STAG-VIO: Stabilized Prompt-to-Geometry Interface for Robust Dynamic Visual–Inertial Odometry

Rui Zhou[1] **Jingbin Liu**[2,†] **Junbin Xie**[1] **Jianyu Zhang**[1] **Yingze Hu**[1] **Jiele Zhao**[1]

[1]Electronic Information School, Wuhan University.
[2]College of Geodesy and Geomatics, Shandong University of Science and Technology

Project page: https://stag-vio.github.io/

[†]Corresponding author

Abstract

Dynamic visual–inertial odometry (VIO) requires reliable suppression of motion-corrupted measurements, yet prior semantic-assisted approaches depend on category-limited segmenters and degrade under partial occlusion. Promptable foundation segmentation models offer category-agnostic dynamic parsing, but their effectiveness in VIO depends critically on the temporal stability of input prompts—a factor largely overlooked in existing pipelines. When prompts derived from raw detection are jittery or intermittent under occlusion, the resulting masks flicker across frames, destabilizing geometric estimation. We propose STAG-VIO, which formulates dynamic robustness as a perception-to-geometry interface stabilization problem. We introduce uncertainty-adaptive multi-object tracking that models prompt generation as state estimation with bounded noise adaptation, producing temporally coherent box prompts. These stabilized prompts drive a lightweight foundation segmenter whose masks undergo geometry-oriented morphological refinement to establish conservative safety margins. A constraint-budget-aware feature redistribution strategy preserves well-conditioned static measurements when dynamic regions dominate the view. Experiments on VIODE and OpenLORIS-Scene show consistent gains over state-of-the-art baselines. Ablation confirms that prompt stabilization is the single most impactful component, reducing trajectory error by up to 83%.

1 INTRODUCTION

Visual–inertial odometry (VIO), which fuses synchronized camera streams and IMU measurements for ego-motion estimation, is a foundational capability for autonomous platforms operating in GPS-denied environments [1, 2, 3]. By exploiting complementary visual and inertial cues, modern VIO systems achieve accurate, drift-controlled pose estimation under a wide range of motions. However, real-world scenes are rarely static: pedestrians, vehicles, and other moving objects frequently dominate the camera's field of view, while partial occlusions and appearance variations further complicate visual perception. Under such highly dynamic conditions, classical static-scene assumptions are violated, leading to corrupted feature tracks, degraded geometric constraints, and—in extreme cases—complete tracking failure. A natural remedy is to incorporate semantic segmentation to suppress dynamic regions before geometric estimation [4, 5, 6, 7], yet this line of work faces a systematic bottleneck that has received insufficient attention.

Prior semantic-assisted VIO systems [4, 6, 5, 7], rely on segmentation networks trained on fixed label sets (*e.g.*, SegNet [8], Mask R-CNN [9]), which limits their ability to generalize to unseen dynamic object categories. Recent promptable foundation models such as SAM [10] and its efficient variants [11, 12, 13] seemingly resolve this limitation by enabling category-agnostic segmentation via spatial prompts. However, our ablation study (Section 4.3) reveals that naïvely feeding a foundation segmenter with raw detection prompts—without explicit temporal stabilization—increases trajectory error by up to 6×, with the error distribution exhibiting a pronounced long tail of catastrophic drift episodes (Fig. 4). The root cause is not the segmentation model itself, but the *temporal instability of its input prompts*: under partial occlusion and detector noise, bounding-box prompts become intermittent or spatially jittery across frames, which propagates through masks into inconsistent dynamic-region estimates and ultimately destabilizes geometric estimation. In other words, the effective bottleneck in foundation-model-assisted dynamic VIO lies at the *perception-to-geometry interface*—specifically, the temporal coherence of the information pathway from object detection to geometric constraints.

This observation leads to a clear design principle: to exploit universal segmentation for robust dynamic VIO, one must explicitly stabilize the entire prompt–mask–geometry information chain rather than loosely coupling off-the-shelf components. We therefore propose **STAG-VIO** (Stabilized Tracking-Assisted Generalized-masking VIO), which formulates dynamic robustness as an *interface stabilization* problem and decomposes it into three tightly coupled stages, each targeting a specific failure mode in the information pathway: **(i)** *Track-to-Prompt* models prompt generation as adaptive state estimation to produce temporally coherent bounding-box prompts under partial occlusion; **(ii)** *Prompt-to-Mask* converts stable prompts into conservative, geometry-oriented dynamic masks via foundation segmentation with asymmetric morphological refinement; and **(iii)** *Mask-to-Geometry* translates refined masks into reliable geometric constraints through budget-aware static feature redistribution that maintains estimator observability when dynamic regions dominate the view. Moreover, the temporal coherence of stabilized prompts enables a practical efficiency gain: the segmentation model's image encoder can be invoked at reduced frequency

while the lightweight mask decoder runs every frame, as the stable prompts compensate for slightly stale image embeddings (Section 3.3). Crucially, the framework is model-agnostic: the detector and segmentation backbone can each be replaced by alternative implementations, provided they satisfy the interface contracts of temporally stable prompts and geometry-reliable masks.

We validate STAG-VIO on two complementary benchmarks: VIODE [14] (outdoor, controlled dynamic levels) and OpenLORIS-Scene [15] (indoor, open-world dynamics with crowds and occlusions), demonstrating consistent accuracy and robustness gains over state-of-the-art VIO baselines in both settings. Crucially, ablation reveals that the proposed prompt stabilization is the single most impactful component, reducing trajectory error by up to 83%—strongly suggesting that temporal prompt coherence, rather than segmentation model capacity, is the dominant factor in robust dynamic VIO. Real-world deployment on a mobile robot in complex urban traffic further confirms the framework's practical effectiveness, with a 57.9% RMSE reduction compared to the baseline. Our contributions are summarized as follows:

- A perception-to-geometry interface framework is proposed that decomposes dynamic VIO robustness into a principled Track-to-Prompt → Prompt-to-Mask → Mask-to-Geometry pipeline, systematically stabilizing the detection-to-constraint pathway.
- We introduce an uncertainty-adaptive prompt stabilization module that models prompt generation as state estimation with bounded noise adaptation, enabling temporally coherent segmentation under occlusion and safe encoder embedding reuse for improved throughput.
- Extensive experiments across controlled, open-world, and real-world scenarios demonstrate consistent improvements over state-of-the-art baselines, with ablation revealing prompt coherence as the dominant accuracy factor in foundation-model-assisted dynamic VIO.

2 RELATED WORK

Handling dynamic objects in VIO and visual SLAM has been approached from three complementary directions: geometry-only filtering, semantic-assisted rejection, and joint perception–estimation frameworks.

Geometry-based methods identify and reject dynamic measurements without external semantic cues. RP-VIO [16] detects inconsistencies between visual reprojection and IMU prediction to flag outliers. RD-VIO [17] proposes an IMU-PARSAC algorithm that performs two-stage robust matching using accumulated motion statistics, handling both dynamic scenes and pure rotation. DynaVINS [6] formulates a novel loss function in bundle adjustment to downweight features whose motion is inconsistent with IMU preintegration, avoiding the need for explicit object detection. DFF-VIO [18] further introduces a dynamic feature fusion strategy that adaptively weights features based on motion consistency, enabling feature reuse rather than outright rejection. While effective in moderately dynamic scenes, these methods rely on motion consistency as their sole discriminative signal; when large portions of the field of view move coherently (e.g., a bus occupying 60% of the image), the consensus set itself is corrupted and estimation degrades.

Semantic-assisted methods leverage learned segmentation to identify known dynamic classes and mask their features before geometric estimation. DS-SLAM [4] augments ORB-SLAM2 with a SegNet [8]-based semantic thread to detect potentially dynamic objects, then applies geometric validation to confirm their motion. DynaSLAM [5] employs Mask R-CNN [9] for multi-class instance segmentation and combines it with multi-view geometry to handle both known and unknown moving objects. PVO [19] takes a joint approach, coupling panoptic segmentation with visual odometry so that segmentation benefits from geometric cues and vice versa. These methods achieve strong results on benchmarks, yet their segmentation networks are tied to fixed label sets and require retraining to accommodate new environments or object categories, limiting deployment generality.

Foundation-model-assisted methods have begun to adopt promptable segmentation models to overcome the category limitation. SamSLAM [20] integrates SAM with epipolar-based dynamic detection within ORB-SLAM3 to filter dynamic features without relying on fixed label sets. However, existing foundation-model-assisted approaches treat prompt generation as a preprocessing step and feed raw detection outputs directly to the segmenter, without considering how detection noise and partial occlusion cause prompts to jitter or vanish across frames. As we show in Section 4.3, this temporal prompt instability—rather than segmentation model capacity—is the dominant factor limiting trajectory accuracy, yet it has received no explicit treatment in the literature.

3 METHODOLOGY

We present STAG-VIO, a dynamic visual–inertial odometry framework that treats robustness in dynamic scenes as an explicit *perception-to-geometry interface* stabilization problem. Given a synchronized image stream $\{I_t\}$ and IMU measurements $\{u_t\}$, STAG-VIO estimates the platform state $\mathbf{x}_t$ (pose, velocity, and IMU biases) while explicitly stabilizing the information pathway from dynamic object detection to geometric constraints—the perception-to-geometry interface whose instability we identified in Section 1 as the dominant accuracy bottleneck.

3.1 Overview: Perception-to-Geometry Information Chain

STAG-VIO decomposes dynamic robustness into three coupled stages along the information chain $\mathcal{D}_t \to \hat{\mathcal{B}}_t \to \mathbf{M}_t^{\text{ref}} \to \mathcal{P}_t^{\text{static}} \to \mathbf{x}_t$, where raw detections $\mathcal{D}_t$ are converted into stabilized prompts $\hat{\mathcal{B}}_t$, then refined masks $\mathbf{M}_t^{\text{ref}}$, and finally a static feature set $\mathcal{P}_t^{\text{static}}$ for the back-end optimizer.

A critical observation is that noise introduced at any stage propagates and *amplifies* downstream: jittery detections produce flickering prompts, which yield temporally inconsistent masks, which in turn inject motion-corrupted features into geometric estimation. Our design inserts an explicit stabilization mechanism at each transition: **(i)** *Track-to-Prompt* converts noisy frame-wise detections into temporally coherent box prompts via uncertainty-adaptive multi-object tracking; **(ii)** *Prompt-to-Mask* obtains category-agnostic pixel masks from stable prompts and applies geometry-oriented morphological refinement to establish

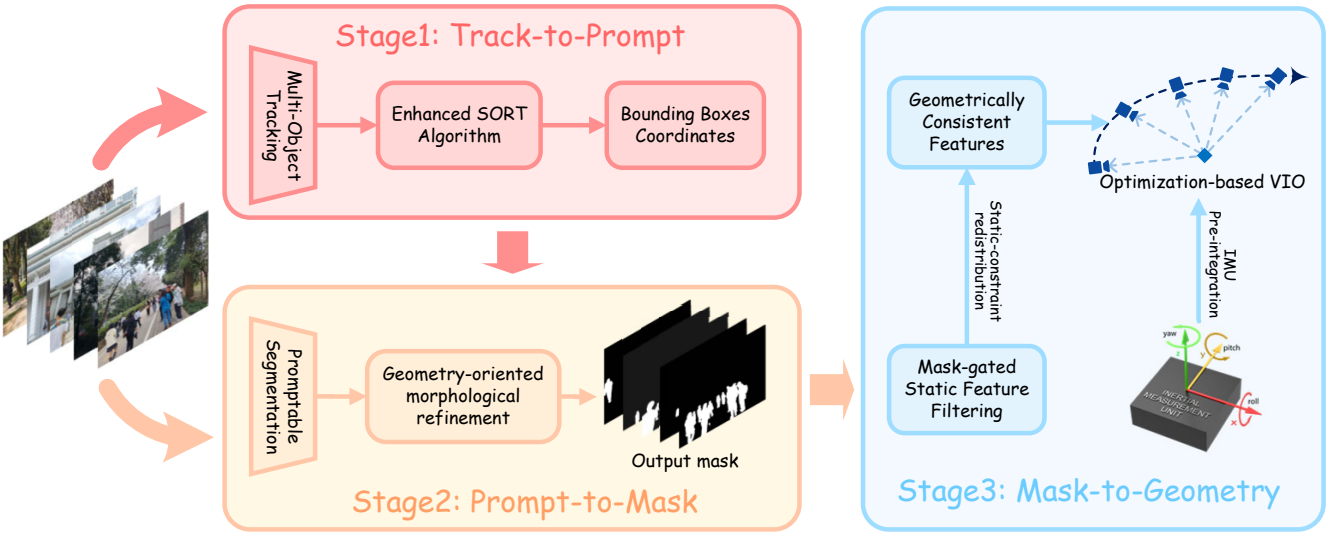


Figure 1: Overview of the STAG-VIO pipeline. The three stages address successive instabilities in the perception-to-geometry information chain: **Stage 1** stabilizes raw detections into temporally coherent box prompts via uncertainty-adaptive tracking; **Stage 2** converts prompts into conservative dynamic masks through promptable segmentation and geometry-oriented morphological refinement; **Stage 3** filters dynamic features and redistributes static constraints to maintain well-conditioned visual–inertial optimization.

conservative safety margins; **(iii)** *Mask-to-Geometry* translates refined masks into reliable geometric constraints through mask-gated feature filtering and a constraint-budget-aware redistribution strategy.

This decomposition explicitly separates open-world perception from metric estimation while enforcing a stable interface between them.

3.2 *Track-to-Prompt: Occlusion-Robust Prompt Stabilization via Uncertainty-Adaptive Tracking*

Let $\mathcal{B}_t = \{\mathbf{z}_t^k\}$ denote the set of candidate bounding boxes at time t produced by a real-time object detector. As discussed in Section 1, directly using $\mathcal{B}_t$ as prompts is brittle: detections become intermittent or jittery under occlusion, producing flickering masks that destabilize geometric estimation. However, the physical motion of tracked objects is inherently smooth, suggesting that prompt generation can be modeled as a *state estimation* problem where a predictive filter bridges detection gaps and suppresses jitter, yielding stabilized prompts $\hat{\mathcal{B}}_t$.

3.2.1 *State model*

For each tracked object k, we maintain a state vector

$$\mathbf{x}_t^k = \begin{bmatrix} x, & y, & w, & h, & v_x, & v_y, & v_w, & v_h \end{bmatrix}^\top, \tag{1}$$

where (x, y) is the box center, (w, h) are the box dimensions, and (v_x, v_y, v_w, v_h) are their respective temporal rates. A constant-velocity Kalman filter predicts $\mathbf{x}_{t|t-1}^k$ and updates it upon association with a matched detection.

3.2.2 *Data association*

We compute inter-frame correspondences between detections and predicted tracks by solving a global assignment problem via Hungarian matching [21] under a geometric IoU cost. This yields a consistent correspondence map that supports multi-object prompt continuity even when individual detections are momentarily missed.

3.2.3 *Uncertainty-adaptive prompt stabilization*

In practice, detection reliability varies significantly across frames. During partial occlusion, the detector may report a shrunken or shifted box; under severe occlusion, the detection may vanish entirely. If the Kalman filter's measurement noise covariance $\mathbf{R}$ is fixed, it either over-trusts unreliable detections (causing the prompt to jump) or under-trusts reliable ones (causing the prompt to lag). We therefore adaptively adjust $\mathbf{R}$ based on recent prediction residuals, implementing an asymmetric trust mechanism: *when detections are consistent with predictions, trust them; when they deviate substantially, downweight them.*

Specifically, for each measurement dimension j we compute the RMSE of the innovations within a sliding window of N frames:

$$r_{\text{RMSE}}^{(j)} = \sqrt{\frac{1}{N} \sum_{i=k-N+1}^{k} (v_i^{(j)})^2}, \tag{2}$$

where $\boldsymbol{\nu}_k = \mathbf{z}_k - \mathbf{H}\,\hat{\mathbf{x}}_{k|k-1}$ is the innovation (prediction residual) at time step k. The residual magnitude is then mapped to a measurement noise value through a bounded, monotonically non-decreasing activation function $\varphi : [0, +\infty) \to [R_{\min}, R_{\max}]$ with $R_{\min} > 0$:

$$\mathbf{R}_k = \beta \cdot \operatorname{diag}\Big(\varphi(\lambda \cdot r_{\text{RMSE}}^{(1)}),\ \ldots,\ \varphi(\lambda \cdot r_{\text{RMSE}}^{(m)})\Big), \tag{3}$$

where β controls the global noise magnitude and λ governs sensitivity to residual changes.

Design rationale for the bounds. The positive lower bound $R_{\min}$ prevents the Kalman gain from diverging when residuals are momentarily small—for instance, immediately after an

occlusion ends, the first matched detection may coincidentally align with the prediction, and an unbounded gain would cause the filter to over-commit to this single observation, amplifying prompt jitter on subsequent frames. Conversely, the upper bound $R_{\max}$ ensures that observations are never entirely discarded: even under severe, prolonged occlusion, the track retains a nonzero update from any matched detection, allowing prompt recovery once the target reappears.

Choice of activation function. We adopt $\varphi(x) = R_{\min} + (R_{\max} - R_{\min})\,\mathrm{erf}(x)$, where $\mathrm{erf}(\cdot)$ is the Gauss error function. Unlike arbitrary logistic curves (*e.g.*, sigmoid or tanh), the error function naturally aligns with the Gaussian noise assumption underlying the Kalman filter: it represents the cumulative probability of the measurement error magnitude exceeding nominal bounds, providing a principled, rather than *ad hoc*, mapping from residual statistics to noise covariance. In practice, the overall filtering behavior is primarily governed by β, λ, $R_{\min}$, and $R_{\max}$; replacing erf with other saturating functions yields comparable results, confirming that the bounded adaptation mechanism matters more than the specific functional form.

3.3 Prompt-to-Mask: Category-Agnostic Dynamic Parsing with Geometry-Oriented Refinement

Given stabilized prompts $\hat{\mathcal{B}}_t$, we obtain pixel-level dynamic masks using a promptable foundation segmentation model with box prompting [11, 22]. For each prompt $\hat{\mathbf{b}}_t^k$, the model predicts an instance mask $\mathbf{m}_t^k \in \{0,1\}^{H\times W}$. The overall dynamic mask is the union across all tracked instances:

$$\mathbf{M}_t = \bigvee_k \mathbf{m}_t^k . \tag{4}$$

3.3.1 Why recognition-optimal masks hurt geometry

Promptable segmentation models are trained to produce pixel-accurate, tightly fitting masks—optimal for recognition but problematic for geometry. Features near the contour of a moving object are often corrupted by motion blur, cast shadows, or minor boundary inaccuracies; if misclassified as static, they inject motion-corrupted measurements into the optimizer. Crucially, the cost is *strongly asymmetric*: a false negative (leaking a dynamic pixel) directly introduces a corrupted residual that can trigger catastrophic drift, whereas a false positive (over-masking a static pixel) merely reduces the constraint budget—a loss explicitly recoverable by redistribution (Section 3.4). This asymmetry motivates a deliberately conservative masking strategy.

3.3.2 Geometry-oriented morphological refinement

To operationalize the conservative masking principle, we refine the raw union mask $\mathbf{M}_t$ using an asymmetric morphological strategy. Instead of the conventional denoise-then-expand pipeline, we *first* apply an aggressive dilation using a large elliptical structuring element $\mathcal{S}_d$ to explicitly inflate the dynamic boundaries, establishing a safety buffer around every moving object:

$$\mathbf{M}_t^{\text{dilate}} = \mathbf{M}_t \oplus \mathcal{S}_d . \tag{5}$$

This is followed by a lightweight erosion with a much smaller element $\mathcal{S}_e$ to smooth out boundary artifacts and jagged edges introduced by the large-kernel dilation:

$$\mathbf{M}_t^{\text{ref}} = \mathbf{M}_t^{\text{dilate}} \ominus \mathcal{S}_e . \tag{6}$$

By design, $\mathcal{S}_d \gg \mathcal{S}_e$, so this "heavy-dilate, light-erode" operation produces masks that *over-envelope* the moving targets, providing a robust geometry-oriented buffer that actively prevents motion-corrupted edge features from leaking into the visual–inertial back-end. The static features lost to this conservative expansion are explicitly recovered by the constraint-preserving redistribution strategy in the next stage, ensuring that the overall constraint budget is maintained.

3.3.3 Embedding reuse via prompt-driven scheduling

The image encoder of the promptable segmentation model dominates per-frame computation, whereas the prompt-conditioned mask decoder is lightweight. Because the Track-to-Prompt stage produces temporally coherent prompts, the visual context relevant to dynamic objects evolves smoothly across adjacent frames, allowing the image embedding to be safely cached and reused. We therefore invoke the encoder only once every K frames; in intermediate frames, the mask decoder runs against the cached embedding with current stabilized prompts.

This decoupling relies on the mask decoder receiving two inputs: the image embedding and the box prompt. When the embedding is slightly stale, the up-to-date stabilized prompt still anchors the decoder to the correct region—without prompt stabilization, both inputs would be unreliable simultaneously, making reuse infeasible. The tolerable staleness depends on scene appearance: well-lit conditions permit longer reuse, whereas low-contrast scenes narrow the safe window. We empirically determine K in Section 4.4.

3.4 Mask-to-Geometry: Constraint-Preserving Feature Processing under Large Dynamic Regions

The refined dynamic mask $\mathbf{M}_t^{\text{ref}}$ is injected into the VIO front-end to gate visual measurements: tracked features propagated to frame t by sparse optical flow are discarded whenever they fall inside $\mathbf{M}_t^{\text{ref}}$, preventing motion-corrupted residuals from entering the estimator.

However, this masking introduces a *second-order failure mode*. When dynamic regions occupy a large portion of the image—a common situation in crowded urban or indoor scenes—naive masking decimates the tracked feature set. The surviving static features tend to cluster in narrow image bands, degrading the spatial conditioning of the visual–inertial optimization and, in extreme cases, rendering it ill-posed. Existing approaches either accept this constraint loss (risking estimator divergence) or relax the mask to re-admit potentially corrupted measurements. We instead adopt a third strategy: *actively managing the constraint budget* by replenishing and redistributing static features within the unmasked region.

3.4.1 Constraint-preserving redistribution

To sustain a robust geometric constraint budget, we replenish the tracked feature set to a target capacity $N_{\max}$ within the static region $\overline{\mathbf{M}_t^{\text{ref}}}$, excluding existing tracked points. Because the remaining static area is often narrow, newly detected features tend to cluster, degrading geometric conditioning. We therefore extract surplus ORB candidates [23] and apply Adaptive Non-Maximal Suppression (ANMS) [24], which selects keypoints by their minimum suppression radius $r_i = \min_{j:\, R_j > R_i} \|\mathbf{p}_i - \mathbf{p}_j\|_2$ to enforce approximately uniform spatial spread. A final dis-

tance filter $D_{\min}$ removes remaining close pairs, ensuring a well-conditioned static constraint set even when the dynamic mask covers the majority of the image.

3.5 Back-End Visual–Inertial Estimation

The back-end follows a standard optimization-based VIO formulation with IMU pre-integration and reprojection residuals, but it operates on the *mask-filtered and redistributed* static feature set:

$$\mathcal{P}_t^{\text{static}} = \{\mathbf{p} \in \mathcal{P}_t \mid \mathbf{M}_t^{\text{ref}}(\mathbf{p}) = 0\}. \tag{7}$$

By suppressing dynamic outliers at the *interface level* rather than relying on the optimizer's internal robustification, the back-end receives a stable stream of reliable geometric constraints even under severe occlusion and diverse moving objects.

Model agnosticism. Our framework imposes no assumptions on the specific detector or segmentation model, as long as they collectively produce temporally stable prompts and geometry-reliable masks. The detector can be replaced by any real-time proposal generator (*e.g.*, YOLO series [25, 26], RT-DETR series [27, 28]); and the segmenter by other promptable foundation models (*e.g.*, EfficientSAM [12], FastSAM [13], SAM 2 [29]). This modularity enables flexible trade-offs between accuracy and computational efficiency depending on deployment constraints.

4 EXPERIMENTS

4.1 Experimental Protocol

Datasets. We evaluate STAG-VIO in both outdoor and indoor dynamic settings.

(1) VIODE [14] provides controlled dynamic levels (*none / low / mid / high*) across urban environments, enabling systematic stress tests as moving objects increasingly dominate the field of view. We report results on three environments—*City day*, *City night*, and *Parking lot* (12 sequences in total)—which together cover the key challenges of high dynamic density, significant illumination variation, and diverse scene geometry.

(2) OpenLORIS-Scene [15] contains long-horizon indoor sequences across five diverse settings (*office, corridor, home, cafe, market*) with realistic challenges including lighting variations, crowds, and frequent partial occlusions.

Metrics. On VIODE, we report **RMSE of Absolute Trajectory Error (ATE)** in meters. On OpenLORIS-Scene, we additionally report **Correct Rate (CR)** [15] over the full time span, where CR reflects tracking robustness (higher is better) and ATE RMSE reflects accuracy (lower is better).

Implementation details. All experiments are conducted on an NVIDIA RTX 4090D GPU. We instantiate STAG-VIO with YOLOv11x [26] for object detection and MobileSAM [11] for prompt-guided segmentation. For the uncertainty-adaptive tracker, the sliding window size is $N = 10$, sensitivity $\lambda = 0.01$, and noise scale $\beta = 1.0$; the bounded activation function φ is instantiated with the Gauss error function as in Eq. (3). For geometry-oriented morphological refinement, the dilation kernel $\mathcal{S}_d$ is a 25×25 elliptical element and the erosion kernel $\mathcal{S}_e$ is 3×3. Feature parameters are set to $N_{\max} = 150$ and $D_{\min} = 15$ pixels unless otherwise stated. The image encoder is invoked every $K = 2$ frames by default .

4.2 Main Results on Public Benchmarks

4.2.1 VIODE: accuracy under controlled dynamics

Table 1 reports ATE RMSE across all four dynamic levels (none/low/mid/high) for the three VIODE environments. STAG-VIO achieves the best result in 8 of 12 settings and, more importantly, never fails and never exceeds 0.246 m ATE in any configuration—including the most adversarial high-dynamic sequences—whereas every baseline exhibits at least one catastrophic failure or divergence.

Graceful degradation vs. catastrophic collapse. The defining difference between STAG-VIO and the baselines is not average accuracy but the *shape* of degradation as dynamics intensify. Purely geometric or fixed-vocabulary methods stay competitive up to a threshold and then collapse abruptly: VINS-Mono is accurate on Parking lot at none/low (0.102/0.109 m) but jumps to 2.915 m at mid and 4.933 m at high—a 45× increase—while ORB-SLAM3 fails outright (∗) on City day–high and on all dynamic City night sequences. RD-VIO (0.187→1.221 m from mid to high on City day) and VINS-Fusion (0.161→1.201 m on Parking lot) show the same cliff behavior. In contrast, STAG-VIO degrades gracefully: across the full none→high sweep its error stays within 0.094–0.246 m in every environment, a spread of under 2.2×, indicating that the perception-to-geometry interface removes the dominant failure mode rather than merely lowering the mean.

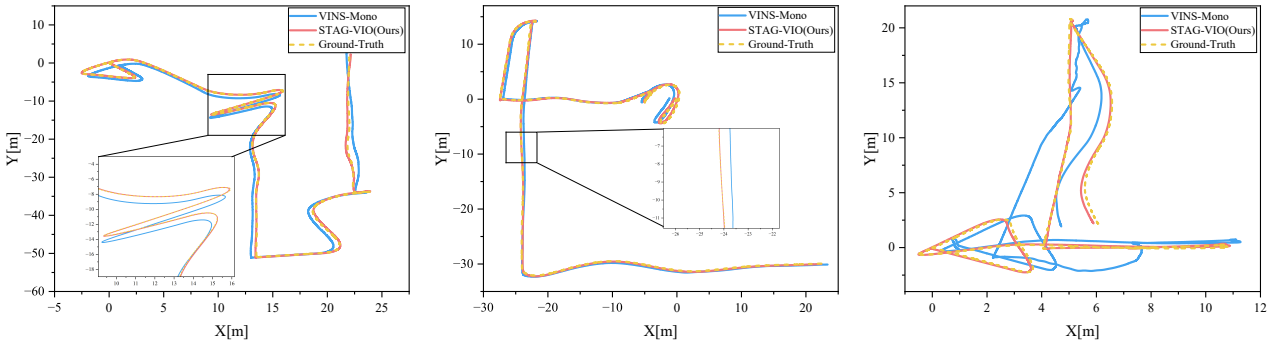
High-dynamic regime. The advantage is most pronounced under high dynamics, where moving objects dominate the field of view. Using VINS-Mono [2] as a consistent reference, STAG-VIO reduces ATE RMSE from 3.169 m to 0.204 m on City day–high (93.56%), from 0.575 m to 0.210 m on City night–high (63.48%), and from 4.933 m to 0.094 m on Parking lot–high (98.09%). These gains are not merely relative to a weak baseline: even against the *strongest* competing method in each high-dynamic column (Dyna-VINS in all three), STAG-VIO still improves accuracy by 16.7% (City day), 32.5% (City night), and 10.5% (Parking lot), confirming that the benefit stems from the stabilized interface rather than from an unfavorable baseline choice.

The low-dynamic regime. In the four settings where STAG-VIO does not rank first—City day–low, City night–low, Parking lot–low, and Parking lot–mid—the scenes are predominantly low-dynamic, where dynamic masking offers little benefit and the modest overhead of the segmentation pipeline can marginally affect estimation. Crucially, the absolute gaps are small in every case (≤0.051 m; e.g., 0.156 vs. 0.148 m on City day–low), and the methods that edge out STAG-VIO in these near-static conditions are precisely those that collapse once dynamics increase (VINS-Mono wins Parking lot–low at 0.109 m but diverges to 4.933 m at high). STAG-VIO therefore trades a negligible, bounded overhead in easy scenes for a decisive robustness advantage across the operating envelope—a favorable trade-off for deployment, where worst-case rather than best-case behavior governs safety. Figure 2 visualizes estimated trajectories on the high sequences, where STAG-VIO consistently aligns with the ground truth while VINS-Mono exhibits severe drift.

Table 1: Comparison with State-of-the-art Methods (RMSE of ATE in [M]). We highlight the best of each column in red. The STAG-VIO row is highlighted in light yellow, and the three *high* columns are highlighted in light blue.

Method	VIODE											
	City day				City night				Parking lot			
	none	low	mid	high	none	low	mid	high	none	low	mid	high
ORB-SLAM3 [1]	2.179	5.301	2.204	*	0.176	*	*	*	0.287	2.897	6.742	7.482
VINS-Fusion [3]	0.203	0.148	0.261	0.321	0.319	0.368	0.433	0.490	0.120	0.113	0.161	1.201
VINS-Mono [2]	0.186	0.237	0.263	3.169	0.314	0.436	0.727	0.575	0.102	0.109	2.915	4.933
VINS-Mah [30]	0.149	0.193	0.222	0.264	0.417	0.610	0.651	0.556	0.136	0.151	0.246	0.128
RP-VIO [16]	0.382	0.229	0.435	0.536	0.263	0.509	0.652	0.577	0.981	1.334	0.375	0.713
RD-VIO [17]	0.305	0.148	0.187	1.221	0.305	0.871	0.635	0.560	0.299	0.359	0.312	0.527
Dyna-VINS [6]	0.349	0.330	0.258	0.245	0.687	0.207	0.251	0.311	0.106	0.167	0.181	0.105
STAG-VIO (Ours)	**0.107**	**0.156**	**0.138**	**0.204**	**0.125**	**0.246**	**0.222**	**0.210**	**0.099**	**0.160**	**0.204**	**0.094**

• *: Failure case.



(a) City day (b) City night (c) Parking lot

Figure 2: Trajectory comparison on the high-dynamic VIODE sequences, including City day, City night, and Parking lot. The ground truth is shown as a dashed yellow line. STAG-VIO closely follows the ground-truth trajectories across all scenes, while VINS-Mono exhibits significant drift under strong dynamic interference.

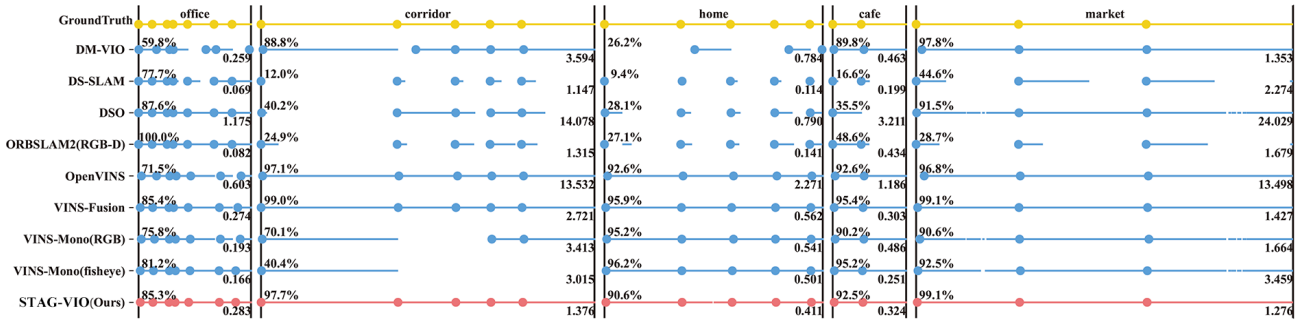


Figure 3: Experiment results with the OpenLORIS-Scene datasets. For every algorithm, the dots represent successful initiation instances, and the lines show the span of successful tracking. The percentage in the top left corner of each scene indicates the average correct rate; a higher correct rate suggests greater algorithm robustness. The float value on the first line below depicts the average ATE RMSE, with lower values indicating higher accuracy. Some experimental results from prior methods are referenced from [15]

4.2.2 *OpenLORIS-Scene: long-horizon robustness in open-world indoor scenes*

Figure 3 reports both Correct Rate (CR) and ATE RMSE across five indoor scenes. Unlike the controlled VIODE benchmark, OpenLORIS-Scene consists of long-horizon sequences recorded in everyday environments, where the dominant challenges are sustained tracking over extended durations, open-world object categories, and frequent partial occlusions rather than a single controllable dynamic level. We therefore read the two metrics jointly: CR measures how large a fraction of the trajectory is tracked without losing initialization (robustness, higher is better), while ATE RMSE measures accuracy over the tracked portion (lower is better).

Robustness under heavy dynamics. The two most demanding scenes are *market* and *corridor*. In *market*, dominated by dense crowds and diverse moving objects, STAG-VIO sustains 99.1% CR with 1.276 m ATE, tracking almost the entire sequence despite pervasive dynamic occlusion—the regime our

perception-to-geometry interface is designed for, and where fixed-vocabulary baselines tend to lose tracking as unseen or heavily occluded objects evade their segmenters. In *corridor*, characterized by low texture, illumination changes, and frequent occlusions, STAG-VIO maintains 97.7% CR with 1.376 m ATE, showing resilience to the compounded stress of weak visual constraints and dynamic interference.

Balanced performance across scenes. We note that the highest-CR scenes (*market*, *corridor*) also show larger absolute ATE than shorter, more static scenes such as *office*; this is expected rather than contradictory, as sustaining tracking through long, heavily dynamic trajectories accumulates more drift than terminating early would, and high CR paired with meter-level drift is the more deployment-relevant outcome. Across the remaining scenes (*office*, *home*, *cafe*), STAG-VIO preserves a favorable robustness–accuracy balance, sustaining high CR while keeping ATE competitive with state-of-the-art methods. Taken with the VIODE results, this consistency across controlled outdoor dynamics and uncontrolled open-world indoor conditions indicates that stabilizing the perception-to-geometry interface generalizes beyond any single benchmark.

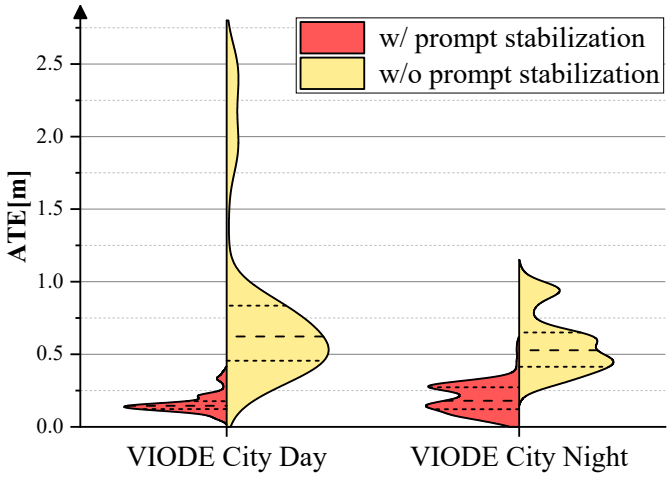


Figure 4: Violin plots of ATE on VIODE (City day/night) comparing STAG-VIO with and without the proposed prompt-stabilization tracker, showing that stabilized prompts yield lower and more consistent ATE.

4.3 Analysis of the Perception-to-Geometry Interface

4.3.1 Impact of prompt stabilization

To isolate the contribution of the Track-to-Prompt stage, we ablate the proposed prompt-stabilization tracker—feeding raw per-frame detections directly to the segmenter—and evaluate on VIODE. Figure 4 compares the per-frame ATE distributions with and without stabilization on City Day and City Night; full distributional statistics are reported in Appendix A.2.

Accuracy gain. Prompt stabilization reduces the RMSE of the ATE sequence by 83.28% on City Day and 63.99% on City Night. The improvement is equally evident in central tendency: the median ATE drops from 0.622 m to 0.145 m (76.7%) on City Day and from 0.528 m to 0.180 m (65.8%) on City Night, confirming that the gain reflects a systematic shift of the entire distribution rather than a change in a single summary statistic.

Elimination of catastrophic drift. Beyond the shift in central tendency, the shape of the distribution changes qualitatively—the property most visible in the violin plots. On City Day, the long upper tail collapses: the 95th-percentile ATE falls from 2.373 m to 0.326 m and the maximum from 2.608 m to 0.406 m, while the fraction of frames exceeding 1 m drops from 16.6% to 0%. Concentration tightens accordingly, with the interquartile range shrinking 7.0× (0.380 m → 0.054 m) and the standard deviation 9.5× (0.609 m → 0.064 m). City Night shows the same trend more moderately (IQR 1.6×, std 2.0× tighter), because its baseline distribution lacks the heavy day-time tail to begin with. This confirms that prompt stabilization does not merely lower average error but eliminates the catastrophic drift episodes that populate the upper tail—a property critical for deployment, where occasional large excursions are more dangerous than uniformly moderate error.

Taken together, these results establish prompt temporal coherence as the single largest contributor to STAG-VIO's accuracy, consistent with our central premise that stabilizing the perception-to-geometry interface—rather than increasing segmentation-model capacity—is the key to robust dynamic VIO.

4.3.2 Robustness to the static-constraint budget

A key practical failure mode in dynamic VIO is that aggressive dynamic masking leaves too few or poorly distributed static features. To stress-test constraint sensitivity, we densely sweep the maximum feature count $N_{\max} \in [40, 200]$ and the minimum feature distance $D_{\min} \in [5, 40]$ px, yielding 2,916 parameter configurations per method, and evaluate ATE RMSE on the VIODE high-dynamic sequence. Figure 5 visualizes the resulting error landscapes, and full statistics are reported in Appendix A.1.

Sensitivity of the baseline. VINS-Mono is highly sensitive to the constraint budget. Its RMSE spans 0.113–7.668 m across the grid—over an order of magnitude—with a median of 2.334 m and an interquartile range (IQR) of 4.127 m. Critically, 78.9% of all configurations exceed 1 m error and only 13.6% stay below 0.5 m, indicating that acceptable accuracy occupies a narrow region of the parameter space. The error correlates strongly and negatively with $D_{\min}$ ($\rho = -0.77$): without a mechanism to enforce spatial spread, the baseline depends critically on a well-chosen minimum distance to avoid feature clustering, and mis-setting it triggers sharp degradation.

Insensitivity of STAG-VIO. In contrast, STAG-VIO remains consistently accurate across the same space, with a median RMSE of 0.229 m (10.2× lower than VINS-Mono) and an IQR of just 0.165 m—a typical spread 25× tighter than the baseline. 95.2% of all configurations stay below 0.5 m and only 0.8% exceed 1 m, so nearly the entire parameter space is deployable rather than a narrow sweet spot. Its error is essentially uncorrelated with both $N_{\max}$ ($\rho = 0.11$) and $D_{\min}$ ($\rho = -0.06$), showing that the mask-aware redistribution strategy does not merely remove outliers but actively sustains a sufficient and well-distributed static constraint set regardless of how the budget is parameterized. We attribute this flatness to ANMS-based replenishment, which enforces spatial coverage internally and thereby removes the baseline's dependence on $D_{\min}$.

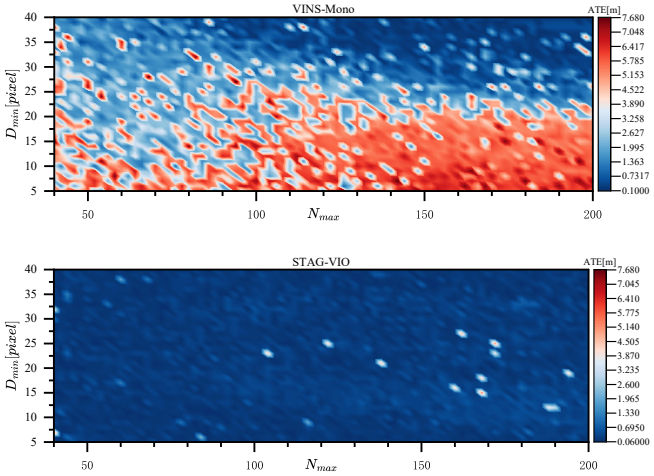


Figure 5: The heatmap illustrates RMSE ATE in relation to the maximum feature points N_{max} and minimum feature point distance D_{min} for the "high" sequence of the VIODE dataset. The color bar represents the range of ATE values.

Table 2: Effect of SAM encoder interval K on accuracy and throughput (VIODE *high* dynamic level, RTX 4090D).

K	ATE RMSE [m] ↓		FPS ↑	
	City day	City night	City day	City night
1 (base)	0.197	0.257	14.59	15.99
2	**0.204**	**0.210**	**18.22**	**19.55**
5	0.192	0.473	20.78	21.77
10	0.271	0.347	21.82	23.13

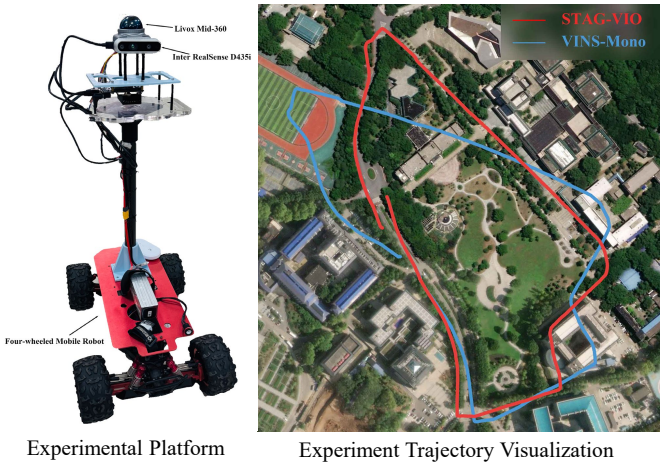
This robustness has direct practical value, substantially reducing the need for careful per-scenario parameter tuning. For completeness, a small fraction of STAG-VIO configurations (0.8%, concentrated at the largest $N_{\max}$) still produce elevated error, where over-replenishment re-admits weakly conditioned features; this residual sensitivity is far milder than the baseline's and is easily avoided by keeping $N_{\max}$ in a moderate range.

4.4 *Runtime and Efficiency Analysis*

Table 2 reports throughput and accuracy under different encoder intervals K on the most challenging VIODE sequences (*high* dynamic level). At $K = 1$, the full pipeline already operates at ∼15 FPS, exceeding the 10 Hz camera rate of both evaluation datasets. Setting $K = 2$ improves throughput to ∼19 FPS—a **23% speedup with no accuracy loss** (average ATE shifts from 0.227 m to 0.207 m), validating the design in Section 3.3: temporally coherent prompts make embedding reuse not only safe but effectively free in terms of accuracy cost. For $K \geq 5$, throughput gains saturate (21–23 FPS) while nighttime accuracy degrades notably (ATE: 0.210 → 0.473 m at $K = 5$), confirming that low-contrast scenes narrow the safe reuse window. We adopt $K = 2$ as the default, achieving a favorable accuracy–efficiency balance that demonstrates prompt stabilization as both an accuracy enabler and a practical efficiency mechanism.

4.5 *Real-World Deployment*

Beyond public benchmarks, we evaluate STAG-VIO on a real outdoor sequence to verify that the proposed perception-to-geometry interface transfers to uncontrolled deployment conditions.



Experimental Platform — Experiment Trajectory Visualization

Figure 6: (a) The sensor suite mounted on a four-wheeled mobile robot, consisting of a Livox Mid-360 LiDAR and an Intel RealSense D435i stereo camera. (b) Trajectory estimates from STAG-VIO (red) and VINS-Mono (blue) on a large-scale outdoor sequence. STAG-VIO produces a more consistent trajectory that closely follows the actual path, while VINS-Mono exhibits noticeable drift.

Table 3: Real-world experiment results.

Method	ATE [m] ↓			
	RMSE	Mean	Max	Std
VINS-Mono [2]	34.609	29.266	75.215	18.475
STAG-VIO (Ours)	**14.563**	**12.656**	**31.736**	**7.205**
Improvement	↓57.9%	↓56.8%	↓57.8%	↓61.0%

Platform and ground truth. Data is collected with a four-wheeled mobile robot equipped with an Intel RealSense D435i camera (1280×720 at 30 Hz, BMI085 IMU at 200 Hz) and a Livox Mid-360 LiDAR. Camera–IMU extrinsics are calibrated with Kalibr [31]. Ground-truth poses are obtained by running FAST-LIO2 [32] on the LiDAR data, providing a reliable reference trajectory for evaluation.

Scenario. The sequence traverses complex urban traffic containing a diverse mix of dynamic objects—pedestrians, cars, buses, motorcycles, and tricycles—presenting the combined challenges of object diversity, varying scales, and frequent mutual occlusions.

Quantitative results. Table 3 reports trajectory accuracy. STAG-VIO reduces RMSE ATE from 34.61 m to 14.56 m compared with the VINS-Mono baseline, a **57.9%** error reduction. The improvement is consistent across all statistics: the maximum error drops by 57.8% (from 75.21 m to 31.74 m) and the standard deviation decreases by 61.0% (from 18.47 m to 7.20 m), indicating that STAG-VIO suppresses both average drift and large occasional excursions.

Qualitative results. Figure 6 aligns estimated trajectories with satellite imagery. Compared with VINS-Mono, which exhibits noticeable drift in segments with dense moving traffic, STAG-VIO produces a visibly more consistent path that better

conforms to the road geometry, confirming that the proposed framework transfers effectively to real-world deployments with complex dynamics and partial occlusions.

5 CONCLUSION

We have shown that dynamic robustness in visual–inertial odometry is fundamentally an interface stabilization problem. STAG-VIO instantiates this principle through a Track-to-Prompt → Prompt-to-Mask → Mask-to-Geometry pipeline, achieving consistent improvements on both VIODE and OpenLORIS-Scene benchmarks, with ablation confirming prompt stabilization as the single most impactful component (up to **83%** error reduction) and real-world deployment validating practical effectiveness (**57.9%** RMSE reduction in complex urban traffic). Current limitations include the category-bounded upstream detector, resolution-dependent morphological parameters, and ∼19 FPS throughput that leaves limited headroom for higher-frequency sensors; integrating open-vocabulary detectors [33], adaptive morphological kernels, and TensorRT deployment are promising directions.

ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China under Grant No. 42474060, the National Key R&D Program of China under Grant No. 2023YFB3906101, and the Natural Science Foundation of Hubei Province under Grant No. 2024AFD403.

We would like to express our sincere gratitude to the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University, for their support with the experimental platform and facilities.

REFERENCES

[1] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. M. Montiel, and J. 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. ISSN 1941-0468. doi: 10.1109/TRO.2021.3075644.

[2] T. Qin, P. Li, and S. Shen. Vins-mono: A robust and versatile monocular visual-inertial state estimator. *IEEE Transactions on Robotics*, 34(4):1004–1020, 2018. ISSN 1941-0468. doi: 10.1109/TRO.2018.2853729.

[3] Tong Qin, Jie Pan, Shaozu Cao, and Shaojie Shen. A general optimization-based framework for local odometry estimation with multiple sensors, January 01, 2019 2019. URL https://ui.adsabs.harvard.edu/abs/2019arXiv190103638Q.

[4] Chao Yu, Zuxin Liu, Xinjun Liu, Fugui Xie, Yi Yang, Qi Wei, and Qiao Fei. Ds-slam: A semantic visual slam towards dynamic environments, 2018. URL https://arxiv.org/abs/1809.08379.

[5] Berta Bescos, Jose M. Facil, Javier Civera, and José Neira. Dynaslam: Tracking, mapping and inpainting in dynamic scenes, 2018. URL https://arxiv.org/abs/1806.05620.

[6] Seungwon Song, Hyungtae Lim, Alex Junho Lee, and Hyun Myung. Dynavins: A visual-inertial slam for dynamic environments. *IEEE Robotics and Automation Letters*, 7(4):11523–11530, 2022. doi: 10.1109/LRA.2022.3203231.

[7] Jianheng Liu, Xuanfu Li, Yueqian Liu, and Haoyao Chen. Rgb-d inertial odometry for a resource-restricted robot in dynamic environments. *Ieee Robotics and Automation Letters*, 7(4):9573–9580, 2022. ISSN 2377-3766. doi: 10.1109/lra.2022.3191193. URL <GotoISI>://WOS:000831182500036. Times Cited: 24 Liu, Jianheng/0000-0002-3124-5559; Liu, Yueqian/0000-0002-1994-6408 0 24.

[8] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation, November 01, 2015 2015. URL https://ui.adsabs.harvard.edu/abs/2015arXiv151100561B.

[9] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn, March 01, 2017 2017. URL https://ui.adsabs.harvard.edu/abs/2017arXiv170306870H. open source; appendix on more results.

[10] A. Kirillov, E. Mintun, N. Ravi, H. Z. Mao, C. Rolland, L. Gustafson, T. T. Xiao, S. Whitehead, A. C. Berg, W. Y. Lo, P. Dolla'r, R. Girshick, and Ieee. Segment anything. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE International Conference on Computer Vision, pages 3992–4003, 2023. ISBN 979-8-3503-0718-4. doi: 10.1109/iccv51070.2023.00371. URL <GotoISI>://WOS:001159644304024. Kirillov, Alexander Mintun, Eric Ravi, Nikhila Mao, Hanzi Rolland, Chloe Gustafson, Laura Xiao, Tete Whitehead, Spencer Berg, Alexander C. Lo, Wan-Yen Dolla'r, Piotr Girshick, Ross 1550-5499.

[11] Chaoning Zhang, Dongshen Han, Yu Qiao, Jung Uk Kim, Sung-Ho Bae, Seungkyu Lee, and Choong Seon Hong. Faster segment anything: Towards lightweight sam for mobile applications, June 01, 2023 2023. URL https://ui.adsabs.harvard.edu/abs/2023arXiv230614289Z. First work to make SAM lightweight for mobile applications.

[12] Yu Xiong, Bohao Zhang, Rui Huang, Xiaodan Liang, Dahua Lin, et al. Efficientsam: Leveraged masked image pretraining for efficient segment anything, 2023. URL https://arxiv.org/abs/2312.00863.

[13] Xinyu Zhao, Wenguan Wang, et al. Fast segment anything, 2023. URL https://arxiv.org/abs/2306.12156.

[14] K. Minoda, F. Schilling, V. Wüest, D. Floreano, and T. Yairi. Viode: A simulated dataset to address the challenges of visual-inertial odometry in dynamic environments. *IEEE Robotics and Automation Letters*, 6(2):1343–1350, 2021. ISSN 2377-3766. doi: 10.1109/LRA.2021.3058073.

[15] Xuesong Shi, Dongjiang Li, Pengpeng Zhao, Qinbin Tian, Yuxin Tian, Qiwei Long, Chunhao Zhu, Jingwei Song, Fei Qiao, Le Song, Yangquan Guo, Zhigang Wang, Yimin Zhang, Baoxing Qin, Wei Yang, Fangshi Wang,

Rosa H. M. Chan, and Qi She. Are we ready for service robots? the openloris-scene datasets for lifelong slam. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3139–3145, 2020. doi: 10.1109/ICRA40945.2020.9196638.

[16] K. Ram, C. Kharyal, S. S. Harithas, K. M. Krishna, and Ieee. Rp-vio: Robust plane-based visual-inertial odometry for dynamic environments. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE International Conference on Intelligent Robots and Systems, pages 9198–9205, 2021. ISBN 978-1-6654-1714-3. doi: 10.1109/iros51168.2021.9636522. URL <GotoISI>://WOS:000755125507029. Ram, Karnik Kharyal, Chaitanya Harithas, Sudarshan S. Krishna, K. Madhava Krishna, Madhava/JYO-4800-2024 2153-0858.

[17] Jinyu Li, Xiaokun Pan, Gan Huang, Ziyang Zhang, Nan Wang, Hujun Bao, and Guofeng Zhang. Rd-vio: Robust visual-inertial odometry for mobile augmented reality in dynamic environments, 2023. URL https://arxiv.org/abs/2310.15072.

[18] Nan Luo, Zhexuan Hu, Yuan Ding, Jiaxu Li, Hui Zhao, Gang Liu, and Quan Wang. Dff-vio: A general dynamic feature fused monocular visual-inertial odometry. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(2):1758–1773, 2025. doi: 10.1109/TCSVT.2024.3482573.

[19] Weicai Ye, Xinyue Lan, Shuo Chen, Yuhang Ming, Xingyuan Yu, Hujun Bao, Zhaopeng Cui, and Guofeng Zhang. Pvo: Panoptic visual odometry. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9579–9589, 2023. doi: 10.1109/CVPR52729.2023.00924.

[20] Xianhao Chen, Tengyue Wang, Haonan Mai, and Liangjing Yang. Samslam: A visual slam based on segment anything model for dynamic environment. In *2024 8th International Conference on Robotics, Control and Automation (ICRCA)*, pages 91–97, 2024. doi: 10.1109/ICRCA60878.2024.10649185.

[21] Fatemeh Rajabi-Alni and Alireza Bagheri. Computing a many-to-many matching with demands and capacities between two sets using the hungarian algorithm. *Journal of mathematics*, 2023(1):7761902, 2023.

[22] Chaoning Zhang, Kang Han, Chang Xu, et al. Mobilesamv2: Faster segment anything to everything, 2023. URL https://arxiv.org/abs/2312.09579.

[23] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *2011 International Conference on Computer Vision*, pages 2564–2571, 2011. doi: 10.1109/ICCV.2011.6126544.

[24] M. Brown, R. Szeliski, and S. Winder. Multi-image matching using multi-scale oriented patches. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 510–517 vol. 1, 2005. doi: 10.1109/CVPR.2005.235.

[25] Muhammad Hussain. Yolov5, yolov8 and yolov10: The go-to detectors for real-time vision. *arXiv preprint arXiv:2407.02988*, 2024.

[26] Rahima Khanam and Muhammad Hussain. YOLOv11: An Overview of the Key Architectural Enhancements. *arXiv e-prints*, art. arXiv:2410.17725, October 2024. doi: 10.48550/arXiv.2410.17725.

[27] Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, and Jie Chen. Detrs beat yolos on real-time object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16965–16974, 2024.

[28] Wenyu Lv, Yian Zhao, Qinyao Chang, Kui Huang, Guanzhong Wang, and Yi Liu. Rt-detrv2: Improved baseline with bag-of-freebies for real-time detection transformer. *arXiv preprint arXiv:2407.17140*, 2024.

[29] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

[30] Yuquan Hu and Alessandro Gardi. Vins-mah: A robust monocular visual-inertial state estimator for dynamic environments. *IEEE Robotics and Automation Letters*, 11(3): 3614–3621, 2026. doi: 10.1109/LRA.2026.3656724.

[31] Paul Furgale, Joern Rehder, and Roland Siegwart. Unified temporal and spatial calibration for multi-sensor systems. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1280–1286, 2013. doi: 10.1109/IROS.2013.6696514.

[32] Wei Xu, Yixi Cai, Dongjiao He, Jiarong Lin, and Fu Zhang. Fast-lio2: Fast direct lidar-inertial odometry. *IEEE Transactions on Robotics*, 38(4):2053–2073, 2022. doi: 10.1109/TRO.2022.3141876.

[33] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pages 38–55. Springer, 2024.

Table 4: ATE RMSE distribution across the full ($N_{\max}, D_{\min}$) parameter sweep (2,916 configurations, VIODE high-dynamic sequence). All ATE values in meters. Best value in each row in bold.

Statistic	VINS-Mono	STAG-VIO (Ours)
Median	2.334	**0.229**
Mean	3.069	**0.274**
Std. deviation	2.157	**0.303**
IQR ($Q_3 - Q_1$)	4.127	**0.165**
Minimum	0.113	**0.069**
Maximum	7.668	**4.943**
Configs. < 0.5 m	13.6%	**95.2%**
Configs. > 1.0 m	78.9%	**0.8%**
Corr. with $N_{\max}$ (ρ/R^2)	0.04 / 0.1%	0.11 / 1.2%
Corr. with $D_{\min}$ (ρ/R^2)	−0.77 / 58.7%	−0.06 / **0.3%**

Table 5: Mean ATE RMSE within $N_{\max}$ bands (averaged over all $D_{\min}$ values), VIODE high-dynamic sequence. All values in meters. Best value in each column in bold.

$N_{\max}$ band	[40, 80)	[80, 120)	[120, 160)	[160, 200)
VINS-Mono	2.874	3.122	3.149	3.128
STAG-VIO (Ours)	**0.235**	**0.251**	**0.268**	**0.339**

A Detailed Experimental Statistics

This appendix reports the full distributional statistics underlying two analyses in the main text: the static-constraint-budget robustness sweep (Section 4.3.2, Figure 5) and the prompt-stabilization ablation (Section 4.3, Figure 4). All statistics are computed directly from the raw per-configuration and per-frame ATE values.

A.1 Static-Constraint-Budget Robustness Sweep

We sweep the maximum feature count $N_{\max} \in [40, 200]$ (step 2) and the minimum feature distance $D_{\min} \in [5, 40]$ px (step 1), giving 2,916 configurations per method, and evaluate ATE RMSE on the VIODE high-dynamic sequence. Table 4 reports the distribution of ATE RMSE over the full grid; "Configs. < 0.5 m" and "> 1.0 m" are the fractions of configurations falling in each range. The correlation rows give the Pearson correlation ρ between ATE RMSE and each swept parameter, with the coefficient of determination R^2 (the fraction of error variance explained by that parameter). Table 5 disaggregates the mean ATE RMSE into $N_{\max}$ bands, averaged over all $D_{\min}$ values.

A.2 Prompt-Stabilization Ablation

We compare STAG-VIO with the prompt-stabilization tracker enabled ("w/") against a variant that feeds raw per-frame detections directly to the segmenter ("w/o"), on the VIODE City Day and City Night sequences. Table 6 reports the full statistics of the per-frame ATE distributions visualized in Figure 4: central-tendency measures (RMSE, mean, median), spread and tail statistics (standard deviation, IQR, 95th percentile, maximum, and the fraction of frames exceeding 1 m), and the number of tracked frames n over which each distribution is computed.

Table 6: Per-frame ATE distribution statistics for the prompt-stabilization ablation on VIODE (data underlying Figure 4). "w/o" feeds raw detections to the segmenter; "w/" enables the proposed stabilization tracker. All ATE values in meters. Best value of each pair in bold.

	City Day		City Night	
Statistic	w/o	w/	w/o	w/
RMSE	0.868	**0.145**	0.531	**0.191**
Mean	0.824	**0.159**	0.569	**0.194**
Median	0.622	**0.145**	0.528	**0.180**
Std. deviation	0.609	**0.064**	0.199	**0.097**
IQR ($Q_3 - Q_1$)	0.380	**0.054**	0.235	**0.151**
95th percentile	2.373	**0.326**	0.954	**0.328**
Maximum	2.608	**0.406**	1.062	**0.583**
Frames > 1 m	16.6%	**0.0%**	1.1%	**0.0%**
Frames n	614	616	568	561

The 83.28% (City Day) and 63.99% (City Night) reductions quoted in the main text correspond to the RMSE row.[1]

[1] These per-frame ATE statistics are computed on the raw estimated trajectories and therefore differ slightly in convention from the aligned ATE RMSE reported in the main comparison table (Table 1), which follows the standard Umeyama-aligned evaluation.