NO EXTERNAL PRIORS NEEDED: UNLOCKING THE NATIVE DYNAMIC ROBUSTNESS OF DROID-SLAM

Anonymous authors
Paper under double-blind review

ABSTRACT

The core challenge of SLAM in dynamic environments is that independent motion violates the static-scene assumption, and even a small fraction of dynamic observations can corrupt tracking and reconstruction. Mainstream solutions resort to external modules such as semantic segmentation or motion masks to remove dynamic content, at the cost of redundant computation and domain-shift sensitivity. This paper gives the opposite answer: the native signals of DROID-SLAM themselves are capable of resisting dynamic interference, without any external priors. Three designs support this claim—a local linearization analysis reveals that independent motion remains in the residual only through the component unexplainable by the static model, yielding a purely geometric discriminator in which reprojection and epipolar residuals complement each other; the Schur-complement curvature of the BA solver is directly converted into per-pixel geometric reliability, replacing learned uncertainty prediction; and temporal re-examination over multiple keyframe intervals drives static Gaussian mapping, enabling clean reconstruction of dynamic scenes. Extensive experiments on a wide range of datasets demonstrate that our framework enjoys clear advantages in inference speed and computational cost, while matching or even surpassing state-of-the-art methods based on external priors in accuracy.

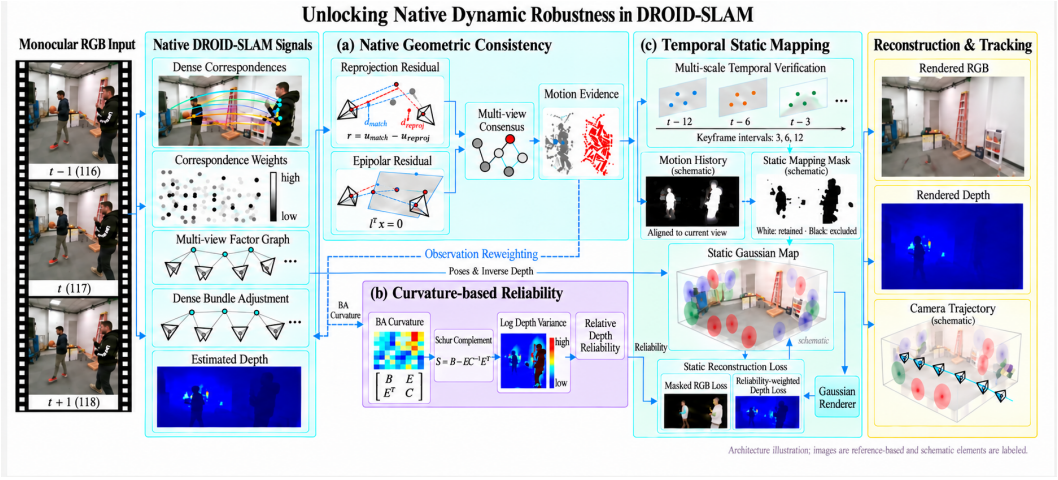


Figure 1: **System overview.** Native DROID-SLAM signals—correspondences, weights, the factor graph, and the BA state—are turned into dynamic robustness without external modules: reprojection and epipolar residuals, fused by multi-view consensus, softly down-weight dynamic observations (a); BA curvature yields per-pixel geometric reliability (b); temporally re-examined evidence forms a static mask that supervises the Gaussian map (c), producing clean color, depth, and camera trajectory.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Simultaneous localization and mapping in dynamic environments is an unavoidable step on the path of 3D vision toward the real world. As cameras move through crowds and drive along streets, independently moving objects are always present in the field of view—these observations violate the static-scene assumption, yet they inevitably enter the estimation loop together with the static background. How to identify and neutralize dynamic interference during estimation fundamentally determines trajectory accuracy and map quality.

Faced with this problem, the prevailing paradigm seeks help from outside the system: semantic segmentation, motion masks, or uncertainty prediction networks first remove dynamic content, and the "purified" observations are then handed over to the geometric backend. This paradigm is costly—external modules bring redundant computation and training dependencies, their failure modes are independent of the geometric estimation itself, and misjudgments under domain shift translate directly into a permanent loss of constraints. More regrettably, this practice silences the estimator's own voice: correspondence-based dense systems such as DROID-SLAM naturally produce residuals, matching weights, and solver curvature in the course of optimization—signals that already carry rich information about dynamic interference, yet are superseded by external hard masks and left underexploited. The root cause is a lack of theoretical understanding of the coupling between optical flow (correspondence) estimation and geometric optimization: once independent motion enters the coupled estimation system, when does it surface in the residual, and when is it silently absorbed by state updates? Existing work has not answered this question.

We argue that the key to resolving this lies in a staged, end-to-end scheme that fuses native signals (Fig. 1). We first perform a local linearization analysis of the coupled pose–depth–correspondence estimation, revealing that the component of independent motion lying in the column space of the geometric Jacobian is absorbed by state updates, and only the component that the static model cannot explain remains in the optimized residual. This understanding directly gives rise to a discriminative criterion in which the reprojection residual and the epipolar residual complement each other: the former examines full geometric consistency, while the latter targets the component that depth cannot explain away; together with an adaptive robust threshold and multi-edge evidence aggregation, dynamic discrimination is achieved without any external supervision. Following the same principle of seeking information from within the estimator, we further turn the Schur-complement curvature of the BA normal equations into a per-pixel geometric reliability estimate, which cooperates with dynamic consistency in a division of labor—the latter softly down-weights inconsistent observations, while the former modulates the contribution of retained observations to map optimization; the two native signals each fulfill their own role, replacing learned uncertainty predictors. Ultimately, the entire framework introduces no external module at all: motion evidence is temporally aggregated over multiple keyframe intervals and continuously re-examined through confirmation and retrospective updates; the re-examined static evidence drives Gaussian mapping, in which color and depth supervision each operate on their own valid domains, so that background that is geometrically under-constrained yet photometrically valid is fully preserved.

Our contributions are summarized as follows:

- **Theoretical characterization of dynamic identifiability in coupled estimation systems.** Through a local linearization analysis, we reveal that the component of independent motion lying in the column space of the geometric Jacobian is absorbed by state updates, and only the component that the static model cannot explain remains in the optimized residual; based on this, we construct a discriminative criterion in which the reprojection residual and the epipolar residual complement each other, achieving dynamic discrimination purely from native geometric signals. This is contrary to existing methods.
- **Replacing learned uncertainty prediction with the second-order structure of the optimizer.** We turn the Schur-complement curvature of the BA normal equations into a per-pixel geometric reliability estimate, decoupled from dynamic consistency in a division of labor—the former modulates the supervision strength of retained observations, while the latter softly down-weights inconsistent ones—without any external uncertainty predictor.
- **A temporally re-examined static Gaussian mapping framework without external modules.** We propose a temporal re-examination mechanism over multiple keyframe intervals, assign color and depth supervision their respective valid domains, and rebuild the

Gaussian map from the re-examined static evidence, yielding clean reconstructions of dynamic scenes. Extensive experiments show that our framework matches or even surpasses prior-based methods in accuracy while achieving a ten-fold speedup. This is in stark contrast to existing methods that rely on external priors.

On five datasets including TUM, KITTI, and BONN, this fully endogenous framework achieves experimental results identical to, or even surpassing, state-of-the-art methods based on external priors. This confirms the central claim of this paper: the native signals of DROID-SLAM themselves possess the capability to resist dynamic interference—what is needed is theoretical insight into their coupled structure and a principled way to decouple and exploit them.

2 METHOD

2.1 DYNAMIC CONSISTENCY FROM CORRESPONDENCE-GEOMETRY AND EPIPOLAR RESIDUALS

Let the input image sequence be $\mathcal{I} = \{I_t\}_{t=1}^{N}$. The system maintains keyframe poses and dense inverse depths over a factor graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where the node set $\mathcal{V}$ represents keyframes and the edge set $\mathcal{E}$ represents image pairs with overlapping fields of view. For an edge (i, j), let $\hat{\mathbf{u}}_{ij}(\mathbf{x})$ denote the predicted correspondence of a source pixel $\mathbf{x}$. Under the static-scene assumption, this correspondence should satisfy the rigid projection relation:

$$\mathbf{u}_{ij}^{\mathrm{rig}}(\mathbf{x}) = \pi\big(\mathbf{T}_j \mathbf{T}_i^{-1} \pi^{-1}(\mathbf{x}, d_i(\mathbf{x}))\big), \tag{1}$$

where $\mathbf{T}_i \in SE(3)$ is the transformation from the world frame to the camera frame of the i-th frame, d_i is the inverse depth, and π and π^{-1} denote projection and back-projection, respectively. The correspondence-geometry residual is then defined as:

$$\mathbf{r}_{ij}(\mathbf{x}) = \hat{\mathbf{u}}_{ij}(\mathbf{x}) - \mathbf{u}_{ij}^{\mathrm{rig}}(\mathbf{x}). \tag{2}$$

In dynamic environments, independent motion, matching errors, and geometric estimation errors can all produce non-zero residuals. Additional visual representations are commonly used to provide discriminative evidence among these error sources; from the perspective of static reconstruction, a discriminative criterion can also be constructed based on whether an observation can be explained by the shared geometric model. It should be noted that the correspondence estimation of DROID is influenced by the current geometric state; therefore, Eq. 2 is a consistency residual within a coupled estimation system, rather than a difference between two independent measurements.

Under local linearization, the stacked residual approximately satisfies:

$$\mathbf{r} \approx \mathbf{J}_\xi \Delta\boldsymbol{\xi} + \mathbf{J}_d \Delta d + \mathbf{m} + \boldsymbol{\epsilon}, \tag{3}$$

where $\mathbf{J}_\xi$ and $\mathbf{J}_d$ are the Jacobians with respect to poses and inverse depths, respectively, $\mathbf{m}$ denotes the correspondence displacement caused by independent motion, and $\boldsymbol{\epsilon}$ denotes observation noise. Let

$$\mathbf{A} = \mathbf{W}^{1/2} \begin{bmatrix} \mathbf{J}_\xi & \mathbf{J}_d \end{bmatrix},$$

then, with correspondences and weights fixed and under an undamped linear least-squares approximation, the optimized residual can be written as:

$$\tilde{\mathbf{r}}^{\,*} \approx \left(\mathbf{I} - \mathbf{A}\mathbf{A}^{\dagger}\right) \mathbf{W}^{1/2} \left(\mathbf{m} + \boldsymbol{\epsilon}\right), \tag{4}$$

where $\mathbf{A}^{\dagger}$ is the Moore–Penrose pseudoinverse. Eq. 4 reveals the local identifiability of dynamic observations: the component of independent motion lying in the column space of the geometric Jacobian may be absorbed into state updates, and only the component that cannot be explained by the static model remains in the residual. Consequently, relying on the reprojection residual alone may either misclassify geometric errors as motion, or miss motion that has been explained away by erroneous geometry.

To enhance the complementarity of the constraints, we further introduce epipolar consistency. Given the fundamental matrix $\mathbf{F}_{ij}$ associated with the relative pose, the distance from the target correspondence to its epipolar line is defined as:

$$r_{ij}^{\mathrm{epi}}(\mathbf{x}) = \frac{\left|\bar{\mathbf{u}}_{ij}^{\mathsf{T}} \mathbf{F}_{ij} \bar{\mathbf{x}}\right|}{\sqrt{(\mathbf{F}_{ij}\bar{\mathbf{x}})_1^2 + (\mathbf{F}_{ij}\bar{\mathbf{x}})_2^2}}, \tag{5}$$

where $\bar{\mathbf{x}}$ and $\bar{\mathbf{u}}_{ij}$ are homogeneous pixel coordinates. When the relative pose is fixed and the translation is non-degenerate, depth variations can only shift the correspondence along the epipolar line and cannot eliminate deviations perpendicular to it. Therefore, the reprojection residual and the epipolar residual examine full geometric consistency and its depth-unexplainable component, respectively. For degenerate cases such as near-zero translation, the epipolar criterion is disabled to avoid mistaking a lack of constraint for satisfaction of the static assumption.

Considering that error scales differ across image pairs, the residual median and a robust scale are computed over the set of valid observations:

$$m = \mathrm{median}(r), \qquad s = 1.4826\,\mathrm{median}(|r - m|),$$

and an adaptive threshold is defined as:

$$\tau = \max\left(0.5,\ 2m,\ m + 2.5s\right). \tag{6}$$

The above criterion is applied to the reprojection residual and the epipolar residual separately, and any valid test exceeding the threshold casts an outlier vote. Here, 0.5 pixels is defined on the correspondence grid at $1/8$ resolution. This robust estimation determines the error scale from the dominant consistent observations; its validity relies on sufficient static observation support and is not guaranteed to hold when dynamic regions dominate.

To reduce the impact of misjudgment from a single factor, evidence from multiple edges of the same source pixel is aggregated:

$$q_i^{(k)}(\mathbf{x}) = (1-\eta) q_i^{(k-1)}(\mathbf{x}) + \eta \frac{\sum_j v_{ij}(\mathbf{x})\, b_{ij}(\mathbf{x})}{\sum_j v_{ij}(\mathbf{x})}, \tag{7}$$

where v_{ij} denotes observation validity and b_{ij} denotes the outlier vote. The score is updated only when the number of valid edges is at least 2; the update coefficient is $\eta = 0.5$, and the dynamic classification threshold is $q_i \geq 0.35$. Invalid correspondences do not enter the vote denominator, preventing occlusions and out-of-bound projections from being incorrectly counted as static support. This score characterizes the degree of inconsistency with the multi-view static model and is not directly interpreted as a calibrated motion probability.

2.2 Geometric Reliability Estimation Based on the Second-Order Structure of the Optimization

Dynamic consistency and geometric reliability describe the model applicability of observations and the degree of state constraint, respectively. Static regions may have unreliable depth due to weak texture, low parallax, or insufficient observations; dynamic regions may also yield numerically stable estimates under an erroneous model. Therefore, compressing these two types of information into a single uncertainty measure tends to conflate dynamic rejection with background preservation. We model observation consistency and local geometric reliability separately, assigning them distinct roles in the optimization.

For observations with dynamic outlier evidence, the joint pose-depth optimization applies soft down-weighting:

$$\min_{\mathbf{T},d} \frac{1}{2} \sum_{(i,j)\in\mathcal{E}} \sum_{\mathbf{x}} \mathbf{r}_{ij}^{\mathsf{T}}(\mathbf{x}) \widetilde{\mathbf{W}}_{ij}(\mathbf{x}) \mathbf{r}_{ij}(\mathbf{x}), \tag{8}$$

$$\widetilde{\mathbf{W}}_{ij}(\mathbf{x}) = \gamma_{ij}(\mathbf{x}) \mathbf{W}_{ij}(\mathbf{x}), \qquad \gamma_{ij} = \begin{cases} 0.1, & \text{observations with dynamic outlier evidence,} \\ 1, & \text{other valid observations.} \end{cases} \tag{9}$$

where $\mathbf{W}_{ij}$ is the weight output by the matching network. Retaining a small non-zero contribution mitigates the constraint loss caused by erroneous rejection when the initial poses are inaccurate. The consistency evidence is recomputed after state updates, forming an alternating correction process between observation weights and the geometric state.

Per-pixel reliability depends not only on local image appearance, but also on camera baseline, the number of observations, and pose-depth coupling. Additional uncertainty predictors typically learn the association between these factors and observation errors, whereas the second-order structure of the joint optimization already contains the actual constraint information of the current observations

on the state. Its local curvature can therefore be exploited to construct an approximation of geometric reliability.

Partitioning the linearized normal equations into pose and inverse-depth variables:

$$\begin{bmatrix} \mathbf{B} & \mathbf{E} \\ \mathbf{E}^\mathsf{T} & \mathbf{C} \end{bmatrix} \begin{bmatrix} \Delta\boldsymbol{\xi} \\ \Delta d \end{bmatrix} = \begin{bmatrix} \mathbf{b}_\xi \\ \mathbf{b}_d \end{bmatrix}, \tag{10}$$

where $\mathbf{B}$ and $\mathbf{C}$ correspond to pose and inverse-depth variables, respectively, and $\mathbf{E}$ describes their coupling. Let $\mathbf{Q} = \mathbf{C}^{-1}$; eliminating the inverse-depth variables yields the Schur complement:

$$\mathbf{S} = \mathbf{B} - \mathbf{E}\mathbf{Q}\mathbf{E}^\mathsf{T}. \tag{11}$$

Provided that gauge freedoms are properly handled and an appropriate local noise approximation is adopted, the inverse-depth covariance is approximated, via the block matrix inversion relation, as:

$$\boldsymbol{\Sigma}_d \approx \mathbf{Q} + \mathbf{Q}\mathbf{E}^\mathsf{T}\mathbf{S}^{-1}\mathbf{E}\mathbf{Q}. \tag{12}$$

The first term of Eq. 12 describes the inverse-depth constraint conditioned on fixed cameras, while the second term describes the propagation of pose uncertainty to depth through the coupling. Ignoring the second term amounts to treating the camera states as perfectly known, which may overestimate depth precision under weak parallax or insufficient pose constraints.

The practical computation adopts a damped form consistent with the corresponding BA solve:

$$\mathbf{S}_{\lambda,\epsilon} = \mathbf{S} + \lambda \operatorname{diag}(\mathbf{S}) + \epsilon\mathbf{I}, \tag{13}$$

where λ and ϵ use the actual damping parameters of the current BA call. Let $\mathbf{S}_{\lambda,\epsilon} = \mathbf{L}\mathbf{L}^\mathsf{T}$; for the p-th depth variable, define $\mathbf{a}_p = \mathbf{E}\mathbf{Q}\mathbf{e}_p$, then:

$$[\boldsymbol{\Sigma}_d]_{pp} \approx Q_{pp} + \left\|\mathbf{L}^{-1}\mathbf{a}_p\right\|_2^2. \tag{14}$$

This form obtains the required diagonal entries through triangular solves, without explicitly constructing the full inverse matrix. In implementation, the computation is performed in blocks of 512 depth variables; when the factorization fails or non-finite values occur, the corresponding estimates are marked as unreliable. Since the matching weights, robust down-weighting, and damping do not necessarily satisfy strict statistical calibration conditions, the result should be interpreted as a local geometric reliability approximation conditioned on the current optimization state, rather than a true error probability.

For depth $z = 1/d$, first-order error propagation gives:

$$\sigma_z^2 \approx \frac{\sigma_d^2}{d^4}. \tag{15}$$

The absolute depth variance is amplified with distance, and directly applying a fixed threshold to exclude high-variance pixels tends to introduce a systematic bias against distant background. To this end, we adopt the relative depth uncertainty:

$$u_i(\mathbf{x}) = \frac{\sigma_{z,i}^2(\mathbf{x})}{z_i^2(\mathbf{x}) + \varepsilon},$$

and construct a bounded weight using the median $\tilde{u}_i$ over valid regions:

$$c_i(\mathbf{x}) = \operatorname{clip}\left(\frac{1}{1 + u_i(\mathbf{x})/\tilde{u}_i},\, 0.1,\, 1\right). \tag{16}$$

This weight modulates the supervision strength of valid depth observations, without directly treating insufficient geometric constraint as motion. Through this decomposition, dynamic consistency suppresses observations that are incompatible with the static model, while geometric reliability modulates the contribution of retained observations to map optimization.

2.3 Static Gaussian Map Update Under Temporal Consistency Constraints

Satisfying static geometric relations within a local temporal window does not imply that an observation can serve as a reliable source for the static map in the long term. For slow or intermittent motion, the displacement between adjacent frames may fall below the detection threshold, while the geometric deviation becomes identifiable only over longer time intervals. Map fusion therefore needs to re-examine earlier judgments using subsequent observations, avoiding equating a single absence of motion evidence with persistent staticness.

Cross-time associations are established along valid correspondences, and motion evidence is aggregated over temporal intervals of $\{3, 6, 12\}$ keyframes. Shorter intervals provide relatively stable correspondences, while longer intervals help accumulate displacements caused by slow motion. Meanwhile, larger temporal intervals also increase the risk of occlusion and mismatches; therefore, historical evidence is propagated only when correspondence validity and visibility conditions hold. The backward correspondences required for bidirectional verification must be computed independently and cannot be replaced by coordinate inversion.

External semantic or feature guidance can supplement insufficient local evidence, while subsequent observations can likewise corroborate or refute earlier static hypotheses. Based on this temporal complementarity, an initial static state requires at least 3 consecutive confirmations; once a motion event is supported by adjacent observations, the preceding 2 frames are retrospectively examined and the fusion states of the related observations are updated. For regions with confirmed motion, a brief absence of motion evidence does not immediately restore their reliability; propagation is interrupted when correspondences fail, preventing stale states from being incorrectly applied to newly revealed background. The above temporal intervals and confirmation counts serve as the initial settings of the temporal module, and their applicability needs to be validated on the corresponding correspondence-propagation implementation.

In 3D Gaussian mapping, let the map parameters be $\mathcal{M}$, and the rendered color and depth be $\hat{I}_i(\mathcal{M})$ and $\hat{z}_i(\mathcal{M})$, respectively. The mapping objective consists of static color reconstruction, reliability-weighted depth reconstruction, and a geometric regularization term:

$$\mathcal{L}_{\text{map}} = \alpha \mathcal{L}_{\text{rgb}} + (1-\alpha)\mathcal{L}_{\text{depth}} + \lambda_{\text{reg}} \mathcal{L}_{\text{reg}}. \tag{17}$$

The color term is defined as:

$$\mathcal{L}_{\text{rgb}} = \frac{\sum_{\mathbf{x}} M_i^{\text{rgb}}(\mathbf{x}) \left\| \hat{I}_i(\mathbf{x}) - I_i(\mathbf{x}) \right\|_1}{\sum_{\mathbf{x}} M_i^{\text{rgb}}(\mathbf{x}) + \varepsilon}, \tag{18}$$

and the depth term is defined as:

$$\mathcal{L}_{\text{depth}} = \frac{\sum_{\mathbf{x}} M_i^{d}(\mathbf{x})\, c_i(\mathbf{x})\, |\hat{z}_i(\mathbf{x}) - z_i(\mathbf{x})|}{\sum_{\mathbf{x}} M_i^{d}(\mathbf{x})\, c_i(\mathbf{x}) + \varepsilon}. \tag{19}$$

M_i^{rgb} denotes the valid static color mask, and M_i^d additionally requires the depth to satisfy multi-view support and numerical validity conditions; the current TUM dynamic configuration uses $\alpha = 0.8$.

Color and depth use different valid domains in order to distinguish appearance usability from geometric observability. Static background may lack reliable depth due to insufficient parallax, yet still contain usable color information. If depth confidence were directly used as an exclusion condition for all color supervision, insufficient geometric estimation would further manifest as missing background texture. Therefore, color supervision is constrained by static consistency, while depth supervision is modulated by both static validity and geometric reliability.

Incremental maps also exhibit significant history dependence. Gaussians generated by early erroneous observations may have already acquired high opacity, and even after they stop receiving new supervision, they continue to occlude the background. Dynamic handling therefore cannot act only on the current-frame loss; it must also correct the contribution of historical observations in the map. After the final geometric optimization and temporal re-examination, the system updates keyframe poses, depths, and static masks, rebuilds the Gaussian map from the retained static observations, and performs a final optimization. The formal run configuration adopts 20000 final optimization iterations.

Table 1: **Tracking performance on the Bonn RGB-D Dynamic Dataset** (ATE RMSE ↓ [cm]). Best results are highlighted as **first** and second. "-" indicates sequences without reported results or unavailable code. † denotes the backbone of our system, listed for reference.

Method	Balloon	Balloon2	Crowd	Crowd2	Person	Person2	Moving	Moving2	**Avg.**
DROID-SLAM†	7.5	4.1	5.2	6.5	4.3	5.4	2.3	4.0	4.91
Splat-SLAM	8.8	3.0	6.8	F	4.9	25.8	1.7	3.0	N/A
DynaMoN (MS)	6.8	3.8	6.1	5.6	2.4	3.5	1.4	2.6	4.02
DynaMoN (MS&SS)	2.8	2.7	3.5	2.8	14.8	2.2	1.3	2.7	4.10
TTT3R	21.5	15.4	9.8	7.7	30.0	21.4	33.4	41.2	22.55
MegaSaM	3.7	2.6	1.6	7.2	4.1	4.0	1.4	3.4	3.51
WildGS-SLAM	2.8	2.4	1.6	2.2	3.9	3.1	1.7	2.5	2.52
DROID-W	2.6	2.5	1.3	1.8	3.3	2.9	1.6	2.3	2.30
RU4D-SLAM	–	–	–	–	–	–	–	–	–
GGD-SLAM	–	–	–	–	–	–	–	–	–
RoSe-SLAM	–	–	–	–	–	–	–	–	–
MoPe	–	–	–	–	–	–	–	–	–
Ours	–	–	–	–	–	–	–	–	–

Table 2: **Tracking performance on the TUM RGB-D Dataset** (ATE RMSE ↓ [cm]). Best results are highlighted as **first** and second. "-" indicates sequences without reported results or unavailable code. † denotes the backbone of our system, listed for reference.

Method	f2/dp	f3/ss	f3/sx	f3/sr	f3/shs	f3/ws	f3/wx	f3/wr	f3/whs	**Avg.**
DROID-SLAM†	0.6	0.5	0.9	2.2	1.4	1.2	1.6	4.0	2.2	1.62
Splat-SLAM	0.7	0.5	0.9	2.3	1.5	2.3	1.3	3.9	2.2	1.71
DynaMoN (MS)	0.6	0.5	0.9	2.1	1.9	1.4	1.4	3.9	2.0	1.63
DynaMoN (MS&SS)	0.7	0.5	0.9	2.4	2.3	0.7	1.4	3.9	1.9	1.63
TTT3R	113.1	3.1	5.8	6.4	24.9	2.0	24.7	15.9	23.1	24.33
WildGS-SLAM	1.4	0.5	0.8	2.4	2.0	0.4	1.3	3.3	1.6	1.51
DROID-W	1.1	0.5	0.8	2.1	1.4	0.5	1.2	3.1	1.6	1.36
RU4D-SLAM	–	–	–	–	–	–	–	–	–	–
GGD-SLAM	–	–	–	–	–	–	–	–	–	–
RoSe-SLAM	–	–	–	–	–	–	–	–	–	–
MoPe	–	–	–	–	–	–	–	–	–	–
Ours	–	–	–	–	–	–	–	–	–	–

To prevent discrete keyframe updates from affecting result observation, mapping nodes and rendering cameras are scheduled separately. Non-keyframes have their poses completed by motion-only BA and are rendered sequentially without adding permanent mapping nodes. During the final optimization, the preview viewpoint is fixed; after the optimization, the final map and the complete trajectory in the same coordinate frame are used to generate a continuous image sequence. This process enables continuous observation of map quality over the full trajectory, but it does not alter the limits of geometric observability: regions with long-term occlusion, degenerate camera motion, or a lack of valid correspondences may still exhibit holes, blurring, or local reconstruction errors.

3 EXPERIMENTS

AI USE STATEMENT

(This section is **required** and does not count toward the page limit.)

In this work, we used generative AI tools for [tasks with required disclosure]. We have not used generative AI tools for [other tasks with required disclosure], and [the rest of the required disclosure tasks] are not applicable to this work. Additionally, we used generative AI tools for [tasks with recommended disclosure]. We have reviewed all AI-assisted work. [Elaborate. For example, "we checked LLM-generated research ideas for potential plagiarism through a manual literature survey", "LLM-generated code was verified and tested for correctness by 2 authors", etc.]. We take respon-

Table 3: **Rendering quality on dynamic sequences of TUM and Bonn** (PSNR [dB] ↑ / SSIM ↑ / LPIPS ↓). Best results are in **bold**, second best underlined. "-" indicates unreported results. Values for MonoGS, SplaTAM and WildGS-SLAM are transcribed from GGD-SLAM (ICRA 2026); verify sequence split before submission.

	f3/w/xyz			f3/w/static			f3/w/rpy			f3/w/half			**Avg.**		
Method	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
MonoGS	13.97	0.467	0.591	16.57	0.551	0.476	19.56	0.726	0.454	18.07	0.695	0.567	17.04	0.610	0.522
SplaTAM	14.29	0.527	0.486	15.53	0.478	0.493	17.81	0.596	0.344	15.79	0.517	0.592	15.81	0.526	0.479
WildGS-SLAM	21.53	0.821	0.141	20.89	0.829	0.136	23.35	0.904	0.148	23.15	0.885	0.264	22.23	0.855	0.172
GGD-SLAM	22.68	0.751	0.117	22.25	0.848	0.134	22.92	0.902	0.162	24.27	0.918	0.219	23.03	0.859	0.158
RU4D-SLAM	–	–	–	–	–	–	–	–	–	–	–	–	25.95	–	–
DROID-W	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Ours	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–

Table 4: **Efficiency comparison**: average FPS ↑ and peak GPU memory (MB) ↓. All methods are evaluated on the same hardware.

Method	External module	Bonn (FPS)	TUM (FPS)	Memory (MB)
DROID-SLAM	None	19.89	26.97	–
WildGS-SLAM	Uncertainty MLP (DINOv2)	0.22	0.32	–
DROID-W	Monocular depth + DINOv2	10.57	14.92	–
Ours	**None**	–	–	–

sibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

See the ICLR 2027 AI Policy for Authors for more details. This statement should not be more than 1 page.

Ethics statement

(This section is **recommended** and does not count toward the page limit.)

If authors feel that their paper submission raises questions regarding the Code of Ethics, they are encouraged to include a paragraph of Ethics Statement (at the end of the main text before references) to address potential concerns where appropriate. Topics include, but are not limited to, studies that involve human subjects, practices to data set releases, potentially harmful insights, methodologies and applications, potential conflicts of interest and sponsorship, discrimination/bias/fairness concerns, privacy and security issues, legal compliance, and research integrity issues (e.g., IRB, documentation, research ethics). This statement should not be more than 1 page.

Reproducibility statement

(This section is **recommended** and does not count toward the page limit.)

It is important that the work published in ICLR is reproducible. Authors are strongly encouraged to include a paragraph-long Reproducibility Statement at the end of the main text (before references) to discuss the efforts that have been made to ensure reproducibility. This paragraph should not itself describe details needed for reproducing the results, but rather reference the parts of the main paper, appendix, and supplemental materials that will help with reproducibility. For example, for novel models or algorithms, a link to an anonymous downloadable source code can be submitted as supplementary materials; for theoretical results, clear explanations of any assumptions and a complete proof of the claims can be included in the appendix; for any datasets used in the experiments, a complete description of the data processing steps can be provided in the supplementary materials. Each of the above are examples of things that can be referenced in the reproducibility statement.

Author Contributions

If you'd like to, you may include a section for author contributions as is done in many journals. This is optional and at the discretion of the authors.

Table 5: **Ablation studies** on the Bonn and TUM datasets (average ATE RMSE ↓ [cm]). Each component of the proposed framework is disabled in turn.

Configuration	Bonn	TUM
a. w/o epipolar consistency	–	–
b. w/o multi-edge evidence aggregation	–	–
c. w/o curvature-based reliability	–	–
d. w/o temporal re-examination	–	–
e. w/o reliability-weighted depth loss	–	–
Full model	–	–

ACKNOWLEDGMENTS

Use unnumbered third level headings for the acknowledgments. All acknowledgments, including those to funding agencies, go at the end of the paper.

REFERENCES

A APPENDIX

You may include other additional sections here.