Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.079663

ARTICLE

CF2-SLAM: Conformal-Calibrated Foundation-Factor Graph SLAM across Modalities and Domains

Xiangqin Chen*

College of Engineering, Pennsylvania State University, University Park, PA, USA

*Corresponding Author: Xiangqin Chen. Email: rexc@alumni.psu.edu

Received: 26 January 2026; Accepted: 10 April 2026; Published: 15 June 2026

ABSTRACT: Simultaneous localization and mapping (SLAM) must remain reliable when sensing suites and operating conditions vary across platforms and deployments. Beyond correspondence degradation, a dominant deployment failure mode is *misweighted* constraints: under distribution shift, uncertainty estimates can become miscalibrated, allowing a small set of overconfident factors to dominate iterative optimization and destabilize inference. This article presents conformal-calibrated foundation-factor graph SLAM (CF^2-*SLAM*), a sensor-agnostic framework that combines frozen foundation representations with lightweight probabilistic factor heads that emit explicit residuals and covariances, and a classical factor-graph back-end for principled multi-modal fusion. To mitigate systematic misweighting under shift, an online conformal calibration layer is introduced to rescale factor covariances by aligning empirical residual quantiles with target quantiles on a per-factor-family basis. Loop closure is further integrated through foundation-descriptor retrieval for candidate proposal and conservative geometric verification for graph insertion, controlling false loop constraints without relying on dataset-specific place-recognition supervision. Across heterogeneous benchmarks spanning monocular, stereo, red-green-blue-depth (RGB-D), and visual-inertial settings, CF^2-*SLAM* operates without retraining and shows improved robustness trends under zero-shot transfer, consistent with stabilized factor weighting.

KEYWORDS: Simultaneous localization and mapping (SLAM); factor graph optimization; foundation models; uncertainty estimation; online calibration; conformal calibration

1 Introduction

Simultaneous localization and mapping (SLAM) is a core state-estimation module for embodied platforms ranging from aerial robots to autonomous vehicles and augmented reality/virtual reality (AR/VR) devices. In practice, deployments vary in sensing configuration (monocular/stereo/red-green-blue-depth (RGB-D)/inertial measurement unit (IMU)) and operating conditions (illumination, weather, scene layout, motion, and dynamics), inducing distribution shifts that degrade robustness.

Classical geometric pipelines remain attractive due to transparent objectives and auditable components, but they rely on brittle data association and typically assume stationary noise models. Learned SLAM mitigates perceptual brittleness via learned representations and stronger visual priors [1,2], yet a less visible failure mode often dominates in deployment: *miscalibrated confidence under shift*. In iterative back-ends (bundle adjustment or factor graphs), relative factor weighting controls conditioning and convergence. A small subset of overconfident, incorrect constraints can dominate the normal equations and cause divergence or persistent bias, especially when mixing factor types of different dimensions and noise profiles.

The target of this work is a unified SLAM framework that operates across modalities while retaining the interpretability of a probabilistic back-end. Two principles guide the design. First, frozen foundation models can provide more transferable representations than task-specific encoders [3]. Second, if learned components emit explicit residuals and covariances, inference can be posed as a classical factor-graph maximum a posteriori (MAP) problem [4], enabling principled fusion. However, transferability alone does not prevent systematic misweighting under shift. An online conformal-style calibration mechanism is therefore introduced to adjust factor covariance magnitudes using residual quantiles [5], aiming to support more reliable optimization behavior over time under distribution shift.

Because widely used SLAM datasets differ in available sensor fields and calibration metadata, Section 5 documents modality availability and Section 6.2 fixes evaluation protocols to reduce inadvertent modality leakage.

Contributions.

This article makes three contributions: (1) CF^2-*SLAM* is formulated as a factor-graph SLAM framework in which frozen foundation features feed lightweight probabilistic factor heads, producing explicit residuals and covariances in a transparent MAP objective; (2) an online conformal calibration layer rescales factor covariances by matching observed residual quantiles to target quantiles, mitigating systematic misweighting under domain shift during sequential optimization; (3) descriptor-topological loop closure, i.e., descriptor-space candidate proposal followed by geometric verification and loop-closure (LC)-factor insertion, uses foundation descriptors for candidate proposal and geometric verification for conservative graph insertion, avoiding dataset-specific place-recognition supervision while controlling false loop closures.

2 Related Work

Classical geometric SLAM and visual-inertial odometry (VIO) remain strong baselines when sensing assumptions are matched to deployment. Feature-based systems such as ORB-SLAM3 [6] remain highly competitive, while recent geometry-aware learned systems such as Photo-SLAM [1] and IBD-SLAM [2] illustrate the continued value of combining stronger visual representations with explicit optimization back-ends. However, their accuracy and stability still depend strongly on data association quality and on appropriately tuned noise models across visual and inertial factors.

Learned front-ends improve correspondence robustness under viewpoint, illumination, and texture variation. Representative recent examples include DINOv2 [3], LightGlue [7], RoMa [8], and IBD-SLAM [2]. These methods show that learned representations and learned residual models can substantially strengthen SLAM front-ends, but they also expose a recurring weakness: confidence or covariance estimates that are reliable in-domain can become miscalibrated under cross-dataset, cross-sensor, or synthetic-to-real transfer.

Loop closure is commonly structured as candidate retrieval followed by geometric validation before graph insertion. In this setting, the proposed descriptor-topological module follows the same principle: foundation descriptors are used only for candidate proposal, while accepted loop constraints are instantiated as verified LC factors after dense geometric checking. This is distinct from metric-semantic SLAM in the sense of explicit object- or scene-level labeling, and it is also distinct from recent dense mapping systems such as Gaussian Splatting SLAM [9] and SplaTAM [10], which emphasize reconstruction quality and often assume RGB-D inputs and higher compute budgets.

Recent work on post-hoc neural calibration and conformal prediction has strengthened uncertainty quantification in supervised learning [5,11,12], but their role in SLAM is less straightforward because SLAM involves sequential correlation, heterogeneous factor dimensions, and solver-level sensitivity to relative factor weighting. The emphasis here is therefore not only uncertainty reporting but solver conditioning under

distribution shift. A related perspective on resilience under changing operating conditions appears in graph-structured logistics routing, where learned spatiotemporal risk prediction is integrated with dynamic edge weighting to maintain robust decision making under congestion and demand fluctuations [13]. The proposed online conformal layer complements standard robust losses: robust losses suppress large instantaneous outliers, whereas conformal rescaling corrects systematic covariance scale mismatch across factor families so that no single miscalibrated modality dominates the normal equations after transfer.

3 Problem Formulation

This work considers heterogeneous sensor streams (monocular/stereo/RGB-D/IMU) and estimates a trajectory (and optionally map parameters) over a horizon. Let the state at time t be

$$\mathbf{x}_t = \{\mathbf{T}_t \in \mathrm{SE}(3),\ \mathbf{v}_t \in \mathbb{R}^3,\ \mathbf{b}_t \in \mathbb{R}^6,\ \boldsymbol{\theta}_t\}, \tag{1}$$

where $\mathbf{T}_t$ is the pose, $\mathbf{v}_t$ velocity, $\mathbf{b}_t$ IMU biases (if applicable), and $\boldsymbol{\theta}_t$ optional map parameters.

A factor graph over $\mathbf{x} = \{\mathbf{x}_t\}_{t=1}^{T}$ defines the MAP objective:

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \sum_k \rho_k\left(\mathbf{r}_k(\mathbf{x})^\top \mathbf{W}_k^{-1} \mathbf{r}_k(\mathbf{x})\right), \tag{2}$$

where factor k contributes residual $\mathbf{r}_k(\mathbf{x})$, covariance matrix $\mathbf{W}_k$, and robust loss $\rho_k(\cdot)$ [4]. The standard convention is followed that $\mathbf{W}_k$ is a covariance matrix; correspondingly, $\mathbf{W}_k^{-1}$ is the information (weight) matrix used inside Eq. (2). Beyond uncertainty reporting, $\mathbf{W}_k$ directly controls factor influence and conditioning; thus systematic miscalibration under shift can destabilize optimization. In CF^2-*SLAM*, each learned factor family predicts an initial $\mathbf{W}_k$, and the online conformal calibration layer rescales $\mathbf{W}_k$ (hence $\mathbf{W}_k^{-1}$) to improve weighting reliability across modalities and domains. The method below retains interpretability of Eq. (2) while improving weighting reliability across modalities and domains. Throughout the paper, foundation-feature (FF), depth/disparity (D), inertial measurement unit (IMU), and loop-closure (LC) denote the four factor families used in the MAP objective. FF/D/LC correspond to visual or verified loop constraints whose residual blocks are paired with learned or learned-assisted covariance estimates, while IMU denotes the standard analytic preintegration factor with its usual inertial covariance when IMU measurements are available.

4 Method: CF^2-*SLAM*

4.1 Overview

CF^2-*SLAM* pairs a learned front-end with a classical factor-graph back-end. A frozen foundation encoder produces transferable representations; lightweight factor heads emit probabilistic constraints (residuals and covariances); an online calibration layer rescales covariances per factor type; inference uses Gauss-Newton/Levenberg-Marquardt (GN/LM) on a sliding-window graph. Modality changes do not require retraining: analytic factors (e.g., IMU preintegration) are enabled only when the corresponding sensor fields exist. An overview is shown in Fig. 1.

4.2 Foundation Representations

A frozen foundation model $\phi(\cdot)$ (e.g., DINOv2 [3]) is used to compute (i) dense token features $\mathbf{F}_t \in \mathbb{R}^{H\times W\times C}$ and (ii) a global descriptor $\mathbf{g}_t \in \mathbb{R}^C$ for retrieval. Token-wise normalization improves stability for similarity-based residuals. Freezing ϕ limits dataset-specific overfitting and focuses trainable capacity on small heads that translate features into residual and covariance predictions.

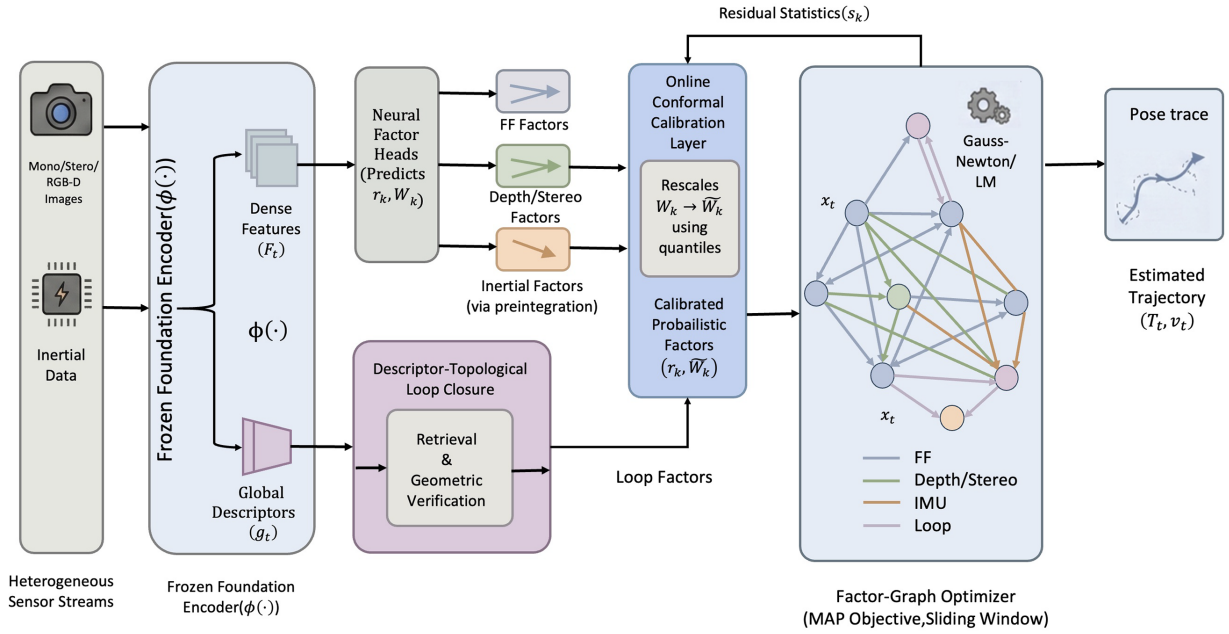


Figure 1: System overview of conformal-calibrated foundation-factor graph SLAM (CF^2-*SLAM*). Heterogeneous sensors → frozen foundation features → probabilistic factor heads → online conformal calibration → factor-graph optimization for trajectory and optional map.

4.3 Graph Construction and Edge Selection

Optimization is performed over a sliding window of N states with optional loop-closure edges. Temporal edges ensure local observability; sparse covisibility edges add redundancy without quadratic connectivity. Loop candidates are proposed by retrieving neighbors in descriptor space and inserted only after geometric verification (Section 4.5), since false high-confidence loops can bias the entire graph.

4.4 Probabilistic Learned Factors

Factor types and modality dependencies are summarized in Fig. 2; only factors supported by dataset fields (Section 5) are enabled to avoid modality leakage.

Four factor families are used: a foundation-feature (FF) factor, a depth/disparity (D) factor when depth is present, an IMU preintegration factor when IMU exists [14], and a loop-closure (LC) factor after verification. The descriptor-topological loop-closure module described below is therefore not a fifth factor family; rather, it is the proposal-and-verification pipeline whose accepted outputs are instantiated as LC factors in the graph. For FF, given an estimated relative pose $\mathbf{T}_{ij}$ and optional depth $\mathbf{D}_i$, pixel u is warped from i to j and features are compared:

$$\mathbf{r}_{ij}^{\mathrm{FF}}(u) = \mathbf{F}_i(u) - \mathbf{F}_j(\pi(\mathbf{T}_{ij}\,\Pi(u, \mathbf{D}_i))). \tag{3}$$

For RGB-D depth consistency:

$$r_{ij}^{\mathrm{D}}(u) = D_i(u) - \hat{D}_{i\leftarrow j}(u; \mathbf{T}_{ij}, D_j), \tag{4}$$

and for stereo an analogous disparity form applies. For loop closure, after verification a relative pose constraint is added in $\mathfrak{se}(3)$:

$$\mathbf{r}_{ij}^{\mathrm{LC}} = \log\left(\mathbf{T}_{ij}^{-1}\,\mathbf{T}_i^{-1}\mathbf{T}_j\right) \in \mathbb{R}^6. \tag{5}$$

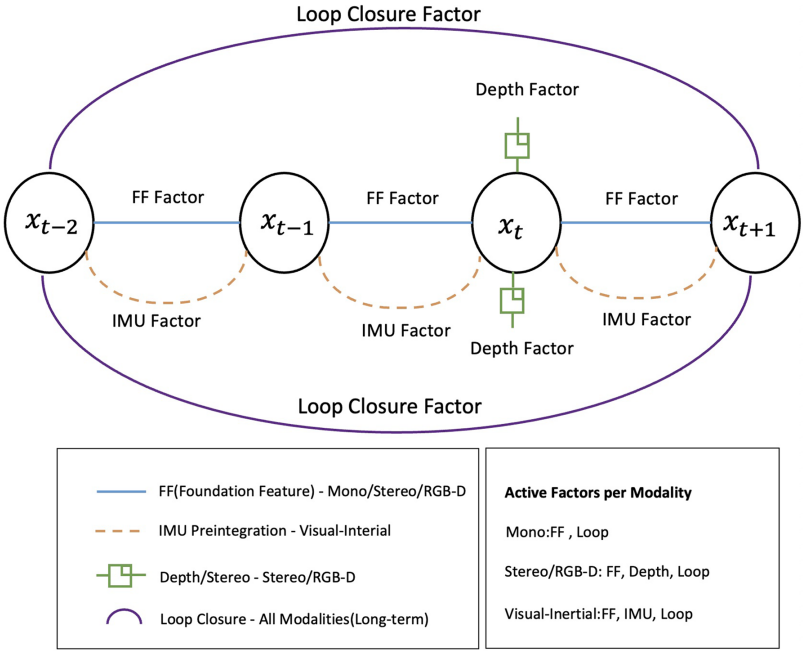


Figure 2: Unified factor graph for CF^2-*SLAM*. Edge availability depends on sensor modalities (monocular/stereo/red-green-blue-depth (RGB-D)/inertial measurement unit (IMU)) and verified loop closures.

The IMU family uses the standard preintegrated residual over pose, velocity, and bias states; d_τ in the calibration section below always denotes the dimension of the corresponding residual block for factor family τ.

Each factor head outputs a positive-definite covariance $\mathbf{W}_k$. Diagonal $\mathbf{W}_k = \mathrm{diag}(\exp(\mathbf{s}_k))$ (log-variances $\mathbf{s}_k$) or a Cholesky parameterization is used when needed. Robust losses $\rho_k(\cdot)$ (e.g., Huber/Cauchy) reduce sensitivity to sporadic outliers; calibration (below) targets systematic misweighting under shift. In particular, the covariance parameterization is chosen so that the initial $\mathbf{W}_k$ is symmetric positive definite (SPD); the robust loss is applied to the Mahalanobis energy in Eq. (2), whereas the conformal layer only rescales the covariance magnitude for a factor family and does not change the residual definition itself.

4.5 Descriptor-Topological Loop Closure with Geometric Verification

Loop closure separates candidate proposal from constraint insertion (Fig. 3). In this article, "descriptor-topological" refers to this retrieval-and-verification pipeline, while the actual graph element added after a successful check is the LC factor defined above. Candidates are retrieved using $\mathbf{g}_t$ and a memory bank; verification uses dense matching (e.g., LoFTR [15]) and robust Perspective-n-Point (PnP)/SE(3) estimation. Only verified loops are inserted, controlling false-loop contamination while avoiding dataset-specific place-recognition supervision.

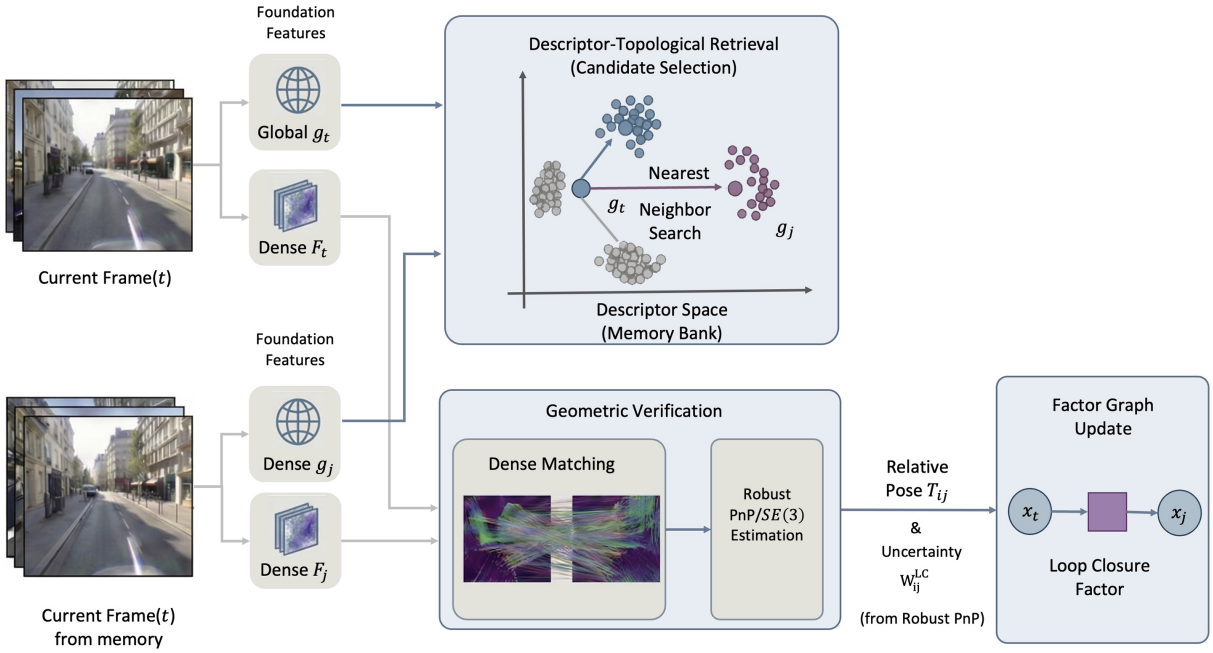


Figure 3: Descriptor-topological loop closure. Foundation descriptors propose candidates; dense matching verifies geometry; accepted closures become loop-closure (LC) factors in the graph.

To balance accuracy and efficiency, proposal and verification are decoupled: descriptor retrieval generates hypotheses, and dense matching with robust pose estimation acts as a conservative gate, executed sparsely (e.g., on keyframes) with a capped number of candidates per query; the reported frame rate, measured in frames per second (FPS), includes the amortized verification cost. Descriptor retrieval alone is high-recall but insufficiently conservative for graph insertion under perceptual aliasing and repeated structures; dense geometric verification is therefore required before adding a loop-closure factor.

In dynamic scenes, non-stationarity mainly impacts retrieval and correspondence; the verification gate rejects inconsistent hypotheses before graph insertion, and dynamic-aware masking or temporal-consistency checks can be incorporated when needed.

4.6 Online Conformal Calibration of Factor Covariances

Under domain shift, predicted uncertainties can become systematically miscalibrated, altering factor influence in GN/LM. Covariances are therefore rescaled online using residual statistics (Fig. 4).

For factor k, define the score

$$s_k = \sqrt{\mathbf{r}_k^\top \mathbf{W}_k^{-1} \mathbf{r}_k}. \tag{6}$$

For each factor type τ, a window of scores is maintained and the empirical $(1-\alpha)$ quantile q_τ^{obs} is computed. If well calibrated and approximately Gaussian, then s_k^2 follows $\chi^2(d_\tau)$ with residual dimension d_τ. The target quantile is set to $q_\tau^{\text{tar}} = \sqrt{\chi^2_{d_\tau}(1-\alpha)}$ and the covariance is rescaled as:

$$\mathbf{W}_k \leftarrow \gamma_\tau \mathbf{W}_k, \qquad \gamma_\tau = \left(\frac{q_\tau^{\text{obs}}}{q_\tau^{\text{tar}}} \right)^2. \tag{7}$$

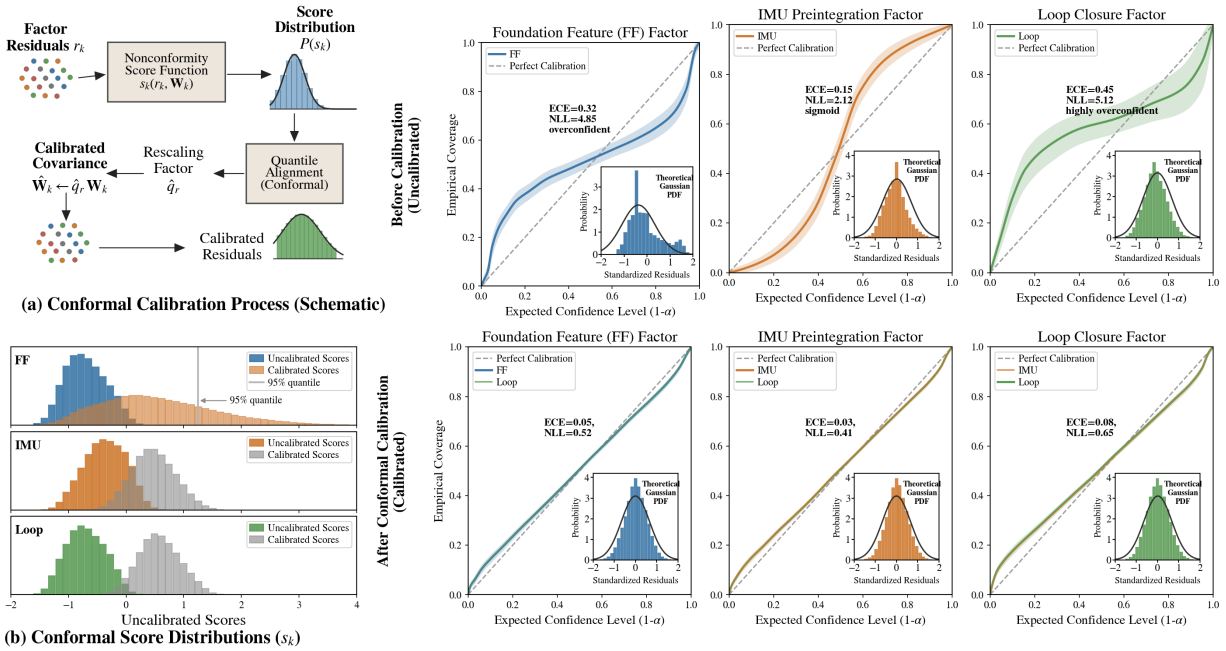


Figure 4: Online conformal calibration. For each factor type, match observed residual quantiles to target quantiles and rescale covariances accordingly.

This increases covariance when residuals are larger-than-expected and decreases it when residuals are smaller-than-expected. Warm-up and caps on γ_τ are used to reduce sensitivity to early transients and abrupt regime changes. Because $\gamma_\tau > 0$, the rescaled covariance remains SPD whenever the initial $\mathbf{W}_k$ is SPD; the update therefore changes only the overall scale of a factor family, not its internal correlation structure. In this implementation, robust losses and conformal rescaling play complementary roles: the robust loss suppresses large instantaneous outliers at the factor level, whereas conformal rescaling corrects slower covariance-scale mismatch accumulated over a window. It is also emphasized that Eq. (7) is used here as an online conformal-style calibration rule rather than as a strict conformal-prediction guarantee. Sequential SLAM violates exchangeability through temporal correlation, state-feedback, and changing operating regimes, so the χ^2 target should be interpreted as an approximate reference for residual-scale matching rather than a finite-sample coverage guarantee.

4.7 Training Details

Factor heads are trained in two stages, with pretraining on TartanAir to initialize residual and covariance behavior. Optimization uses AdamW (learning rate 1×10^{-4}, batch size 16, 100k iterations). A probabilistic objective aligned with inference is minimized:

$$\mathcal{L} = \sum_k \left(\frac{1}{2} \mathbf{r}_k^\top \mathbf{W}_k^{-1} \mathbf{r}_k + \frac{1}{2} \log |\mathbf{W}_k| \right) + \lambda \mathcal{L}_{\text{aux}}, \tag{8}$$

where $\mathcal{L}_{\text{aux}}$ includes smoothness priors and self-supervised consistency terms. All experiments use NVIDIA RTX 4090 graphics processing units (GPUs).

The inference procedure is summarized in Algorithm 1.

Algorithm 1: CF^2-*SLAM* inference (windowed factor graph with conformal calibration)

Require: Sensor stream $\{\mathcal{M}_t\}$, calibration, window size N, Gauss-Newton (GN) iterations K
Initialize states in the active window (e.g., constant velocity)
for each new time step t **do**
    Extract foundation features $(\mathbf{F}_t, \mathbf{g}_t) \leftarrow \phi(\mathcal{M}_t)$
    Select edges (temporal neighbors, covisibility, loop candidates) per Section 4.3
    Build probabilistic factors $\{\mathbf{r}_k, \mathbf{W}_k\}$ via factor heads per Section 4.4
    Update conformal statistics per factor type; rescale $\mathbf{W}_k$ via Eq. (7)
    for $k = 1$ to K **do**
        Linearize residuals; solve GN/LM update
        Update states on SE(3)(3)
    end for
    Output pose and optional map
end for

5 Datasets and Protocols

Sensor fields are documented to support cross-modal comparisons, and only factors supported by available measurements are enabled. Evaluation is conducted on seven public benchmarks that jointly cover outdoor driving (KITTI Odometry, KITTI-360), aerial visual-inertial simultaneous localization and mapping on EuRoC Micro Aerial Vehicle (MAV), and indoor RGB-D tracking/relocalization/mapping (TUM RGB-D, ScanNet, 7-Scenes), with TartanAir used for broad pretraining and stress-testing under diverse simulated conditions. Specifically, KITTI/KITTI-360 provide rectified stereo driving sequences for odometry and loop closure under appearance change; EuRoC provides synchronized stereo + IMU for visual-inertial odometry (VIO) evaluation under aggressive motion; TUM RGB-D and 7-Scenes emphasize indoor tracking and relocalization with depth; ScanNet supports large-scale indoor RGB-D tracking and dense reconstruction; TartanAir supplies diverse environments and modalities to initialize factor heads for transfer. Dataset fields and typical tasks are summarized in Table 1.

Table 1: Dataset fields and supported tasks (based on official documentation). "RGB" denotes the color image stream, "RGB-D" denotes red-green-blue-depth, "IMU" denotes inertial measurement unit, "GT Traj" indicates ground-truth (GT) trajectory/state sufficient for absolute trajectory error (ATE)/relative pose error (RPE), "Intr." indicates published intrinsics/calibration, and "SLAM" denotes simultaneous localization and mapping.

Dataset	RGB	Stereo	Depth	IMU	Intr.	GT Traj/State	Typical SLAM Tasks
KITTI Odometry [16,17]	✓	✓	–	–	✓	✓ (00–10)	Visual odometry/SLAM, loop closure
KITTI-360 [18]	✓	✓	–	(global positioning system/inertial navigation system)	✓	✓	Long-term odometry, mapping
EuRoC MAV [19]	✓	✓	–	✓	✓	✓	Visual-inertial odometry (VIO), loop closure
TUM RGB-D [20]	✓	–	✓	(robot operating system (ROS) bag)	✓	✓ (most)	RGB-D SLAM, dense mapping
ScanNet [21]	✓	–	✓	–	✓	✓ (poses)	Tracking, dense reconstruction
7-Scenes	✓	–	✓	–	default depth intrinsics	✓	Relocalization
TartanAir	✓	(varies)	✓	(varies)	✓	✓	Pretraining, stress tests

6 Experimental Setup

6.1 Baselines

Baselines and modality requirements are summarized in Table 2. Comparison is made only where required fields exist (Table 1), and runtime/hardware is reported when available.

Table 2: Baseline systems and modality requirements. Only compare on datasets where required fields exist (see Table 1). "SLAM" denotes simultaneous localization and mapping, "VO" denotes visual odometry, "RGB-D" denotes red-green-blue-depth, "IMU" denotes inertial measurement unit, and "mono" denotes monocular.

Category	Methods	Required Fields
Feature-based SLAM	ORB-SLAM3 [6]	Mono/stereo/RGB-D, intrinsics; optional IMU
Direct VO/SLAM	DSO [22], SVO2 [23]	Mono/stereo, intrinsics
Visual-Inertial	OKVIS [24], VINS-Mono [14], VINS-Fusion	Stereo/mono + IMU, intrinsics, IMU calibration
Deep VO/SLAM	DROID-SLAM [25]	Mono/stereo/RGB-D (images; optional depth)
Dense RGB-D mapping	BundleFusion [26], NICE-SLAM [27]	RGB-D, intrinsics
Learned matching front-end	SuperPoint [28], LoFTR [15]	Images

6.2 Evaluation Protocol and Reporting

Each sequence is evaluated over R independent runs; mean ± std is reported when applicable. Failure rate is Fail% = $\frac{\#\text{failed runs}}{\#\text{total runs}} \times 100\%$, where a run is counted as failed if optimization becomes numerically invalid (e.g., NaN/Inf states or normal equations) or if tracking/trajectory estimation breaks down irrecoverably before a complete trajectory can be aligned to ground truth. Absolute trajectory error (ATE) uses a single global least-squares alignment with the standard Umeyama procedure; stereo, RGB-D, and visual-inertial settings use SE(3)(3) alignment, whereas monocular settings use Sim(3)(3) when scale is not observable. Key settings are listed in Table 3.

Table 3: Evaluation settings used throughout (unless otherwise stated). "GN" denotes Gauss-Newton, "Seqs" denotes sequences, "seq" denotes sequence, "iters" denotes iterations, "MH" denotes Machine Hall, "V" denotes Vicon Room, "Val" denotes validation, and "Sel." denotes selected.

Dataset	Seqs	Runs/seq (R)	Resolution	Window (N)	GN Iters (K)
KITTI	00–10	5	1241 × 376	7	12
KITTI-360	00, 02-06, 09	5	1408 × 376	7	12
EuRoC	MH + V (All)	10	752 × 480	10	10
TUM	Fr1/2/3 (Sel.)	5	640 × 480	10	10
7-Scenes	All 7	5	640 × 480	12	8
ScanNet	Val (10 seqs)	3	640 × 480	12	8

Reported runtime/frames per second (FPS) includes feature extraction, factor construction, windowed optimization, and the amortized loop-closure cost under the configured proposal/verification schedule. Per-stage timing for retrieval and verification is additionally logged to support reproducibility.

6.3 Metrics

Absolute trajectory error (ATE)/relative pose error (RPE) [20], loop-closure and relocalization precision/recall, mapping quality where depth exists, uncertainty metrics, and efficiency (FPS/memory) are reported. For uncertainty, let $\mathbf{e}_t \in \mathbb{R}^d$ be the pose error and Σ_t the posterior covariance of the pose block at the solved state, approximated by the corresponding block of the inverse GN/LM normal matrix (local Laplace approximation, not an exact posterior). Assuming Gaussian errors, the average negative log-likelihood (NLL) is

$$\mathrm{NLL} = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{2}\mathbf{e}_t^{\top}\Sigma_t^{-1}\mathbf{e}_t + \frac{1}{2}\log|\Sigma_t|\right) + C, \tag{9}$$

with constant C shared across methods. Expected calibration error (ECE) follows standard binning of nominal vs. empirical coverage; reliability curves and score distributions are reported in Section 7.5. False loops per kilometer use traveled distance along the reference trajectory.

7 Results and Analysis

7.1 Outdoor Odometry (KITTI/KITTI-360)

Outdoor trajectory estimation is evaluated under Section 6.2. Results are summarized in Table 4, and representative trajectories are visualized in Fig. 5. On KITTI-360, CF^2-*SLAM* shows reduced drift trends under broader appearance variation, aligning with conservative verified loop insertion and moderated factor influence via online covariance rescaling.

Table 4: Outdoor trajectory accuracy on KITTI Odometry (averaged over sequences 00–10) and KITTI-360. Monocular uses Sim(3) alignment; stereo uses SE(3). Absolute trajectory error (ATE) root-mean-square error (RMSE) [m]; relative pose error (RPE): translation [%]/rotation [°/100 m]; runtime is reported in frames per second (FPS).

Method	Modality	KITTI ATE↓	KITTI RPE-t↓	KITTI RPE-r↓	KITTI-360 ATE↓	Runtime (FPS)↑
ORB-SLAM3 [6]	Stereo	0.82 ± 0.15	0.95	0.28	3.42 ± 1.1	32.5
DSO [22]	Mono	2.51 ± 0.85	2.15	0.54	12.8 ± 4.2	28.0
DROID-SLAM [25]	Stereo	0.65 ± 0.12	0.78	0.21	2.15 ± 0.6	14.5
CF^2-*SLAM*	Mono	1.05 ± 0.22	1.25	0.38	3.25 ± 1.1	18.2
CF^2-*SLAM*	Stereo	0.64 ± 0.09	0.75	0.19	2.12 ± 0.5	17.5

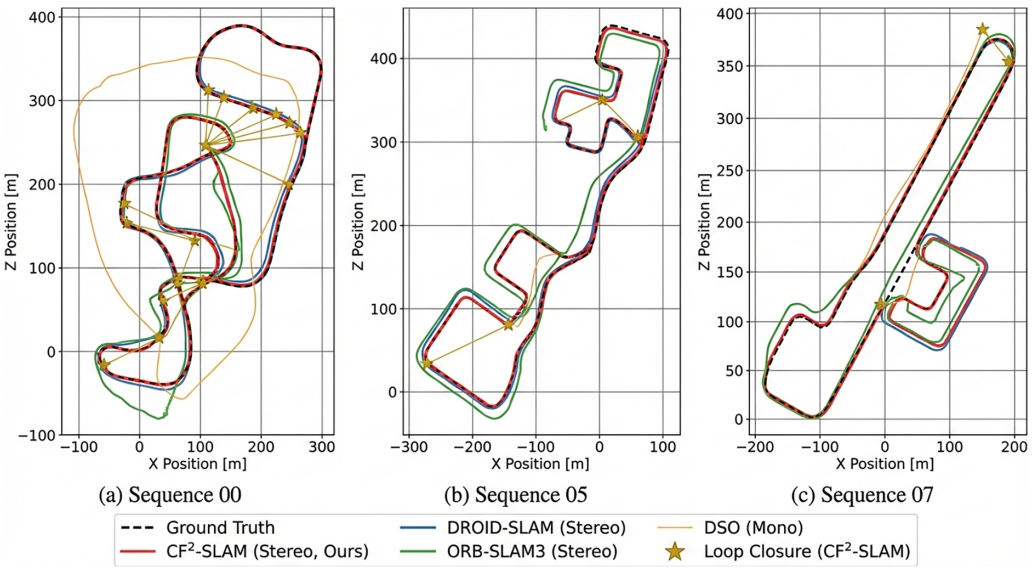


Figure 5: Qualitative trajectories on (**a**) KITTI-360 Sequence 00, (**b**) KITTI Odometry Sequence 02, and (**c**) KITTI Odometry Sequence 05. Alignment follows Section 6.2.

7.2 Visual-Inertial SLAM (EuRoC MAV)

On EuRoC with stereo + IMU, CF^2-*SLAM* attains errors comparable to established VIO baselines (Table 5) and remains stable across runs. Calibration primarily acts as a moderation mechanism when visual residuals become temporarily unreliable (e.g., motion blur), reducing their dominance relative to inertial constraints.

Table 5: EuRoC Micro Aerial Vehicle (MAV) visual-inertial results. Absolute trajectory error (ATE) root-mean-square error (RMSE) [m] is averaged over Machine Hall (MH 01–05) and Vicon Room (V1–V2). "IMU" denotes inertial measurement unit, "Fail%" denotes failure rate computed over total runs (N = 50 for MH, N = 20 for V), and runtime is reported in frames per second (FPS).

Method	Sensors	ATE (MH)↓	Fail% (MH)↓	ATE (V)↓	Fail% (V)↓	Runtime (FPS)↑
OKVIS [24]	Stereo + IMU	0.052	0.0%	0.085	5.0%	25.0
VINS-Fusion	Stereo + IMU	0.044	0.0%	0.038	5.0%	22.0
ORB-SLAM3 [6]	Stereo + IMU	0.033	0.0%	0.041	0.0%	24.5
CF^2-*SLAM*	Stereo + IMU	0.034	0.0%	0.037	0.0%	16.8

7.3 Indoor RGB-D Tracking and Auxiliary Mapping Evidence (TUM, ScanNet, 7-Scenes)

Indoor RGB-D scenes contain textureless regions and perceptual aliasing. Across TUM, ScanNet, and 7-Scenes (Table 6), CF^2-*SLAM* achieves competitive tracking accuracy and strong relocalization, indicating that foundation features and verified loop insertion help maintain consistency under ambiguity. Because the main contribution of this work is conformal factor calibration rather than dense reconstruction itself, the ScanNet mesh results below are treated as auxiliary evidence that better-calibrated factor weighting also preserves downstream geometric consistency, rather than as the primary proof of the method.

Table 6: Indoor trajectory accuracy on TUM red-green-blue-depth (RGB-D), ScanNet, and 7-Scenes. Values are absolute trajectory error (ATE) root-mean-square error (RMSE) [cm], runtime is reported in frames per second (FPS), and "Reloc." denotes relocalization.

Method	Input	TUM ATE↓	ScanNet ATE↓	7-Scenes ATE↓	Reloc. Rate↑	FPS↑
ORB-SLAM3 [6]	RGB-D	1.8	7.2	4.5	82.5%	30.0
BundleFusion [26]	RGB-D	2.5	8.5	5.1	–	10.5
NICE-SLAM [27]	RGB-D	2.1	9.8	6.5	–	0.8
DROID-SLAM [25]	RGB-D	2.0	6.4	3.8	91.0%	12.5
CF^2-*SLAM*	RGB-D	1.9	6.3	3.75	92.5%	16.0

As auxiliary evidence of graph consistency on RGB-D data, quantitative dense mapping results on ScanNet are reported in Table 7, and qualitative examples are shown in Fig. 6.

Table 7: Dense mapping quality on ScanNet (ground-truth mesh). Accuracy (Acc.) [cm] is distance to reference mesh; Completeness (Comp.) [%] uses a 5 cm threshold; Chamfer is in cm; peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM) are also reported.

Method	Dataset	Acc. (cm)↓	Comp. (%)↑	Chamfer↓	PSNR↑	SSIM↑
BundleFusion [26]	ScanNet	6.8	72.1	7.5	20.1	0.78
CF^2-*SLAM*	ScanNet	6.5	74.5	7.2	24.2	0.84

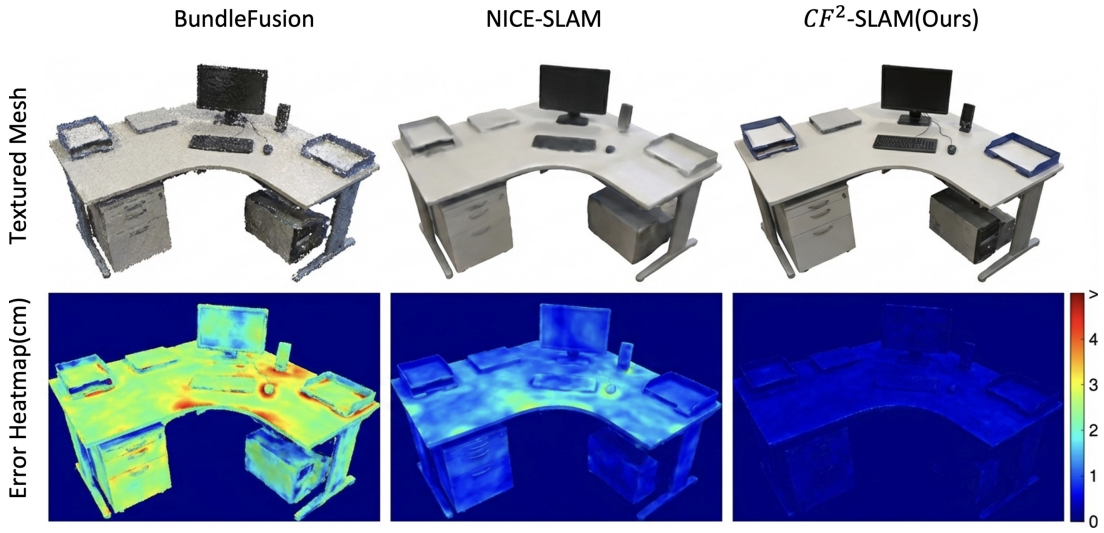


Figure 6: Qualitative dense reconstruction on ScanNet. Top: reconstructed meshes. Bottom: distance-to-mesh error heatmaps (shared scale).

7.4 Loop Closure and Relocalization

Table 8 reports loop closure precision/recall and relocalization. Geometric verification filters most spurious retrieval candidates, while descriptor-based proposal improves recall under large viewpoint/appearance changes. To directly isolate the necessity of the second stage, an additional comparison

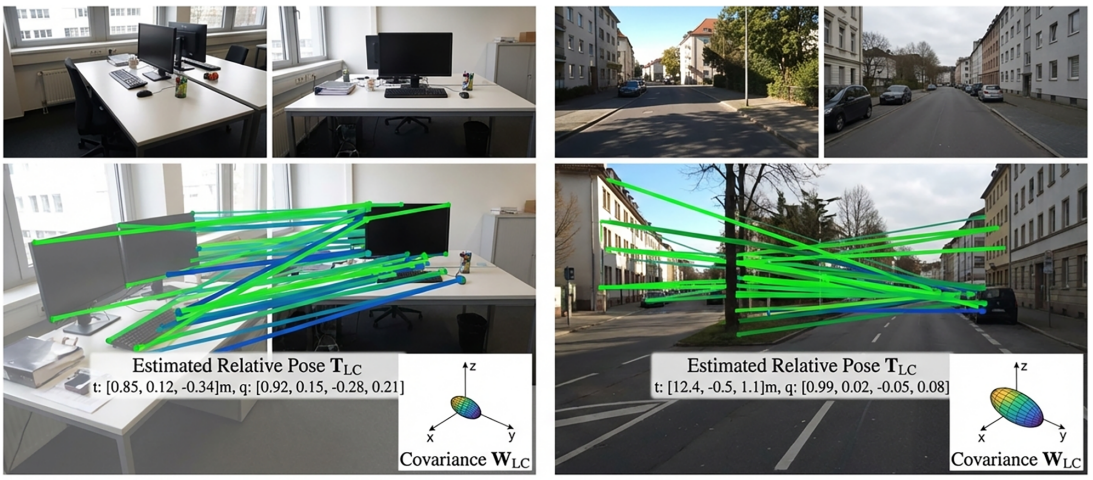
between descriptor-only loop closure and descriptor retrieval followed by geometric verification is reported in Table 9. The verification stage is intended as a conservative gate before loop-factor insertion, since even a small number of false loop constraints can bias subsequent graph optimization. Qualitative verified examples are shown in Fig. 7.

Table 8: Loop closure (LC) and relocalization metrics computed on loop factors inserted after geometric verification. "Prec." denotes precision, "Reloc@d" denotes relocalization within distance threshold d, and False LC/km denotes false loop closures per traveled kilometer.

Method	Dataset	LC Prec.↑	LC Recall↑	Reloc@5 cm ↑	Reloc@10 cm ↑	False LC/km↓
ORB-SLAM3 [6]	KITTI-360	100%	45.2%	65.5%	78.2%	0.00
DROID-SLAM [25]	KITTI-360	95.5%	72.1%	82.1%	88.5%	0.05
CF^2-*SLAM*	KITTI-360	99.5%	73.2%	83.5%	89.8%	0.04
ORB-SLAM3 [6]	TUM Loop	100%	52.8%	58.4%	70.1%	0.00
CF^2-*SLAM*	TUM Loop	99.8%	54.5%	60.2%	72.5%	0.01

Table 9: Direct ablation of descriptor-only loop closure vs. descriptor retrieval + geometric verification on KITTI-360. "LC" denotes loop closure, absolute trajectory error (ATE) is reported in meters, and runtime is reported in frames per second (FPS).

Setting	LC Prec.↑	LC Recall↑	False LC/km↓	ATE [m]↓	FPS↑
Descriptor-only	94.1%	75.6%	0.36	3.38	18.4
Descriptor + geometric verification (CF^2-*SLAM*)	99.5%	73.2%	0.04	2.12	17.5



(a) Difficult Viewpoint Change (TUM Loop) (b) Appearance Change (KITTI-360)

Figure 7: Loop-closure examples passing geometric verification: query/retrieved frames, verified correspondences, and inserted loop constraint.

7.5 Uncertainty, Robustness, and Cross-Sensor Shift

Uncertainty is evaluated with negative log-likelihood (NLL)/expected calibration error (ECE) and robustness with failure rate under zero-shot transfer, with emphasis on how conformal calibration behaves under cross-dataset and cross-sensor shift. The transfer from KITTI to KITTI-360 mainly changes appearance statistics, scene layout, and long-horizon driving context under the same stereo sensing regime, whereas the transfer from TartanAir to EuRoC additionally introduces synthetic-to-real, motion-regime, and stereo + IMU VIO differences. In fixed-noise SLAM systems, such shifts can mis-scale visual, depth, and inertial factor families, causing some modalities to dominate the normal equations. The learned heads provide modality-conditioned initial covariances, and the conformal layer further corrects residual scale mismatch online per factor family. Table 10 and Fig. 8 show that conformal calibration reduces ECE and is accompanied by lower failure rates across both transfer settings, consistent with more reliable factor weighting under shift rather than merely better in-domain fitting.

Table 10: Cross-dataset and cross-sensor uncertainty calibration and robustness under zero-shot transfer. Negative log-likelihood (NLL), expected calibration error (ECE), and absolute trajectory error (ATE) follow Section 6.3; "Fail%" denotes failure rate and "Med." denotes median.

Setting	Transfer	NLL↓	ECE↓	Fail%↓	Med. ATE [m]↓	95% ATE [m]↓
Uncalibrated	KITTI → KITTI-360	2.45	0.32	2.9%	2.20	2.85
Conformal	KITTI → KITTI-360	2.12	0.06	0.0%	2.12	2.65
Uncalibrated	TartanAir → EuRoC	1.80	0.41	5.5%	0.055	0.095
Conformal	TartanAir → EuRoC	1.40	0.10	1.8%	0.037	0.082

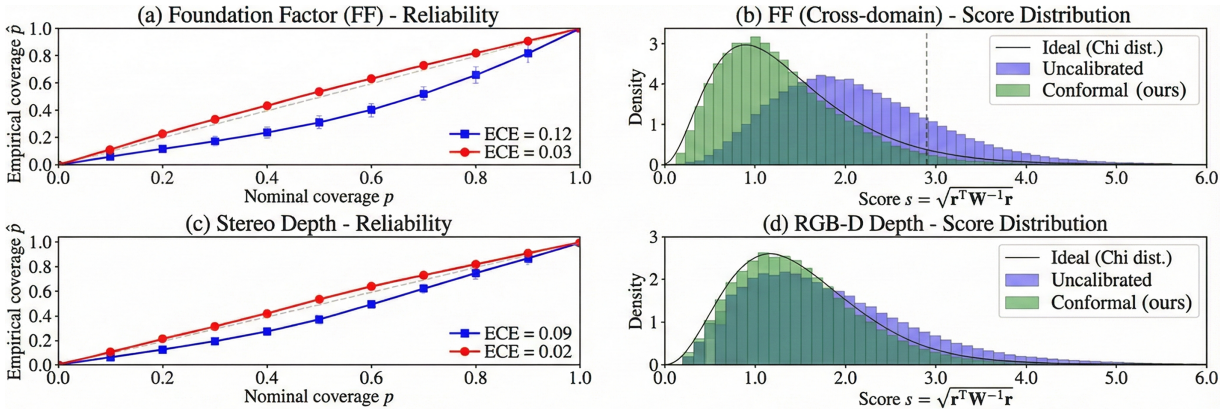


Figure 8: (Continued)

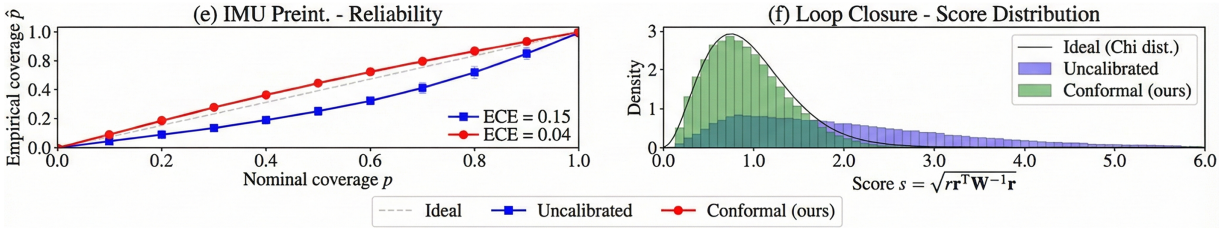


Figure 8: Uncertainty calibration diagnostics. Top: reliability diagrams. Bottom: score distributions under transfer.

8 Ablations

Key components are ablated under the same protocol, with particular emphasis on calibration under shift and on the loop-closure design. Table 11 shows that removing conformal calibration increases both ATE and failure rate (A1 vs. A0), replacing the foundation backbone degrades transfer robustness (A2), and geometric-only loop closure substantially increases drift (A3), indicating the value of descriptor-topological proposal. Together with the cross-shift results in Table 10, these ablations support the claim that conformal reweighting is the primary mechanism improving robustness under transfer. Table 9 further isolates the contribution of the geometric verification stage beyond descriptor retrieval alone.

Table 11: Module ablation on KITTI-360 (Validation). A0: full method. A2 replaces DINOv2 with ResNet50. "LC" denotes loop closure, "ATE" denotes absolute trajectory error, "calib." denotes calibration, "w/o" denotes without, and "descriptor-topo" denotes descriptor-topological, and Fail% corresponds to integer failures (1/35 ≈ 2.9%).

Variant	Backbone	Conformal calib.	LC factor	Toggles	ATE [m]↓	Fail%↓
A0 (full)	DINOv2	✓	Descriptor-topo	Full	2.12	0.0%
A1	DINOv2	–	Descriptor-topo	Full	2.20	2.9%
A2	ResNet50	✓	Descriptor-topo	Full	3.12	5.7%
A3	DINOv2	✓	Geometric-only	Full	4.55	2.9%
A4	DINOv2	✓	Descriptor-topo	w/o depth prior	2.95	2.9%
A5	DINOv2	✓	Descriptor-topo	w/o motion prior	2.25	2.9%

9 Discussion and Limitations

Solver stability in SLAM is tightly coupled to factor weighting. In this work, "stability" is used in an operational sense: fewer failed runs, fewer catastrophic drifts/divergences, and less sensitivity to misweighted factor families during sequential GN/LM updates under transfer. Under domain shift, miscalibrated uncertainties can overweight unreliable constraints and degrade conditioning, producing drift or divergence. The conformal rescaling rule in Eq. (7) provides a simple online mechanism to adjust covariance magnitudes using observed residual statistics, without retraining, and the empirical evidence in Tables 10 and 11 should be interpreted in this operational sense rather than as a stand-alone spectral conditioning proof. A representative divergence-vs.-recovery comparison is shown in Fig. 9.

Limitations include: (i) foundation backbones increase compute relative to compact convolutional neural network (CNN) front-ends, motivating distillation or reduced token resolution for real-time deployment; (ii) calibration relies on bounded windows and approximate stationarity, so abrupt regime changes can challenge residual statistics despite warm-up and caps; (iii) the conformal-style update is empirical and does not provide strict exchangeability-based coverage guarantees in sequential SLAM; and (iv) loop

closure remains subject to the retrieval/verification trade-off under severe viewpoint changes and repeated structures. In highly dynamic scenes with large non-rigid occluders, retrieval and correspondences may degrade; integrating motion- or instance-aware masking with dynamic-aware matching is a natural extension that preserves the current solver formulation. The multi-stage loop-closure pipeline also introduces controllable overhead; sustained real-time operation depends on scheduling choices such as keyframe triggering, candidate caps, and (when available) asynchronous verification.

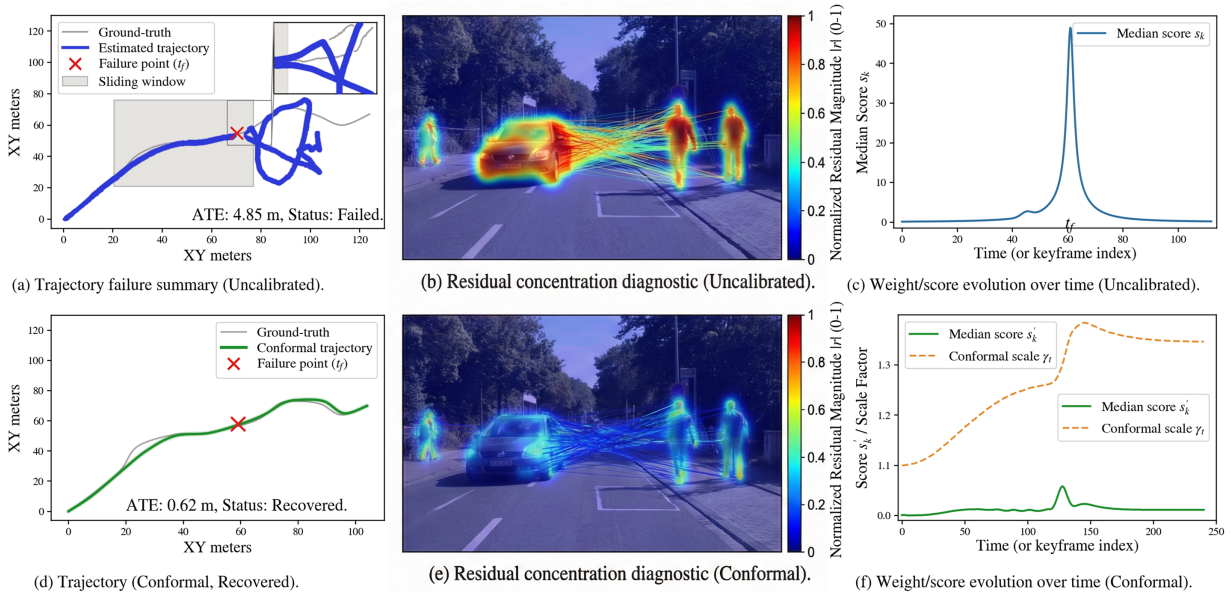


Figure 9: Failure case analysis on KITTI→KITTI-360. Row 1: divergence without calibration. Row 2: recovery with conformal covariance rescaling.

10 Conclusion

This article introduced CF^2-*SLAM*, a sensor-agnostic SLAM framework combining frozen foundation representations, probabilistic factor heads, and a classical factor-graph back-end equipped with online conformal calibration. By rescaling factor covariances using residual quantiles per factor type, the method specifically targets systematic misweighting under cross-dataset and cross-sensor shift and yields more reliable optimization behavior under transfer, reflected in lower failure rates and reduced sensitivity to misweighted factors. Descriptor-topological loop closure was also described based on foundation descriptors with geometric verification for conservative loop insertion, and dataset sensor fields and evaluation protocols were documented for reproducible cross-modal comparison.

Acknowledgement: Not applicable.

Funding Statement: The author received no specific funding for this study.

Availability of Data and Materials: The datasets analyzed in this study are publicly available:

- KITTI Odometry: https://www.cvlibs.net/datasets/kitti/eval_odometry.php
- KITTI-360: https://www.cvlibs.net/datasets/kitti-360/
- EuRoC MAV: https://ethz-asl.github.io/datasets/
- TUM RGB-D: https://cvg.cit.tum.de/data/datasets/rgbd-dataset

- ScanNet: https://github.com/ScanNet/ScanNet
- 7-Scenes: https://www.microsoft.com/en-us/research/project/rgb-d-dataset-7-scenes/
- TartanAir: https://theairlab.org/tartanair-dataset/

Ethics Approval: Not applicable.

Conflicts of Interest: The author declares no conflicts of interest.

References

1. Huang H, Li L, Cheng H, Yeung SK. Photo-SLAM: real-time simultaneous localization and photorealistic mapping for monocular stereo and RGB-D cameras. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE; 2024. p. 21584–93.
2. Yin M, Wu S, Han K. IBD-SLAM: learning image-based depth fusion for generalizable SLAM. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE; 2024. p. 10563–73.
3. Oquab M, Darcet T, Moutakanni T, Vo H, Szafraniec M, Khalidov V, et al. DINOv2: learning robust visual features without supervision. arXiv:2304.07193. 2023.
4. Abdelkarim A, Voos H, Görges D. Factor graphs in optimization-based robotic control—a tutorial and review. IEEE Access. 2025;13(23):28315–34. doi:10.1109/access.2025.3534993.
5. Gibbs I, Candès EJ. Conformal inference for online prediction with arbitrary distribution shifts. J Mach Learn Res. 2024;25(162):1–36.
6. Campos C, Elvira R, Rodríguez JJG, Montiel JM, Tardós JD. ORB-SLAM3: an accurate open-source library for visual, visual-inertial, and multimap SLAM. IEEE Trans Robot. 2021;37(6):1874–90.
7. Lindenberger P, Sarlin PE, Pollefeys M. LightGlue: local feature matching at light speed. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway, NJ, USA: IEEE; 2023. p. 17627–38.
8. Edstedt J, Sun Q, Bökman G, Wadenbäck M, Felsberg M. RoMa: robust dense feature matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE; 2024. p. 19790–800.
9. Matsuki H, Murai R, Kelly PHJ, Davison AJ. Gaussian splatting SLAM. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE; 2024. p. 18039–48.
10. Keetha N, Karhade J, Jatavallabhula KM, Yang G, Scherer S, Ramanan D, et al. SplaTAM: splat track & map 3D Gaussians for dense RGB-D SLAM. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE; 2024. p. 21357–66.
11. Clarté L, Loureiro B, Krzakala F, Zdeborová L. Expectation consistency for calibration of neural networks. In: Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence. Vol. 216. London, UK: PMLR; 2023. p. 443–53.
12. Oliveira RI, Orenstein P, Ramos T, Romano JV. Split conformal prediction and non-exchangeable data. J Mach Learn Res. 2024;25(225):1–38.
13. Xue Z, Zhao S, Qi Y, Zeng X, Yu Z. Resilient routing: risk-aware dynamic routing in smart logistics via spatiotemporal graph learning. arXiv:2601.13632. 2026.
14. Qin T, Li P, Shen S. VINS-Mono: a robust and versatile monocular visual-inertial state estimator. IEEE Trans Robot. 2018;34(4):1004–20.
15. Sun J, Shen Z, Wang Y, Bao H, Zhou X. LoFTR: detector-free local feature matching with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE; 2021. p. 8922–31.
16. Geiger A, Lenz P, Urtasun R. Are we ready for autonomous driving? The KITTI vision benchmark suite. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE; 2012. p. 3354–61.
17. Geiger A, Lenz P, Stiller C, Urtasun R. Vision meets robotics: the KITTI dataset. Int J Robot Res. 2013;32(11):1231–7.

18. Liao Y, Xie J, Geiger A. KITTI-360: a novel dataset and benchmarks for urban scene understanding in 2D and 3D. IEEE Trans Pattern Anal Mach Intell. 2022;45(3):3292–310.
19. Burri M, Nikolic J, Gohl P, Schneider T, Rehder J, Omari S, et al. The EuRoC micro aerial vehicle datasets. Int J Robot Res. 2016;35(10):1157–63. doi:10.1177/0278364915620033.
20. Sturm J, Engelhard N, Endres F, Burgard W, Cremers D. A benchmark for the evaluation of RGB-D SLAM systems. In: 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems. Piscataway, NJ, USA: IEEE; 2012. p. 573–80.
21. Dai A, Chang AX, Savva M, Halber M, Funkhouser T, Nießner M. ScanNet: richly-annotated 3D reconstructions of indoor scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE; 2017. p. 5828–39.
22. Engel J, Koltun V, Cremers D. Direct sparse odometry. IEEE Trans Pattern Anal Mach Intell. 2017;40(3):611–25. doi:10.1109/tpami.2017.2658577.
23. Forster C, Zhang Z, Gassner M, Werlberger M, Scaramuzza D. SVO: semidirect visual odometry for monocular and multicamera systems. IEEE Trans Robot. 2016;33(2):249–65.
24. Leutenegger S, Lynen S, Bosse M, Siegwart R, Furgale P. Keyframe-based visual-inertial odometry using nonlinear optimization. Int J Robot Res. 2015;34(3):314–34. doi:10.1177/0278364914554813.
25. Teed Z, Deng J. DROID-SLAM: deep visual slam for monocular, stereo, and RGB-D cameras. Adv Neural Inf Process Syst. 2021;34:16558–69.
26. Dai A, Nießner M, Zollhöfer M, Izadi S, Theobalt C. BundleFusion: real-time globally consistent 3D reconstruction using on-the-fly surface reintegration. ACM Trans Graph. 2017;36(4):1.
27. Zhu Z, Peng S, Larsson V, Xu W, Bao H, Cui Z, et al. Nice-SLAM: neural implicit scalable encoding for SLAM. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE; 2022. p. 12786–96.
28. DeTone D, Malisiewicz T, Rabinovich A. SuperPoint: self-supervised interest point detection and description. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Piscataway, NJ, USA: IEEE; 2018. p. 224–36.