MOTIONAWAREGS-SLAM: TEMPORAL-CONSISTENCY-AWARE OBSERVATION INTEGRATION FOR DYNAMIC GAUSSIAN SLAM

Anonymous authors
Paper under double-blind review

ABSTRACT

Observations useful for camera tracking and those suitable for persistent Gaussian mapping operate at fundamentally different temporal scales: a transient observation may provide a reliable pose constraint for the current frame, yet become temporally inconsistent when integrated into a persistent Gaussian map. Existing dynamic Gaussian SLAM methods mainly determine observation usability from current-frame reliability, implicitly assuming that observations beneficial for tracking are also suitable for persistent map integration. This mismatch can introduce temporally unstable information into long-term Gaussian representations even when camera tracking remains accurate. We propose MotionAwareGS-SLAM, a temporal-consistency-aware observation integration framework that incorporates temporal consistency evidence into dynamic Gaussian mapping. Specifically, we extract multi-frame temporal consistency evidence, regulate observation participation in persistent Gaussian updates, and perform asymmetric observation routing between camera tracking and persistent mapping. This design allows transient observations to contribute to robust tracking while preventing temporally unstable observations from degrading persistent Gaussian representations. Experiments on four TUM RGB-D dynamic sequences demonstrate the tracking–mapping reliability mismatch through independent future-only RGB-D and official-pose analysis, and validate the proposed observation routing through instrumented tracking and mapping pipelines. The results show that MotionAwareGS-SLAM improves persistent scene consistency while maintaining accurate camera trajectory estimation in dynamic environments.

1 INTRODUCTION

3D Gaussian Splatting (3DGS) gives SLAM explicit primitives for differentiable rendering and dense reconstruction, while persistent updates accumulate observations across frames (Kerbl et al., 2023; Huang et al., 2024; Keetha et al., 2024; Yugay et al., 2023; Yan et al., 2024; Matsuki et al., 2024). In dynamic scenes, a transient surface can constrain the current camera pose yet become inconsistent when fused into a map reused by later frames. The central problem is therefore the temporal contract governing which observations are allowed to alter a persistent Gaussian representation.

Dynamic SLAM uses motion cues, semantics, object reasoning, and learned features to reduce interference from moving content (Scona et al., 2018; Runz et al., 2018; Bescos et al., 2018; Yu et al., 2018; Ji et al., 2021). Dynamic Gaussian methods apply such cues to decide which observations enter tracking or mapping (Kong et al., 2024; Wen et al., 2025; Zheng et al., 2025; Li et al., 2025b; Xu et al., 2024; Li et al., 2025a; Liu et al., 2025). DL-SLAM already distinguishes pixel-level pose evidence from object-level pruning in a static Gaussian map (Xu et al., 2026). Our analysis is narrower and temporal: it tests whether current pose usability predicts future compatibility with a persistent map. We use frozen pretrained features and recent observation history, without dynamic class labels, to inform that integration decision.

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



Figure 1: Conceptual comparison of dynamic Gaussian SLAM pipelines. Existing methods commonly pass current-frame reliability to both tracking and mapping, whereas MotionAwareGS-SLAM combines current observation support with temporal consistency evidence and routes observations according to their distinct tracking and persistent-integration roles. The imagery and overlays are illustrative.

We call the difference between current usability and later compatibility the Tracking–Mapping Reliability Mismatch. Current acceptance is an operational proxy for immediate utility, whereas an independent future-only RGB-D/official- pose reference measures compatibility with later geometry. The two tests can disagree even when an observation satisfies the current-mask rule. This diagnosis creates two linked design requirements: persistent integration needs its own temporal condition, and tracking must retain measurements that still constrain the current pose.

MotionAwareGS-SLAM instantiates Temporal-Consistency-Aware Observation Integration (Fig. 1) as a linked pipeline. Temporal Consistency Evidence Extraction uses frozen DINOv2 features and temporal attention to estimate cross-frame stability. Asymmetric Observation Routing then assigns distinct tracking and persistent-mapping usage, while mapper-side sinks enforce the same eligibility at Gaussian creation, update, and densification. Future-only diagnostics, route-removal tests, temporal-history ablations, and controlled map evaluation measure these successive links. The resulting evidence is a local background-recovery benefit accompanied by sequence-dependent tracking effects and a visible-static rendering cost.

The contributions of this work are summarized as follows:

- We identify a tracking–mapping reliability mismatch: 4.62%–41.85% of currently accepted reference-labeled pixels are future-inconsistent across four dynamic sequences under an independent future-only diagnostic.
- We introduce Temporal-Consistency-Aware Observation Integration, combining frozen-feature evidence with asymmetric routing; instrumented consumers verify the resulting route execution.
- We enforce integration at Gaussian update sinks and quantify its trade-off: in a three-seed walking_xyz window, routing raises revealed-background PSNR by 7.977 dB and reduces ghost excess by 20.095/255, with a 0.573 dB static cost.

2 Related Work

2.1 3D Gaussian Splatting-Based SLAM

3DGS provides explicit primitives and efficient differentiable rendering (Kerbl et al., 2023). Photo-SLAM combines feature-based tracking with Gaussian mapping; SplaTAM, Gaussian-SLAM, GS-SLAM, and Gaussian Splatting SLAM couple camera estimation with rendered or geometric objectives (Huang et al., 2024; Keetha et al., 2024; Yugay et al., 2023; Yan et al., 2024; Matsuki et al., 2024). These systems show that Gaussian maps can support dense reconstruction and pose estimation. They also make observation admission consequential: a measurement used to update a persistent primitive can affect later rendering and tracking, beyond its contribution to the current frame.

Most of these mapping formulations assume that incorporated observations remain compatible with a static scene (Huang et al., 2024; Keetha et al., 2024; Yugay et al., 2023; Yan et al., 2024; Matsuki et al., 2024). In a dynamic scene, locally useful observations can become inconsistent with later views, leaving foreground traces in a persistent map. The relevant decision is therefore not simply whether a frame can be tracked, but whether its observations should create, update, or densify lasting Gaussians. Our method treats temporal consistency as a condition on those map operations.

2.2 Dynamic SLAM in Dynamic Environments

Dynamic SLAM detects moving content through multi-view geometry, motion estimation, semantic classes, or instance cues (Scona et al., 2018; Runz et al., 2018; Bescos et al., 2018; Yu et al., 2018; Ji et al., 2021). These approaches reduce the influence of moving regions when a static-scene assumption breaks down. Their cues differ in the observations they classify and in how those observations are consumed. The present analysis targets a later decision: the future compatibility of a currently accepted measurement with a map whose primitives must remain useful across future frames.

Dynamic Gaussian SLAM combines these cues with explicit map optimization (Kong et al., 2024; Wen et al., 2025; Zheng et al., 2025; Li et al., 2025b; Xu et al., 2024; Li et al., 2025a; Liu et al., 2025). DL-SLAM distinguishes pixel-level pose evidence from object-level pruning (Xu et al., 2026); MotionAwareGS-SLAM instead estimates cross-frame feature stability without dynamic class labels and assigns observation-level authority to tracking and persistent mapping separately. The comparison is organized around the integration rule, with quantitative results reported under the documented local protocols.

2.3 Observation Selection and Map Integration in SLAM

SLAM systems select observations through feature association, keyframes, support masks, confidence, and residual tests (Campos et al., 2021; Huang et al., 2024; Keetha et al., 2024). These choices manage noise and computational cost in pose estimation and map optimization. Gaussian mapping adds a persistent consequence: accepted measurements can create, update, or densify primitives that later frames reuse. Immediate estimation suitability and long-term integration suitability are therefore distinct decisions, even when they share an input mask.

Current-stage selection does not by itself measure compatibility with later observations. Our integration policy uses frozen pretrained features and a recent-frame memory to estimate that compatibility, then retains distinct tracking and mapping routes. The future-only reference is an independent diagnostic rather than an online label, and mapper instrumentation tests whether the chosen routes reach persistent update operations. This separates evidence construction, routing, and execution while keeping the feature score separate from a calibrated motion probability (Campos et al., 2021; Huang et al., 2024; Keetha et al., 2024).

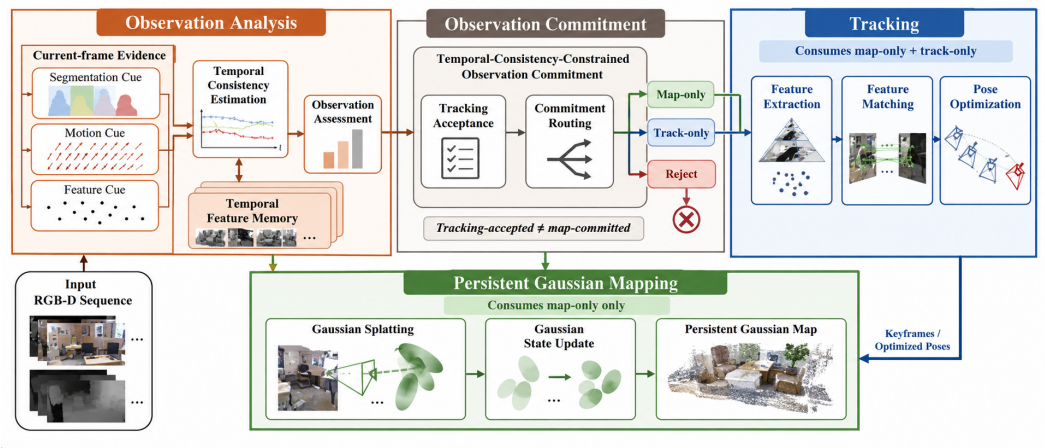


Figure 2: Overview of MotionAwareGS-SLAM. Current-frame cues and feature memory provide temporal evidence; asymmetric routing combines it with current support to assign tracking availability and mapping eligibility. The integration gate enforces mapping eligibility at Gaussian creation, update, and densification while track-only observations remain available for tracking.

3 METHOD

3.1 OVERVIEW OF MOTIONAWAREGS-SLAM

Given synchronized RGB-D frames $\{(I_t, D_t)\}_{t=1}^{T}$, MotionAwareGS-SLAM estimates camera poses while maintaining a persistent 3D Gaussian map. Current observation support determines whether an input can constrain the present pose; temporal evidence additionally determines its authority to alter that map. Temporal-consistency-aware observation integration therefore separates two consumer decisions that a shared current-frame mask would otherwise couple.

Figure 2 follows the execution path. Temporal Consistency Evidence Extraction combines current cues with a multi-frame feature memory. Asymmetric Observation Routing combines this evidence with current support and assigns mapping eligibility and tracking availability. Temporal-Consistency-Aware Observation Integration enforces the resulting mask at Gaussian creation, update, and densification sinks. Tracking can still consume accepted observations that are withheld from the persistent map.

3.2 TRACKING–MAPPING RELIABILITY MISMATCH

Let $A_t(p) = 1$ denote that observation p satisfies the current acceptance condition. Let $S_{t+}(p) = 1$ denote compatibility with an independent future-only RGB-D/official-pose reference. The reference measures temporal compatibility under its geometric construction. The reliability-mismatch set is

$$\mathcal{F}_t = \{p \mid A_t(p) = 1,\ S_{t+}(p) = 0\}. \tag{1}$$

Observations in $\mathcal{F}_t$ are currently accepted but future-inconsistent under the geometric reference. Here acceptance proxies current usability; it does not measure each pixel's marginal contribution to pose accuracy. Sharing that decision with mapping may integrate evidence before its later compatibility has been established. Persistent Gaussian updates thus require a separate temporal condition.

The future-only RGB-D and official-pose analysis in Sec. IV measures $\mathcal{F}_t$ independently of candidate masks, mapper counters, and renders. It is an offline diagnostic, not an input to the online route. At runtime, recent feature history provides the temporal evidence used below.

3.3 TEMPORAL CONSISTENCY EVIDENCE EXTRACTION

We extract cross-frame feature evidence without dynamic class labels. The local real-mapper entrypoint defaults to the frozen `facebook/dinov2-small` encoder (Oquab et al., 2024), which

maps frame I_t to a feature grid $F_t \in \mathbb{R}^{H_f \times W_f \times C}$. Memory $\mathcal{M}_t = \{F_{t-L}, \ldots, F_{t-1}\}$ stores the latest L grids. After resizing and normalizing historical features, temporal attention computes similarity $s_t(p)$, residual distance $d_t(p)$, and entropy $h_t(p)$.

These signals define temporal inconsistency evidence and its complementary consistency prior:

$$\ell_t(p) = \alpha(1 - s_t(p)) + \beta d_t(p) + \gamma h_t(p), \tag{2}$$

$$U_t(p) = \sigma(\ell_t(p)), \qquad C_t(p) = 1 - U_t(p), \tag{3}$$

where U_t is a feature-derived integration-risk score. The coefficients α, β, γ default to 4.5, 2.4, and 0.5 in the real-mapper entrypoint; they are fixed rather than learned in these experiments. The full configuration uses $L = 6$; ablations test $L = 0$, $L = 1$, $L = 6$, and a last-frame no-attention $L = 6$ variant. The score is not a calibrated motion probability; learning its weights is a future extension, not a reported result.

3.4 Asymmetric Observation Routing

For each observation, the framework assigns

$$r_t(p) \in \{\text{map-only}, \text{track-only}, \text{reject}\}. \tag{4}$$

The implementation name *map-only* means mapping-authorized, not tracking-forbidden: these observations can also support tracking. *track-only* observations remain available for pose estimation but cannot modify persistent Gaussians; *reject* observations enter neither path. Current support and the temporal prior jointly determine these routes, which are materialized separately for consumer-side audit.

The mapper exclusion mask is the complement of integration eligibility:

$$M_t^{\text{exclude}}(p) = 1 - \mathbb{1}[r_t(p) = \text{map-only}]. \tag{5}$$

It is exported as a binary image together with per-frame route fractions, temporal evidence statistics, and filename-to-frame-index mappings.

3.5 Temporal-Consistency-Aware Observation Integration

Persistent updates consume only mapping-authorized observations. The mapper excludes track-only observations from RGB-D, Gaussian-depth, and final-update sinks while the frontend retains their tracking support; reject observations enter neither consumer. Creation, update, and densification are thus governed by the same integration eligibility.

Hard commitment loads the exclusion mask for every processed frame and blocks ineligible Gaussian operations. The instrumented mapper records mask requests and loads, blocked operations, non-finite values, and creation/update/ densification leaks. These records verify execution at the measured consumers; the independent future reference and rendering study address different links of the evidence chain.

The mapper-side DynamicConsensus path provides a softer integration probe. A float-confidence variant based on C_t fails the trajectory gate on fr3/walking_half. We report that failed variant alongside the discrete and consensus routes rather than treating a continuous confidence score as an established improvement.

4 Experiments

4.1 Experimental Setup

We evaluate three aspects: *(Q1)* the effect of routing on camera tracking; *(Q2)* changes in persistent-map usage and rendering; and *(Q3)* agreement between the independent future reference and the instrumented consumer routes. The benchmark contains four TUM Freiburg3 dynamic RGB-D sequences (Sturm et al., 2012): walking_xyz, walking_half, walking_static, and sitting_half. Tracking uses official poses, association within 0.01 s, and fixed-scale SE(3) Umeyama alignment.

Table 1: Local tracking accuracy on four TUM RGB-D dynamic sequences: ATE RMSE in cm (↓). Values come from the runs plotted in Fig. 3 under one shared local evaluator.

Method	fr3/w/xyz	fr3/w/half	fr3/w/static	fr3/s/half
Unweighted baseline	1.532	2.419	0.769	1.877
MotionAware: hard	1.624	2.612	0.858	1.871
MotionAware: consensus	1.538	2.630	0.653	1.835

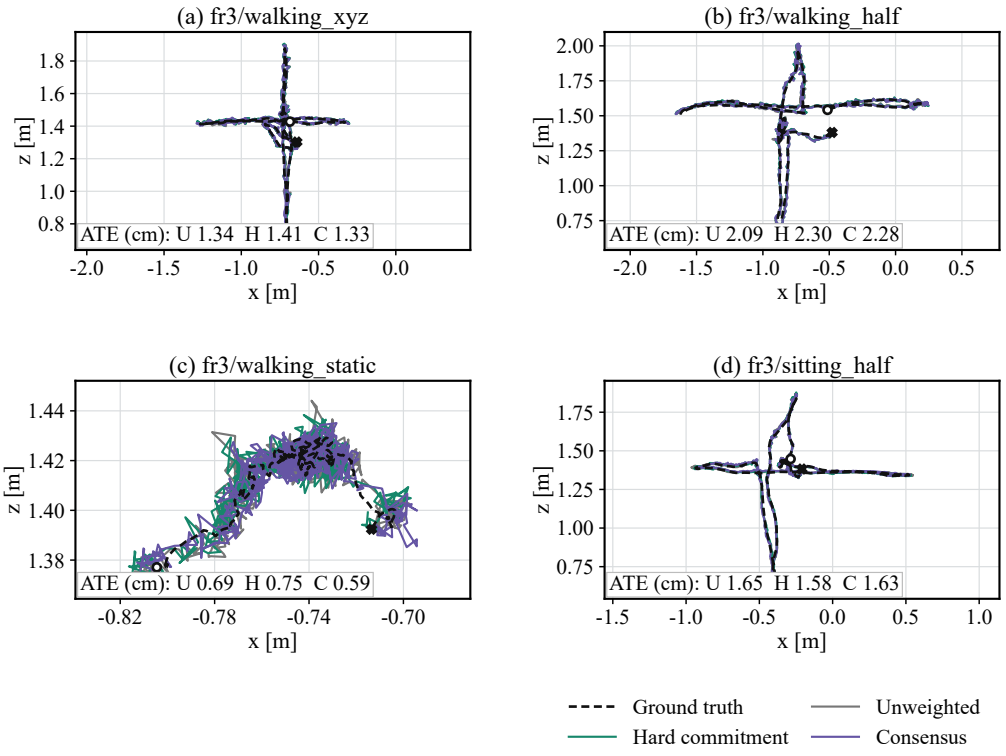


Figure 3: Camera trajectory comparison on four TUM RGB-D dynamic sequences. The three local variants are associated with official poses within 0.01 s and aligned using fixed-scale SE(3) Umeyama, with no scale correction. Insets report mean translational ATE in cm for U (unweighted), H (hard commitment), and C (consensus); Table 1 reports RMSE for the same runs. The overlays support centimeter-level trajectory preservation under this local protocol.

Four evidence scopes remain separate. The 12 representative local runs compare unweighted integration, hard commitment, and mapper-side consensus; published external values are retained separately in Appendix Table 6. The 24-run feature and 36-run fixed-mapper-seed repeat packages assess system behavior separately. A paired track-only removal package contains eight online runs, and the temporal ablation contains 32 four-sequence runs. The offline future reference retains 165 timestamps across four sequences. The controlled route/map study covers 84 valid timestamps in one requested 120-frame walking_xyz window with distinct mapper seeds 0, 11, and 22. Geometry-derived labels provide the stated temporal and map references.

4.2 Local Tracking Evaluation

Table 1 compares the three local variants under one evaluator. Hard commitment raises ATE RMSE on all three walking sequences and slightly lowers it on sitting_half; consensus lowers RMSE on walking_static and sitting_half but raises it on walking_xyz and walking_half. Figure 3 overlays the corresponding trajectories with official ground truth. Its insets show mean translational error,

Table 2: Paired track-only removal. Positive Δ means removal worsens the metric. The arms share binary, seed, sequence, initialization, and Gaussian-admission masks; the mapper runs online in both arms.

Sequence	Full ATE	Removed ATE	ΔATE	ΔRPE	ΔORB
walking_xyz	1.391	1.580	+0.189	+0.243	-7,396
walking_half	2.242	2.769	+0.527	+0.576	-46,566
walking_static	0.813	0.640	-0.173	-0.038	-22
sitting_half	1.818	2.750	+0.932	+0.588	-22,411

Table 3: Mapping quality and map size for the local runs. U: unweighted; H: hard commitment; C: consensus. Gaussian count reports the final number of vertices.

Sequence	Variant	PSNR ↑ [dB]	SSIM ↑	Gaussians
fr3/w/xyz	U	16.589	0.6016	40,785
	H	16.047	0.5759	24,665
	C	16.303	0.5922	35,335
fr3/w/half	U	16.327	0.6104	63,658
	H	16.051	0.5826	29,323
	C	16.286	0.5980	51,981
fr3/w/static	U	14.848	0.6608	33,636
	H	14.867	0.6561	24,924
	C	14.871	0.6519	30,921
fr3/s/half	U	18.773	0.6882	59,131
	H	18.445	0.6782	33,985
	C	18.044	0.6466	49,097

whereas the table reports RMSE from the same records. Published external values appear only in Appendix Table 6, under different execution protocols.

All 24 feature runs complete with maximum per-run mean translational error below 2.31 cm; the 36-run fixed-mapper-seed package records a 2.35 cm maximum. The local RMSE changes in both directions, supporting centimeter-level trajectories rather than a uniform tracking improvement.

Table 2 tests the system effect of retaining track-only support. Removing it worsens ATE on walking_xyz, walking_half, and sitting_half, but improves walking_static. All eight online runs remain finite and pass the diagnostic stability gate. Both arms evolve separate maps; the comparison therefore measures the net effect of the frontend intervention, including downstream map changes, rather than a frozen-map marginal effect.

4.3 Mapping Quality Evaluation

Table 3 reports full-frame rendering and final Gaussian counts for the same local runs. Hard commitment stores fewer Gaussians on all four sequences; its SSIM is lower on all four, and PSNR is lower on three. The separate repeat package reports mean PSNR/SSIM of 17.06/0.654 for unweighted, 16.26/0.618 for hard commitment, and 16.53/0.630 for consensus. Selective integration thus changes map size and sacrifices full-frame fidelity in most measured settings. Region-specific background and ghost metrics below test the complementary persistent-map behavior.

Appendix D shows a 4×5 cross-method mapping grid. Each row uses one physical timestamp selected using an independent future-only reference and MotionAware renders. MotionAware has the lowest ghost excess and highest static PSNR on walking_xyz and walking_static; Photo-SLAM+Mask leads on walking_half, and SplaTAM on sitting_half. Different checkpoint horizons make this a qualitative context panel, not a matched method ranking.

Table 4: Mechanism validation at instrumented consumers in one 120-frame walking_xyz window. Percentages pool 84 valid timestamps over mapper seeds 0, 11, and 22; ATE RMSE is reported as mean ± standard deviation across seeds. Stable/unstable labels are independent geometry-derived references.

Method	Unstable tracked [%]	Unstable → map-only [%]	Unstable → track-only [%]	Unstable → reject [%]	Stable → map-only [%]	ATE RMSE ↓ [cm]
All-map control	100.00	100.00	0.00	0.00	100.00	38.662 ± 2.514
Object route	94.40	2.62	91.78	5.60	98.45	1.626 ± 0.069

Table 5: Temporal-history and attention ablation averaged over four sequences. Each cell contains four successful runs. ATE and RPE are in cm; PSNR is in dB.

Setting	Hard		Consensus		Hard render		Consensus render	
	ATE	RPE	ATE	RPE	PSNR	SSIM	PSNR	SSIM
Attention $L = 0$	1.462	1.572	1.432	1.616	16.776	0.647	16.596	0.635
Attention $L = 1$	1.400	1.556	1.405	1.581	16.522	0.634	16.570	0.630
Attention $L = 6$	1.453	1.605	1.349	1.554	16.385	0.621	16.372	0.627
Last-frame $L = 6$	1.421	1.625	1.371	1.576	16.525	0.629	16.526	0.631

The controlled walking_xyz window compares object routing with all-map integration over mapper seeds 0, 11, and 22. Routing raises revealed-background PSNR by 7.977 dB and reduces ghost excess by 20.095/255, while visible-static PSNR falls by 0.573 dB. Paired hierarchical 95% intervals are $[7.61, 8.33]$ dB, $[18.43, 21.65]$, and $[-0.98, -0.07]$ dB, respectively. This matched local comparison captures the trade-off between revealed-background recovery and static-region fidelity; it does not replace the mixed full-frame results.

4.4 Mechanism Validation

Future-consistency analysis. We test whether current acceptance implies future compatibility. RGB-D frames at offsets $20, 40, 80, 120, 180$ are warped to each target using official poses. A pixel is future-inconsistent if its current depth is at least 0.12 m closer than a future-supported background surface, and future-consistent if it lies within 0.06 m. Labels require two future supports and a 3×3 erosion. The reference reads no candidate masks, routes, mapper counters, or renders. Here A_t is a current-mask acceptance proxy, not measured pose usefulness.

Appendix Fig. 4 records future inconsistency in 4.62%–41.85% of currently accepted, reference-labeled pixels, or 4,518–303,041 conflict pixels per sequence. With temporal integration, the three walking ratios become 19.84%, 40.74%, and 30.59%; sitting-half changes from 4.62% to 4.83%. Each policy has its own accepted-pixel denominator. The diagnostic establishes a nonzero mismatch under the geometry-derived reference and exposes a sitting-half boundary case.

Routing and update execution. The real-consumer study materializes frontend tracking support and instruments persistent RGB-D, Gaussian-depth, and final-update paths. In the measured window, Table 4 shows that routing retains 94.40% of independently unstable support for tracking, sends 91.78% to track-only, and sends 2.62% to mapping-authorized map-only. It retains 98.45% of stable support for mapping. These percentages verify asymmetric usage at the instrumented consumers; they do not label each feature anomaly as motion.

Temporal-context variants also change commitment behavior. The last-frame no-attention $L = 6$ and attention $L = 1$ variants each block 180.13M operations, versus 291.70M for attention $L = 6$; masks load for all 3,639 processed frames. The no-attention variant summarizes the last frame, so this contrast does not isolate attention alone. Operation counts describe mapper activity, not unique dynamic pixels. The future-only diagnostic and consumer verification support different links of the integration argument.

Table 5 shows a non-monotone response to temporal history. Attention $L = 6$ has the lowest consensus ATE; attention $L = 1$ has the lowest hard ATE, and last-frame $L = 6$ also improves hard ATE relative to attention $L = 6$. Longer memory lowers full-frame PSNR/SSIM in these

aggregates. The ablation establishes a tracking–rendering operating-point change, not an isolated attention benefit. Region-specific ghost comparisons across these variants remain to be measured.

4.5 Runtime and Resource Profile

For an $H \times W$ image, temporal matching costs $O(LNC)$ time and history storage and routing costs $O(HW)$; the full setting uses $L = 6$. Across 3,639 profiled frames, the DINOv2 frontend records 71.05 ms mean, 63.54 ms median, and 78.48 ms 95th-percentile latency. Separate jobs reach 18.58, 17.26, and 18.52 frame/s for unweighted, hard, and consensus backends; hard commitment adds 7.7% wall time while saving 154 MiB peak VRAM. Hardware identity was not retained, so these are component profiles rather than end-to-end throughput.

5 Discussion and Limitations

The future-only diagnostic measures geometry compatibility, not pose utility; frozen DINOv2 supplies integration-risk evidence, not calibrated motion probabilities. Consumer checks cover instrumented paths, while track-only and history ablations use evolving maps and full-frame metrics. The three-seed walking_xyz study improves revealed-background and ghost metrics at a 0.573 dB static cost; full-frame results are mixed, sitting-half conflict rises slightly, and transfer beyond indoor RGB-D remains untested.

6 Conclusion

MotionAwareGS-SLAM separates tracking from persistent Gaussian integration. Across four dynamic sequences, a future-only RGB-D/official-pose diagnostic finds 4.62%–41.85% conflict among accepted reference-labeled pixels. Temporal evidence and asymmetric routing preserve tracking support while controlling mapper updates. In a three-seed walking_xyz window, routing raises revealed-background PSNR by 7.977 dB and reduces ghost excess by 20.095/255, at a 0.573 dB visible-static cost. The evidence supports a local recovery–fidelity trade-off rather than uniform improvement.

References

Berta Bescos, Jose M. Facil, Javier Civera, and Jose Neira. Dynaslam: Tracking, mapping, and inpainting in dynamic scenes. *IEEE Robotics and Automation Letters*, 3(4):4076–4083, 2018. doi: 10.1109/LRA.2018.2860039. URL https://dblp.org/rec/journals/ral/BescosFCN18.html.

Carlos Campos, Richard Elvira, Juan J. Gomez Rodriguez, Jose M. M. Montiel, and Juan D. Tardos. Orb-slam3: An accurate open-source library for visual, visual-inertial, and multi-map slam. *IEEE Transactions on Robotics*, 2021. doi: 10.1109/TRO.2021.3075644. URL https://arxiv.org/abs/2007.11898.

Huajian Huang, Longwei Li, Hui Cheng, and Sai-Kit Yeung. Photo-slam: Real-time simultaneous localization and photorealistic mapping for monocular, stereo, and rgb-d cameras. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://arxiv.org/abs/2311.16728.

Tete Ji, Chen Wang, and Lihua Xie. Towards real-time semantic rgb-d slam in dynamic environments. In *IEEE International Conference on Robotics and Automation*, 2021. URL https://arxiv.org/abs/2104.01316.

Nikhil Keetha, Jay Karhade, Krishna Murthy Jatavallabhula, Gengshan Yang, Sebastian Scherer, Deva Ramanan, and Jonathon Luiten. Splatam: Splat, track and map 3d gaussians for dense rgb-d slam. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://arxiv.org/abs/2312.02126.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkuehler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 2023. URL https://arxiv.org/abs/2308.04079.

Mangyu Kong, Jaewon Lee, Seongwon Lee, and Euntai Kim. Dgs-slam: Gaussian splatting slam in dynamic environment. *arXiv preprint arXiv:2411.10722*, 2024. URL https://arxiv.org/abs/2411.10722.

Mingrui Li, Yiming Zhou, Hongxing Zhou, Xinggang Hu, Florian Roemer, Hongyu Wang, and Ahmad Osman. Dy3dgs-slam: Monocular 3d gaussian splatting slam for dynamic environments. *arXiv preprint arXiv:2506.05965*, 2025a. URL https://arxiv.org/abs/2506.05965.

Runfa Blark Li, Mahdi Shaghaghi, Keito Suzuki, Xinshuang Liu, Varun Moparthi, Bang Du, Walker Curtis, Martin Renschler, Ki Myung Brian Lee, Nikolay Atanasov, and Truong Nguyen. Dynagslam: Real-time gaussian-splatting slam for online rendering, tracking, motion predictions of moving objects in dynamic scenes. *arXiv preprint arXiv:2503.11979*, 2025b. URL https://arxiv.org/abs/2503.11979.

Yi Liu, Keyu Fan, Bin Lan, and Houde Liu. Dypho-slam: Real-time photorealistic slam in dynamic environments. *arXiv preprint arXiv:2509.00741*, 2025. URL https://arxiv.org/abs/2509.00741.

Hidenobu Matsuki, Riku Murai, Paul H. J. Kelly, and Andrew J. Davison. Gaussian splatting slam. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://arxiv.org/abs/2312.06741.

Maxime Oquab, Timothee Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Martin Runz, Maud Buffier, and Lourdes Agapito. Maskfusion: Real-time recognition, tracking and reconstruction of multiple moving objects. In *IEEE International Symposium on Mixed and Augmented Reality*, 2018. URL https://arxiv.org/abs/1804.09194.

Raluca Scona, Mariano Jaimez, Yvan R. Petillot, Maurice Fallon, and Daniel Cremers. Staticfusion: Background reconstruction for dense rgb-d slam in dynamic environments. In *IEEE International Conference on Robotics and Automation*, 2018. doi: 10.1109/ICRA.2018.8460681. URL https://dblp.org/rec/conf/icra/SconaJPFC18.

Juergen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of rgb-d slam systems. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 573–580, 2012. doi: 10.1109/IROS.2012.6385773. URL https://ieeexplore.ieee.org/document/6385773.

Long Wen, Shixin Li, Yu Zhang, Yuhong Huang, Jianjie Lin, Fengjunjie Pan, Zhenshan Bing, and Alois Knoll. Gassidy: Gaussian splatting SLAM in dynamic environments. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 8471–8477, 2025. doi: 10.1109/ICRA55743.2025.11127678.

Yueming Xu, Haochen Jiang, Zhongyang Xiao, Jianfeng Feng, and Li Zhang. DG-SLAM: Robust dynamic gaussian splatting SLAM with hybrid pose optimization. In *Advances in Neural Information Processing Systems*, volume 37, pp. 51577–51596, 2024. doi: 10.52202/079017-1633.

Ziheng Xu, Qingfeng Li, Xuefeng Liu, Chen Chen, and Jianwei Niu. Dl-slam: Enabling high-fidelity gaussian splatting slam in dynamic environments based on dual-level probability. *arXiv preprint arXiv:2607.01860*, 2026. URL https://arxiv.org/abs/2607.01860.

Chi Yan, Delin Qu, Dan Xu, Bin Zhao, Zhigang Wang, Dong Wang, and Xuelong Li. Gs-slam: Dense visual slam with 3d gaussian splatting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://arxiv.org/abs/2311.11700.

Chao Yu, Zuxin Liu, Xinjun Liu, Fugui Xie, Yi Yang, Qi Wei, and Qiao Fei. Ds-slam: A semantic visual slam towards dynamic environments. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2018. URL https://arxiv.org/abs/1809.08379.

Vladimir Yugay, Yue Li, Theo Gevers, and Martin R. Oswald. Gaussian-slam: Photo-realistic dense slam with gaussian splatting. *arXiv preprint arXiv:2312.10070*, 2023. URL https://arxiv.org/abs/2312.10070.

Jianhao Zheng, Zihan Zhu, Valentin Bieri, Marc Pollefeys, Songyou Peng, and Iro Armeni. Wildgs-slam: Monocular gaussian splatting slam in dynamic environments. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://arxiv.org/abs/2504.03886.

A Reproducibility and AI Use

The paper specifies temporal evidence, route eligibility, mapper sinks, sequences, seeds, alignment, history variants, and the future-only reference. ChatGPT and Codex assisted with language, LaTeX, code, and illustrative figures; the authors verified all content.

B Future-Consistency Diagnostic

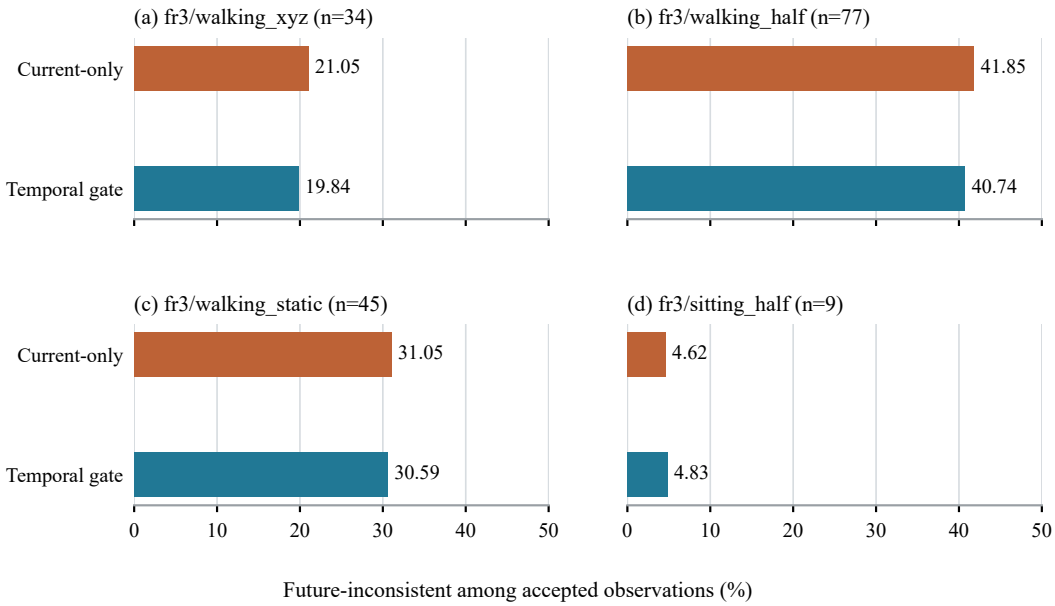


Figure 4: Future-consistency analysis. Each bar reports $|A_t \cap \neg S_{t+}|/|A_t|$ among accepted, geometry-referenced pixels. The four sequences retain 34, 77, 45, and 9 timestamps (165 total). The ratios decrease on the walking sequences and increase slightly on sitting-half under temporal integration; each policy has its own accepted-pixel denominator.

C External Tracking Context

Table 6: Published tracking values on the four TUM dynamic sequences, retained only as protocol-separated context. ATE RMSE is in cm (↓). Values are reported by DyPho-SLAM (Liu et al., 2025); execution protocols are not matched to Table 1, so no cross-source ranking is implied.

Method	fr3/w/xyz	fr3/w/half	fr3/w/static	fr3/s/half
SplaTAM	136.60	185.40	78.30	14.10
Photo-SLAM	88.30	33.50	9.40	3.20
DyPho-SLAM	1.60	2.60	0.60	1.60

D ADDITIONAL MAPPING VISUALIZATION

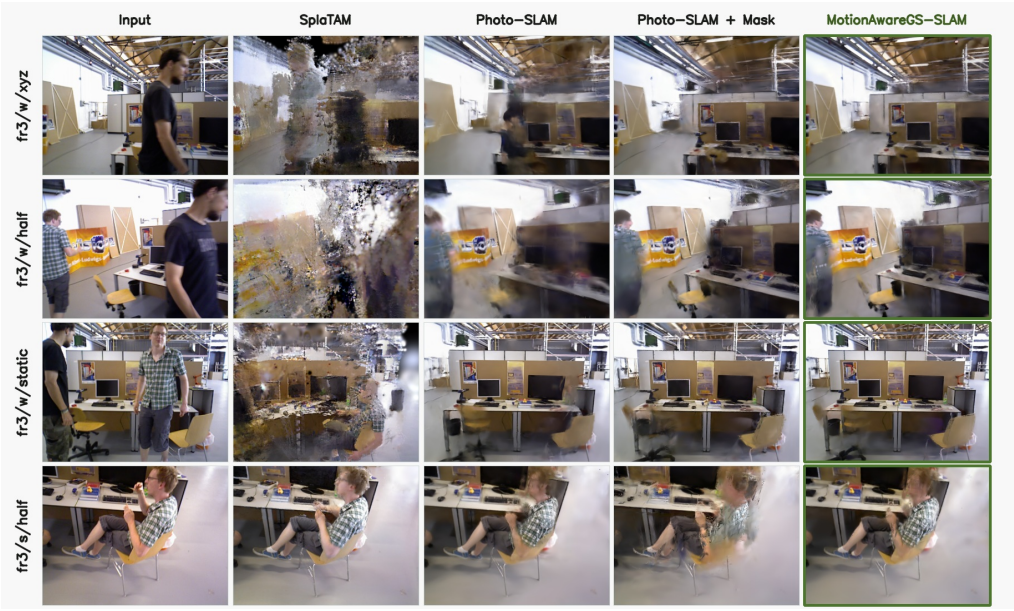


Figure 5: Mapping cross-method comparison: four TUM Freiburg3 dynamic sequences and five columns (input, SplaTAM, Photo-SLAM, Photo-SLAM+Mask, MotionAwareGS-SLAM). Frame selection uses the independent future-only RGB-D/official-pose reference and MotionAware renders, without baseline renders or scores. Each row uses the same physical timestamp. SplaTAM uses local 300-frame checkpoints; the other methods use completed full-sequence maps. The green border identifies our method. This is qualitative evidence, not a protocol-matched quantitative ranking.