	12.39	30.97	28.24
	Full	**74.36**	11.03	12.38	51.09
4RC	DQS	**73.65**	12.81	24.09	48.54
	Full	**72.17**	4.52	12.8	60.92
V-DPM	DQS	**47.64**	16.83	32.52	24.05
	Full	**69.67**	4.42	7.17	47.95
SM4RT	DQS	**47.9**	6.92	11.65	36.69
	Full	**68.08**	4.27	17.06	59.41
TAPIP3D	DQS	**59.66**	13.76	22.95	18.25
	Full	**15.08**	3.84	5.81	7.62
SpatialTrackerV2	DQS	**26.74**	18.02	26.51	0
	Full	**30.4**	5.66	9.2	25.42
DELTA	DQS	**60.66**	23.01	55.33	2.7
	Full	**67.12**	9.4	12.17	35.9
CoTracker3+DA	DQS	**76.72**	65.62	74.01	-1.04
	Full	**71.79**	35	43.24	25.14

The null distribution is constructed by randomly permuting the point-to-group labels while keeping the size of each group unchanged, and is recomputed one hundred times per setting. The observed value is compared with the 95th percentile of this distribution.

The columns of Table A13 are as follows. Observed is the measured statistic. Null mean and Null p95 are the mean and the 95th percentile of the null distribution. Excess over p95 is how far the observed value exceeds the 95th percentile of the null, and because it is computed per sequence before aggregation it does not equal the difference between the Observed and Null p95 columns. The first three columns are in percent, and Excess over p95 is in percentage points.

C.6 All Evaluated Settings

Benchmark construction. The two real-world WorldTrack test sets, Aria Digital Twin and Panoptic Studio, have annotations converted from TAPVid-3D camera-frame trajectories (Koppula et al., 2024) with camera extrinsics, which places them in the world coordinate frame. PointOdyssey plays

two roles. As a benchmark it is evaluated on its original test split under the standard protocol of that dataset, for direct comparison with existing work. As a WorldTrack subset it is reorganized under the unified world-frame protocol, to examine accuracy when geometry, camera motion and point motion are all correct. The two uses differ and do not double count. Throughout the paper, improvements are computed after global alignment of the baseline output, and all values come from a single frozen set of cached predictions.

Pearson correlation coefficients between the three structural metrics and endpoint error improvement are taken over all settings where the structural metrics are available, including settings with negative gain, 58 in total. The correlation of between-group scale variance is $r = 0.61$ ($n = 58$, permutation test $p < 0.0001$), radial energy $r = 0.078$ ($p = 0.56$) and flow direction error $r = 0.058$ ($p = 0.67$). The latter two are not significant. The remaining settings do not meet the motion-magnitude condition required by these metrics and are therefore absent from Figure A1, while their endpoint error improvements are still given in Table 2 and Table A1. Points below the zero line in that figure are settings with negative gain.

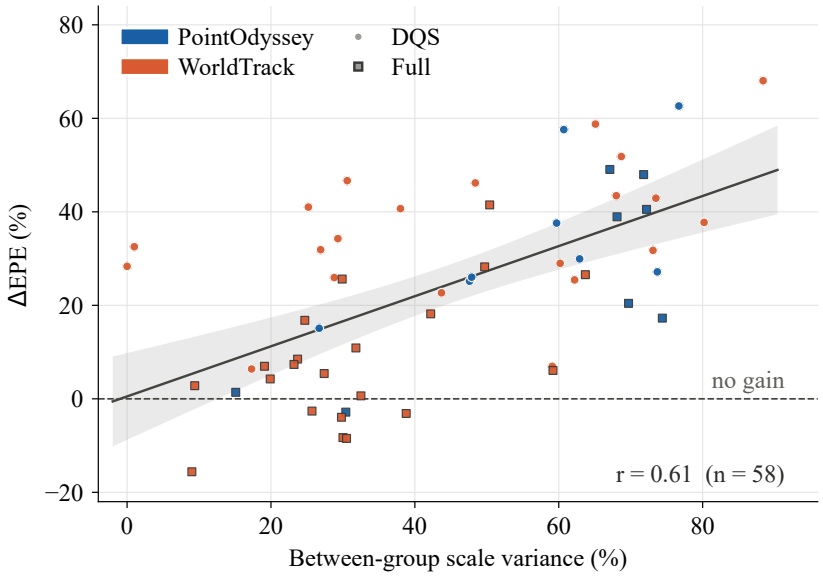


Figure A1: Failure prediction. Relative endpoint error improvement against between-group scale variance, one point per model, split and protocol. The gray line is the least-squares fit over the 58 points and the band its 95% confidence interval.

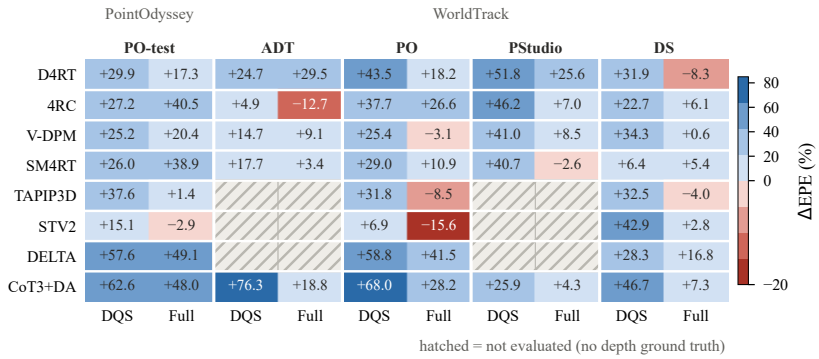


Figure A2: Heatmap of relative endpoint error improvement.

C.7 Complete Error Budget Table

The three medians reported in Section 5.7 have ranges of 7.1% to 82.6%, -2.8% to 62.6% and -11.5% to 92.5%, respectively.

Table A14: Error budget. The upper bound from fitting radial scale and translation per group and per frame on all ground-truth points, against what GAUGE achieves with 5% anchors.

Model	Protocol	Baseline	GAUGE 5%	Ceiling
D4RT	DQS	0.1812	**0.1269**	0.1167
	Full	0.1930	**0.1597**	0.1300
4RC	DQS	0.1421	**0.1035**	0.1004
	Full	0.2471	**0.1470**	0.1203
V-DPM	DQS	0.1451	**0.1086**	0.0994
	Full	0.0778	**0.0619**	0.0566
SM4RT	DQS	0.1351	**0.1000**	0.0919
	Full	0.1538	**0.0939**	0.0683
TAPIP3D	DQS	0.1548	**0.0966**	0.0911
	Full	0.4080	**0.4023**	0.3789
SpatialTrackerV2	DQS	2.0526	**1.7428**	1.4699
	Full	1.3295	**1.3674**	1.0006
DELTA	DQS	0.7013	**0.2974**	0.2337
	Full	0.5493	**0.2798**	0.1793
CoTracker3+DA	DQS	1.6050	**0.5997**	0.2847
	Full	1.1562	**0.6016**	0.2009

C.8 Additional Figures

Figure A3 gives the anchor-budget curve of D4RT referenced in Section 5.4.

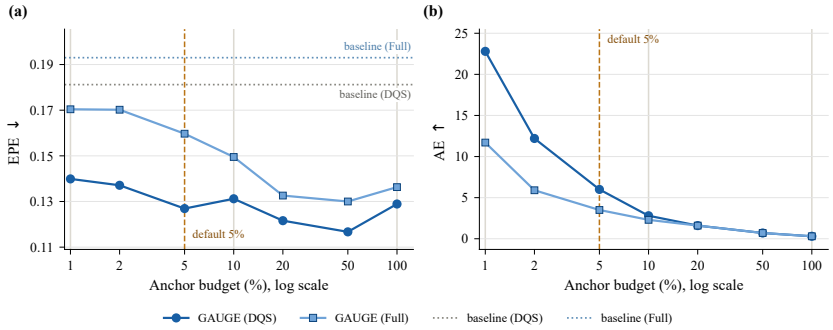


Figure A3: Anchor budget of D4RT. (a) Endpoint error and (b) average accumulated per-anchor efficiency.