OBSERVABILITY-AWARE MANIFOLD FLOW MATCHING FOR DEGRADATION-RESILIENT RADAR-INERTIAL-VISUAL UAV ODOMETRY

Anonymous authors
Paper under double-blind review

ABSTRACT

Reliable metric odometry for small unmanned aerial vehicles is difficult when vision is degraded and radar returns have limited directional coverage. We present Observability-Aware Manifold Flow Matching (OAM-FM), a self-supervised radar-inertial-visual odometry framework that learns conditional transport in local pose coordinates and maps its outputs to $SE(3)$. A calibrated Doppler Jacobian supplies a directional training metric; its undamped information matrix characterizes instantaneous radar constraints without conflating numerical regularization with physical observability. The objective combines flow matching, robust Doppler residuals, and IMU preintegration over poses, velocities, and sensor biases. Optional visual features provide complementary constraints without being required at inference. A separate covariance head estimates local prediction uncertainty, whose calibration is evaluated against reference poses. The formulation distinguishes this uncertainty from the spread of flow samples: a deterministic sensor-based training endpoint alone does not establish a multimodal posterior. The method requires no dense depth or pose labels during training and is evaluated on real-world UAV odometry under radar sparsity, visual degradation, aggressive motion, and inertial bias, together with onboard latency.

1 INTRODUCTION

UAV navigation must remain stable during low-light flight, dust, textureless corridors, and aggressive motion. Cameras can degrade under these conditions, while inertial integration drifts and millimeter-wave radar provides sparse measurements affected by multipath, dynamic objects, and limited directional coverage. Existing radar-inertial estimators commonly optimize trajectories in a factor graph, whereas learning-based methods often regress a pose and fuse modalities with fixed or learned weights. A useful training objective should distinguish motion directions supported by the measurements from directions determined primarily by temporal constraints and priors.

We propose OAM-FM, which estimates an $SE(3)$ increment conditioned on a temporal sensor window, as summarized in Fig. 1. Doppler measurements constrain the velocity of the radar origin along the measured lines of sight. Let $\mathbf{v}$ and $\boldsymbol{\omega}$ be the body-origin linear velocity and angular velocity expressed in the body frame, $\boldsymbol{\ell}_{BR}$ the fixed vector from the body origin to the radar origin, and $\mathbf{n}_i$ the unit line of sight expressed in the same frame. Defining d_i as closing speed (the negative of range rate), a static return obeys

$$d_i = \mathbf{n}_i^\top (\mathbf{v} + \boldsymbol{\omega} \times \boldsymbol{\ell}_{BR}) + \varepsilon_i. \tag{1}$$

The lever arm in Eq. (1) is a sensor extrinsic, not the position of a scatterer. The Jacobian induces an information matrix whose spectrum describes instantaneous Doppler constraints. For one radar it has rank at most three, even with rich three-dimensional return geometry; inertial and temporal information are needed to resolve the remaining motion components (Huang et al., 2024a). We use a normalized, damped version of this matrix to weight a conditional flow-matching objective in local Lie-algebra coordinates.

Our contributions are: (i) a conditional flow-matching formulation with $SE(3)$ pose outputs for radar-inertial-visual odometry; (ii) a Doppler-derived training metric that allocates supervision according to supported motion directions; (iii) a physical objective combining robust radar residuals,

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

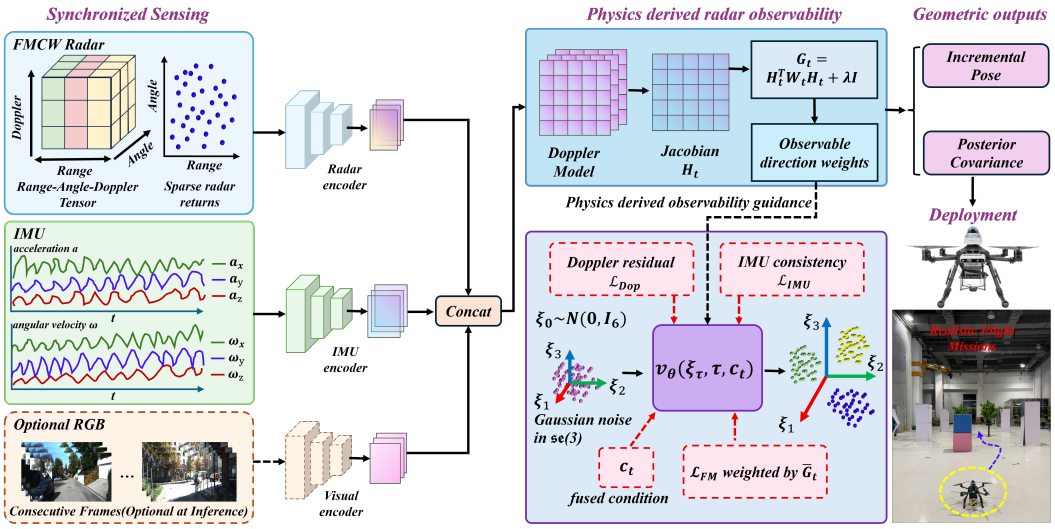

Figure 1: OAM-FM architecture. Radar, IMU, and optional visual features condition a flow in normalized local pose coordinates. Calibrated Doppler geometry weights the training objective. Flow integration and the exponential map generate pose outputs, while a separate head estimates local predictive covariance.

IMU preintegration, optional visual tracks, and a local uncertainty model; and (iv) a UAV evaluation protocol covering radar sparsity, visual corruption, motion estimation, and onboard latency. This problem is distinct from monocular depth-and-pose learning: motion-component approaches decompose optical-flow cues (Zhang et al., 2026), whereas our estimator uses Doppler and inertial physics to constrain metric motion.

2 Related Work

Radar and multimodal odometry. CFEAR and BFAR improve radar registration through conservative filtering and false-alarm control (Adolfsson et al., 2021; Alhashimi et al., 2024); subsequent studies examine their robustness (Frosi et al., 2024; Adolfsson & Hilger, 2024). Radar–inertial systems use EKF updates, multiple radars, factor graphs, or continuous-time motion priors (Michalczyk et al., 2022; Huang et al., 2024a; Michalczyk et al., 2024; Burnett et al., 2025). DeRO and Physical-RIO exploit inertial assistance and physical radar cues (Do et al., 2024; Huang et al., 2024b). LiDAR–radar–inertial fusion adds geometric constraints at the cost of an additional sensor (Nissov et al., 2024). These approaches motivate our sensor factors; our specific contribution is using Doppler directional information to weight learned conditional transport.

Visual and LiDAR estimation. VINS-Mono, VINS-Fusion, and ORB-SLAM3 combine visual geometry with inertial constraints (Qin et al., 2018; 2019; Campos et al., 2021); DROID-SLAM learns dense correspondence (Teed & Deng, 2021). LIO-SAM, FAST-LIO2, and Point-LIO exploit LiDAR–inertial geometry (Shan et al., 2020; Xu et al., 2022; He et al., 2023). Self-supervised monocular methods learn depth and motion through view reconstruction (Zhou et al., 2017; Godard et al., 2019; Guizilini et al., 2020; Watson et al., 2021). These observation models differ from sparse metric Doppler constraints. Dense prediction and driving benchmarks provide complementary context (Ranftl et al., 2021; Caesar et al., 2020; Geiger et al., 2012).

Conditional transport and uncertainty. Flow Matching and Rectified Flow learn transport fields (Lipman et al., 2023; Liu et al., 2023); diffusion and score models offer alternative generative formulations (Ho et al., 2020; Song et al., 2021). Riemannian flow policies address non-Euclidean robot states (Braun et al., 2024). Our flow instead uses a normalized local pose chart, with an exponential-map output and a measurement-derived training metric. A separate error-covariance

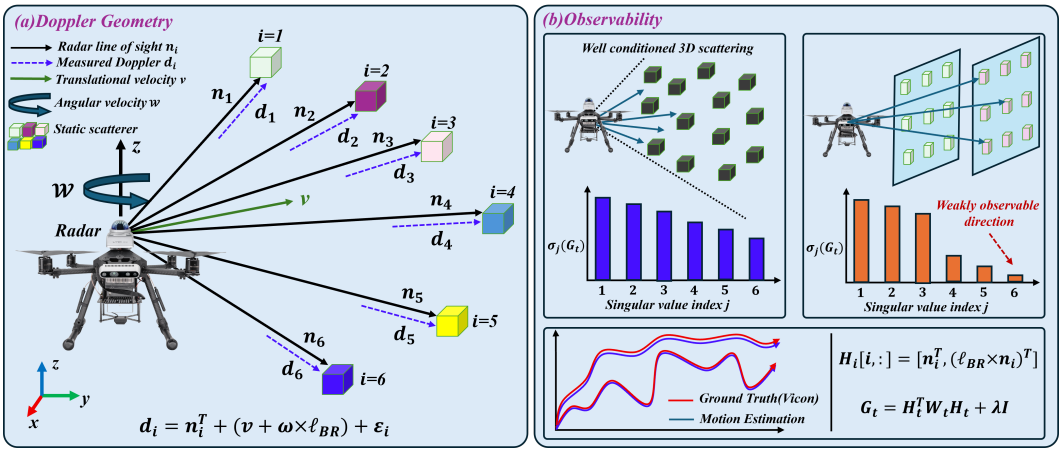

Figure 2: Doppler geometry and directional constraints. The velocity Jacobian uses the body-to-radar lever arm in Eq. (3). Angular coverage affects the undamped information spectrum; an instantaneous single radar constrains at most three twist combinations. Damping provides numerical regularization. The trajectory curves are schematic.

head follows uncertainty-learning principles (Kendall & Gal, 2017); deterministic teacher endpoints do not justify interpreting flow samples as a multimodal sensor posterior.

3 Problem Formulation

Frames, state, and pose coordinates. The transform $\mathcal{T}_{AB}$ maps coordinates from frame B to frame A; B, R, C, and W denote body/IMU, radar, camera, and world frames. The extrinsics $\mathcal{T}_{BR} = (\mathbf{R}_{BR}, \boldsymbol{\ell}_{BR})$ and $\mathcal{T}_{BC}$ are fixed and calibrated. A sensor window contains K consecutive intervals of radar tensors Y_R, IMU samples Y_I, and optional images Y_V. We estimate each relative pose $\mathcal{T}_k = \mathcal{T}_{WB_k}^{-1}\mathcal{T}_{WB_{k+1}}$, conditioned on its window. This is an incremental model, not an explicitly learned joint distribution over all window poses. Sensor timing is calibrated, and a subset of radar returns is assumed static. Ground truth is used only for evaluation.

We use translation-first coordinates $\boldsymbol{\xi} = [\boldsymbol{\rho}^\top, \boldsymbol{\phi}^\top]^\top \in \mathbb{R}^6$, with $\mathrm{Exp}(\boldsymbol{\xi}) = \exp(\boldsymbol{\xi}^\wedge)$ and Log its local inverse. In particular,

$$\mathrm{Exp}(\boldsymbol{\xi}) = \begin{bmatrix} \exp([\boldsymbol{\phi}]_\times) & \mathbf{J}_l^{SO(3)}(\boldsymbol{\phi})\boldsymbol{\rho} \\ \mathbf{0}^\top & 1 \end{bmatrix}, \qquad \mathbf{z} = \mathbf{S}^{-1}\boldsymbol{\xi}, \quad \mathbf{S} = \mathrm{diag}(s_\rho \mathbf{I}_3, s_\phi \mathbf{I}_3). \tag{2}$$

Here $[\mathbf{a}]_\times \mathbf{b} = \mathbf{a} \times \mathbf{b}$, $\mathbf{J}_l^{SO(3)}$ is the left Jacobian, and $s_\rho > 0$ and $s_\phi > 0$ are fixed translation and rotation scales from training endpoints. Thus $\boldsymbol{\rho}$ is a translation *coordinate*, not generally the translation block of the pose. Flow matching uses dimensionless $\mathbf{z}$, preventing an arbitrary mixture of meters and radians. We restrict increments to a local logarithm chart with $\|\boldsymbol{\phi}\| < \pi$. Sensor time and flow time τ are distinct; $\|\mathbf{x}\|_\mathbf{A}^2 = \mathbf{x}^\top \mathbf{A}\mathbf{x}$.

4 Radar–Inertial–Visual Motion Geometry

Doppler model and directional information. For a radar-frame return $\mathbf{p}_i^R$, set $\mathbf{n}_i = \mathbf{R}_{BR}\mathbf{p}_i^R / \|\mathbf{p}_i^R\|$. The scatterer range is not a rotational lever arm: $\mathbf{p}_i^B = \boldsymbol{\ell}_{BR} + \varrho_i \mathbf{n}_i$ implies $\mathbf{p}_i^B \times \mathbf{n}_i = \boldsymbol{\ell}_{BR} \times \mathbf{n}_i$. Differentiating Eq. (1) with respect to the instantaneous body twist $\boldsymbol{\zeta} = [\mathbf{v}^\top, \boldsymbol{\omega}^\top]^\top$ gives

$$\mathbf{H}_{i,:}^\zeta = \left[\mathbf{n}_i^\top, (\boldsymbol{\ell}_{BR} \times \mathbf{n}_i)^\top\right], \qquad \mathbf{H}^\zeta = \mathbf{N}\left[\mathbf{I}_3, -[\boldsymbol{\ell}_{BR}]_\times\right], \tag{3}$$

where $\mathbf{N}$ stacks the unit lines of sight. This is a velocity Jacobian. For a constant body twist over an interval Δt, $\boldsymbol{\xi} = \Delta t \boldsymbol{\zeta}$. With Doppler noise scales $\sigma_i > 0$ and $\mathbf{D} = \mathrm{diag}(\sigma_i)$, the whitened Jacobian

with respect to normalized pose coordinates is

$$\mathbf{H} = \mathbf{D}^{-1}\mathbf{H}^{\zeta}\mathbf{S}/\Delta t, \qquad \mathbf{r} = \mathbf{D}^{-1}(\mathbf{d} - \mathbf{h}). \tag{4}$$

Equation (4) defines a short-interval directional surrogate. During nonconstant motion, physical residuals use velocities at the return timestamps; a trajectory Jacobian must instead apply the chain rule $\sigma_i^{-1}\mathbf{H}^{\zeta}_{i,:}\,\partial\boldsymbol{\zeta}(t_i)/\partial\mathbf{z}$. An interval's endpoint pose alone does not determine its complete velocity history.

For fixed robust weights $w_i \in [0,1]$ and $\mathbf{W} = \mathrm{diag}(w_i)$, define

$$\begin{aligned} &\mathbf{A} = \mathbf{H}^{\top}\mathbf{W}\mathbf{H}, && \mathbf{G} = \mathbf{A} + \lambda\mathbf{I}_6, && \bar{\mathbf{G}} = \mathbf{G}/\operatorname{tr}(\mathbf{G}),\\ &q_j = a_j/(\operatorname{tr}(\mathbf{A}) + \epsilon_g), && a_j = \sigma_j(\mathbf{A}), && \lambda, \epsilon_g > 0. \end{aligned} \tag{5}$$

The eigenvalues a_j describe *undamped* directional information, while $\bar{\mathbf{G}}$ is a positive-definite, unit-trace training metric. Its eigenvectors, not their scalar indices, specify motion directions. The q_j describe relative directional support; absolute information also depends on the scale of $\mathbf{A}$. When no radar information is available, $\mathbf{A} = \mathbf{0}$ and $\bar{\mathbf{G}} = \mathbf{I}_6/6$: this is an optimization floor, not newly observed motion. Robust weighting makes $\mathbf{A}$ a local information surrogate, not in general an exact Fisher matrix or a calibrated inverse covariance.

Figure 2(a) illustrates the line-of-sight projection, while panel (b) relates angular coverage to directional information. Under the instantaneous single-radar model, at most three eigenvalues of $\mathbf{A}$ are nonzero (Proposition 1). Narrow azimuth or elevation coverage can further weaken these constrained combinations. A planar wall alone does not imply coplanar lines of sight through the sensor origin. The trajectory sketch illustrates possible drift, rather than proving an observability rank or an error bound. The corresponding damped normal equation and perturbation convention are derived in Appendix G.

Physical factors. The teacher and training objective use a robust Doppler penalty, bias-corrected IMU preintegration, and optional calibrated visual reprojection. IMU states include rotation, position, world velocity, and both sensor biases; pose alone does not specify a preintegration residual. Visual factors optimize positive inverse depths from valid tracks. Radar residuals evaluate body velocity at each return timestamp, with gyro-corrected lever-arm motion. The Cauchy penalty on the whitened Doppler residual gives $w_i = (1 + r_i^2/c^2)^{-1}$. Complete residuals and their noise assumptions are in Appendix A.

5 Observability-Aware Manifold Flow Matching

Conditioning and sensor-based endpoints. The encoders concatenate window features as $\mathbf{c}_t = E_R(\Upsilon_R)\|E_I(\Upsilon_I)\|E_V(\Upsilon_V)\|m_V$, where the visual feature is masked and m_V indicates its availability. Figure 1 relates this sensor pathway to the flow, geometric losses, and pose output. Geometry is computed from calibrated measurements, while the pose output comes from integrating the learned field and applying the exponential map.

Let $\mathcal{X}$ contain window poses, velocities, biases, and any visual depths. A sensor-based teacher solves

$$\mathcal{X}^{\star} = \arg\min_{\mathcal{X}} \left[\lambda_D\mathcal{L}_{\mathrm{Dop}} + \lambda_I\mathcal{L}_{\mathrm{IMU}} + \lambda_V\mathcal{L}_V + \mathcal{L}_{\mathrm{prior}}\right], \qquad \mathbf{z}_1 = \mathrm{sg}\big(\mathbf{S}^{-1}\operatorname{Log}(\mathcal{T}_k^{\star})\big). \tag{6}$$

The prior anchors the first pose to fix the coordinate gauge and constrains initial velocity and biases using the propagated state, without reference poses. Degenerate directions remain prior-dependent; anchoring the gauge does not make them physically observable. The operator sg stops gradients through the teacher target. The teacher and its robust metric are held fixed during a flow update, preventing optimization from reducing the training loss by moving its target or weights.

Conditional transport. With $\mathbf{z}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_6)$ and $\tau \sim \mathcal{U}(0,1)$, use the flow-matching path (Lipman et al., 2023)

$$\mathbf{z}_\tau = (1-\tau)\mathbf{z}_0 + \tau\mathbf{z}_1, \qquad \mathbf{u}_\tau = \mathbf{z}_1 - \mathbf{z}_0, \qquad \mathcal{L}_{\mathrm{FM}} = \mathbb{E}\Big[\|\mathbf{v}_\theta(\mathbf{z}_\tau, \tau, \mathbf{c}_t) - \mathbf{u}_\tau\|^2_{\mathrm{sg}(\bar{\mathbf{G}})}\Big]. \tag{7}$$

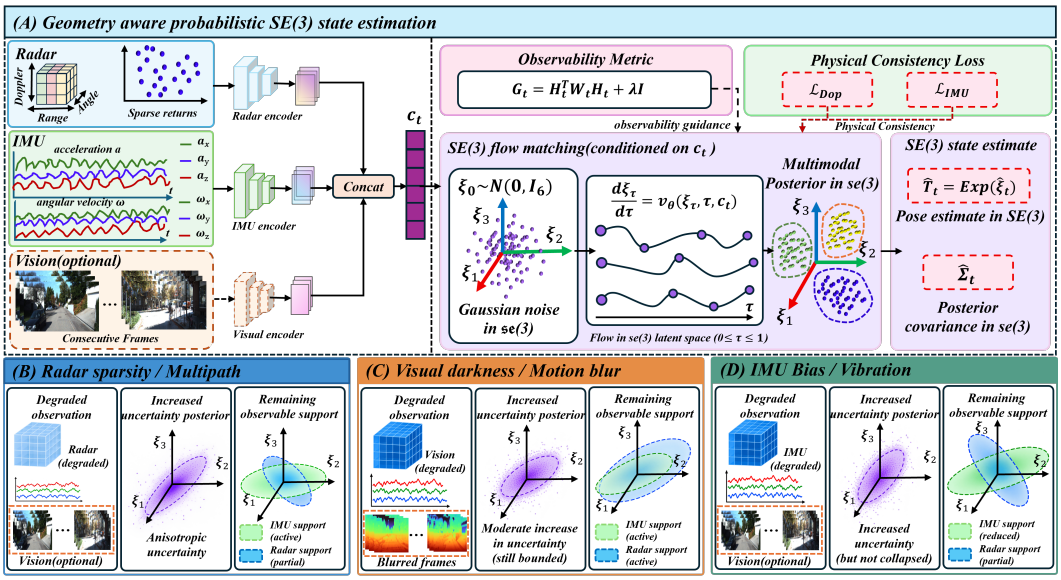

Figure 3: Conditional flow and degradation scenarios. (A) Sensor conditioning, local transport, and pose/uncertainty outputs. (B)–(D) Illustrations of radar sparsity, visual corruption, and inertial corruption. The diagrams motivate complementary sensing; sample clusters are not evidence of a calibrated multimodal posterior or guaranteed recovery under every corruption.

The expectation is over sensor windows, base samples, and flow time. The ODE is integrated in a normalized local chart, followed by $\mathcal{T} = \mathrm{Exp}(\mathbf{Sz})$. This respects the pose-group output structure, but is not an intrinsic Riemannian flow or a geodesic interpolation on $SE(3)$.

For a fixed sensor window and a deterministic endpoint, the conditional path covariance is $(1-\tau)^2\mathbf{I}_6$ and tends to zero as $\tau \to 1$. Thus multiple initializations do not by themselves establish a multimodal conditional motion posterior. Finite-step sample spread is a diagnostic of the learned transport; distributional endpoint supervision and additional validation would be required for posterior claims. Appendix G gives the endpoint limit and explains its relation to ODE regularity.

Physical training and uncertainty. Physical losses are evaluated on predicted window poses together with nuisance velocities, biases, and depths conditioned on those poses by the local factor solve. They cannot be evaluated by substituting a pose logarithm directly for a measured velocity. The stagewise objective is

$$\mathcal{L} = \lambda_F \mathcal{L}_{\mathrm{FM}} + \lambda_D \mathcal{L}_{\mathrm{Dop}} + \lambda_I \mathcal{L}_{\mathrm{IMU}} + \lambda_V \mathcal{L}_V + \lambda_U \mathcal{L}_U, \tag{8}$$

where the coefficients are nonnegative and the visual term is active only for valid images. The covariance head is trained after freezing the pose predictor. For the point estimate $\hat{\mathcal{T}}$ defined below, its target is the normalized right-local discrepancy from the sensor teacher,

$$\begin{aligned} \mathbf{e}^\star &= \mathrm{sg}\left(\mathbf{S}^{-1}\,\mathrm{Log}(\hat{\mathcal{T}}^{-1}\mathcal{T}^\star)\right), & \mathbf{P}_\psi &= \mathbf{L}_\psi\mathbf{L}_\psi^\top + \epsilon_\Sigma\mathbf{I}_6, \\ \mathcal{L}_U &= \tfrac{1}{2}\left[(\mathbf{e}^\star)^\top\mathbf{P}_\psi^{-1}\mathbf{e}^\star + \log\det\mathbf{P}_\psi + 6\log(2\pi)\right], & \hat{\mathbf{\Sigma}}_\psi &= \mathbf{S}\mathbf{P}_\psi\mathbf{S}^\top. \end{aligned} \tag{9}$$

Here $\mathbf{L}_\psi$ is a lower-triangular head output conditioned on the sensor features and predicted increment, and $\epsilon_\Sigma > 0$ is a fixed numerical floor. This Gaussian loss includes the log determinant and a positive-definite covariance. It learns teacher-discrepancy uncertainty; systematic teacher error is not identifiable from this target alone. Calibration against true motion is therefore an empirical evaluation, not a consequence of the loss.

Figure 3(A) depicts conditioning and pose/covariance output; panels (B)–(D) motivate complementary sensing under radar, visual, and inertial degradation. Geometry and physical residuals enter training, while inference integrates the learned field. These illustrations do not establish recovery under arbitrary IMU bias or vibration; measured modality controls are in Table 4.

Algorithm 1: OAM-FM inference. Encode the synchronized window and availability mask into $\mathbf{c}_t$. For each flow sample $m = 1, \ldots, M$, draw $\mathbf{z}_0^{(m)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_6)$ and integrate L Euler steps,

$$\mathbf{z}_{\ell+1}^{(m)} = \mathbf{z}_\ell^{(m)} + \Delta\tau\, \mathbf{v}_\theta(\mathbf{z}_\ell^{(m)}, \ell\Delta\tau, \mathbf{c}_t), \qquad \Delta\tau = 1/L, \quad \ell = 0, \ldots, L-1. \tag{10}$$

Decode $\hat{\boldsymbol{\xi}}^{(m)} = \mathbf{S}\mathbf{z}_L^{(m)}$ and $\hat{\mathcal{T}}^{(m)} = \mathrm{Exp}(\hat{\boldsymbol{\xi}}^{(m)})$. The point estimate is the log-coordinate mean $\hat{\mathcal{T}} = \mathrm{Exp}(\bar{\boldsymbol{\xi}})$, with $\bar{\boldsymbol{\xi}} = M^{-1}\sum_m \hat{\boldsymbol{\xi}}^{(m)}$. It is not in general a Fréchet mean. Return this pose and the head covariance $\hat{\boldsymbol{\Sigma}}_\psi$ in the right-local error convention. Sample covariance is computed from centered right-local deviations with divisor $M-1$, for $M > 1$; it is a diagnostic distinct from the learned covariance (Appendix C).

Theoretical scope. The factorization in Eq. (3) preserves the rank of $\mathbf{N}$, giving $\mathrm{rank}(\mathbf{A}) \leq 3$. Appendix B proves this and states a conditional linear-Gaussian covariance-reduction result. Neither result guarantees that physical losses reduce every realized neural prediction error.

6 Experiments

The evaluation is organized around four questions: which components improve motion estimation, whether the observability metric captures directional degeneracy, how the estimator behaves when a modality degrades, and whether flow inference fits the onboard computation budget. Training uses sensor-based self-supervision; reference poses are reserved for evaluation.

6.1 Experimental Setup and Evaluation Protocol

6.1.1 Experimental Platform

We use a customized quadrotor with a BMI088 IMU, a RealSense D435i camera, and a Jetson Orin NX (16 GB) for onboard inference (Fig. 4(a)). OAM-FM uses FMCW radar, IMU, and optional RGB inputs. The shared platform's Livox Mid-360 is excluded from the radar–inertial–visual ablations. A 32-camera Vicon Vantage system provides six-degree-of-freedom reference poses in a $50 \times 30 \times 8\,\mathrm{m}^3$ arena (Fig. 4(b,c)), using a rigid marker constellation and calibrated infrared tracking. Reference poses are time-synchronized and transformed to the body frame; intervals without valid tracking are excluded consistently.



(a) (b) (c)

Figure 4: Experimental platform and motion-capture environment. (a) Customized quadrotor with the Livox Mid-360, Pixhawk 6C, RealSense D435i, Orin NX, and reflective markers. (b) Flight arena instrumented with 32 Vicon Vantage cameras. (c) Ceiling-mounted Vicon cameras for infrared marker tracking.

Data and modality compatibility. The evaluation uses synchronized UAV radar–IMU–RGB sequences with Vicon reference poses. Entire flights and recording sessions are assigned to disjoint training, validation, and test sets to prevent overlapping windows from leaking across splits. The reported averages concern the UAV sequences. Public benchmarks differ in sensor availability: the Oxford Radar RobotCar Dataset provides polar radar intensity scans rather than per-return Doppler

labels (Barnes et al., 2020), while EuRoC MAV provides visual–inertial observations without radar. These modality differences limit direct evaluation of the full Doppler-based model on such benchmarks.

Metrics and matched controls. We compute position ATE as a trajectory RMSE after one rigid $SE(3)$ alignment, without scale correction. Translational RPE is $100\|\mathbf{t}_{\mathrm{err}}\|/s$ and rotational RPE is $\angle(\mathbf{R}_{\mathrm{err}})/s$ in degrees per meter, averaged over $s \in \{5, 10, 20\}$ m reference subtrajectories sampled every meter. Only lengths supported by a sequence are used. Errors are first averaged within each flight and then equally across flights. A run is classified as a failure if its position error exceeds 5 m for more than 1 s or the estimator stops producing poses. Ablation variants share the available sensor streams, splits, initialization, preprocessing, optimizer budget, and test-time corruptions unless the named intervention changes them. Evaluation excludes loop closures and reference-pose selection of flow samples. The full configuration uses window size $K = 4$, $L = 6$ Euler steps, and $M = 8$ flow samples.

6.2 Odometry Accuracy and 3D Real-Flight Trajectory Comparison

Table 1 compares odometry accuracy and latency across sensing modalities. OAM-FM achieves the lowest ATE and translational RPE among the evaluated methods, with a latency of 24 ms. Sensor configurations differ across groups: MRIO uses multiple radars, and scanning-radar baselines operate on scanning-radar observations (Huang et al., 2024a; Adolfsson et al., 2021). These differences are relevant when interpreting the comparison. The RIV configuration of OAM-FM uses radar, IMU, and vision; its RI counterpart in Table 4 isolates the contribution of the additional visual modality.

Table 1: Odometry accuracy and computational latency grouped by sensing modality. ATE and RPE denote absolute trajectory error and translational relative pose error, respectively. Lower values indicate better performance.

Group	Method	ATE (m) ↓	RPE (%) ↓	Latency (ms) ↓
Radar-only	CFEAR (Adolfsson et al., 2021)	2.43	3.77	9
	BFAR (Alhashimi et al., 2024)	1.21	1.89	12
	Adv-Radar (Frosi et al., 2024)	1.63	2.46	16
Radar–inertial	EKF-RIO (Michalczyk et al., 2022)	1.84	2.91	18
	MRIO (Huang et al., 2024a)	1.16	1.74	37
	FG-RIO (Michalczyk et al., 2024)	1.39	2.06	29
	CT-RIO (Burnett et al., 2025)	1.08	1.62	34
	DeRO (Do et al., 2024)	1.52	2.27	14
	Physical-RIO (Huang et al., 2024b)	1.31	1.98	21
RIV	OAM-FM (RIV)	0.82	1.18	24

Figure 5 presents eight three-dimensional trajectory comparison views arranged in two rows. The upper row shows looped trajectories and repeated path segments, while the lower row shows curved paths with altitude changes. In each view, the black dashed curve denotes the reference trajectory, and the estimated trajectories are identified by the local legend. The spatial views reveal horizontal and vertical deviations around turns, loops, and connecting segments. They complement the aggregate ATE and translational RPE in Table 1 by showing where the estimates separate from the reference; rotational accuracy is assessed separately using rotational RPE.

6.3 Ablation Studies

The four studies below separate architectural effects, geometric information, sensor availability, and inference cost. Physical-RIO motivates testing Doppler, inertial, and robust-residual components (Huang et al., 2024b); MRIO motivates separating return count from directional coverage (Huang et al., 2024a); geometry-aware flow policies motivate the pose-parameterization and sampling controls (Braun et al., 2024). All ablation results are evaluated on the same UAV split.

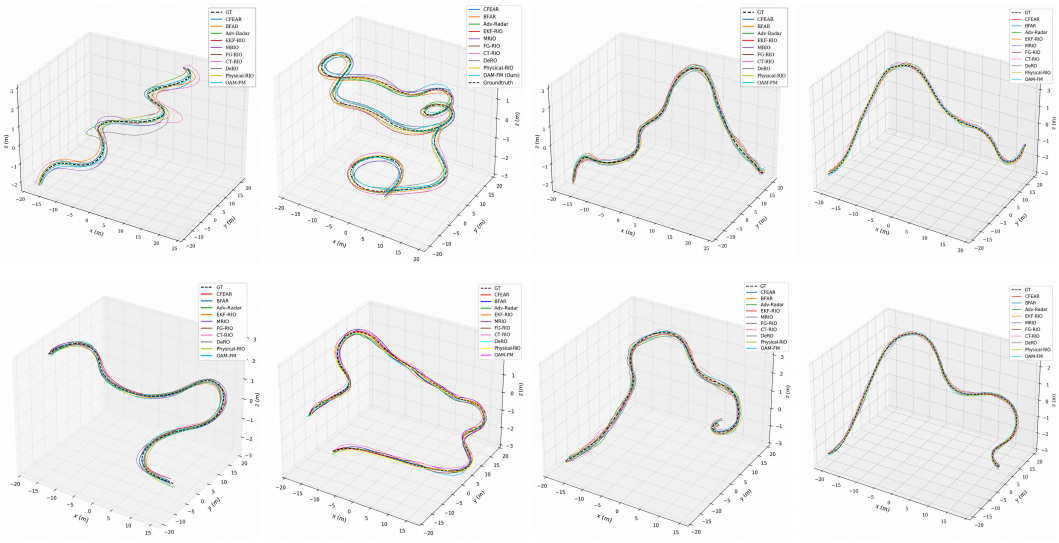

Figure 5: Three-dimensional real-flight trajectory comparisons across eight representative trials. Each view shows the trajectory estimates produced by different odometry methods under diverse flight paths and motion patterns. Black dashed curves indicate the Vicon ground truth; estimated trajectories follow the legend in each view. All position axes are in meters.

6.3.1 Contribution of the Motion Model and Physical Factors

Table 2 compares a parameter-matched deterministic regressor, Cartesian/Euler transport, the local solver alone, and physical-factor ablations. All learned variants are retrained with matched encoders and budgets. Removing a factor also removes it from the teacher and recomputes endpoints; its sensor encoder remains available. The quadratic control replaces both the Cauchy penalty and its robust weights. Appendix D specifies the remaining controls.

The full model attains an ATE of 0.82 m, compared with 1.18 m for the local solver and 1.08 m for deterministic regression. Removing the IMU factor produces the largest deterioration in both translation and rotation, increasing rotational RPE from 0.31°/m to 0.88°/m. The Cartesian/Euler, isotropic-metric, and quadratic-penalty controls also increase trajectory error, supporting the contributions of the pose parameterization, directional weighting, and robust Doppler treatment.

Table 2: Component ablation on the UAV evaluation split. RPE_t and RPE_r denote translational and rotational relative pose errors. Lower values indicate better performance.

Variant	ATE (m) ↓	RPE_t (%) ↓	RPE_r (°/m) ↓
OAM-FM (full RIV)	0.82	1.18	0.31
Local factor solver only	1.18	1.71	0.43
Deterministic pose regressor	1.08	1.55	0.39
Cartesian/Euler flow	0.95	1.38	0.48
Isotropic metric ($\mathbf{I}_6/6$)	1.04	1.49	0.39
Without Doppler residual	1.21	1.73	0.42
Without IMU factor	1.67	2.41	0.88
Quadratic Doppler penalty	0.89	1.29	0.34

6.3.2 Observability versus Generic Reweighting

Table 3 compares the full metric with isotropic, diagonal, shuffled, and learned metrics, matching trace, damping, coordinate scales, and training budget. Besides 25% random retention, we restrict returns to a 30° boresight sector and pair each restricted frame with a random subset of exactly the same size N_c. The outlier condition corrupts 20% of Doppler values. Masks and seeds are shared across variants; full protocols appear in Appendix D.

Angular restriction increases ATE more than count-matched random thinning for every metric, showing that return count alone does not explain the degradation. The full metric achieves the lowest ATE in each condition. Under angular restriction, its ATE is 1.58 m, compared with 2.24 m for isotropic weighting and 1.79 m for the learned metric. Shuffling the metric between windows further degrades accuracy, supporting the importance of correspondence between the weighting and the current radar geometry.

Table 3: Observability ablation under radar sparsity, angular restriction, and Doppler outliers. Entries report ATE in meters (↓). The two N_c columns use the same number of returns in every paired frame.

Metric	Clean	25% retained	Random N_c	Sector N_c	20% outliers
Isotropic	1.04	1.96	1.48	2.24	1.86
Diagonal	0.93	1.67	1.36	1.83	1.67
Window-shuffled	1.07	2.02	1.53	2.31	1.89
Learned positive-definite	0.90	1.64	1.31	1.79	1.63
Full radar metric	0.82	1.52	1.24	1.58	1.46

Physical rank diagnostic. The diagnostic uses the undamped spectrum of $\mathbf{A}$; damping adds no sensor information. Figure 2 distinguishes angular coverage from return count. Improved fused trajectories do not imply full instantaneous Doppler observability (Proposition 1).

6.3.3 Modality Robustness, Calibration, and Runtime

The modality controls in Appendix E isolate visual dropout, radar corruption, and calibration errors (Table 4). The full model remains usable without RGB, while its RI counterpart measures the benefit of visual information under matched radar and IMU inputs. Table 5 compares sampling budgets and covariance controls at fixed pose predictions. At $L = 6, M = 8$, the learned head reaches 93.0% empirical coverage for a nominal 95% six-dimensional Gaussian region and an NLL of 2.14. Fixed diagonal and sample covariance achieve 82.1% and 86.4% coverage, respectively, with identical ATE. These results concern local error calibration, not a multimodal posterior.

On the Orin NX, complete window processing takes 24.0 ms median and 31.6 ms at the 95th percentile, including all samples. Increasing to $L = 8$ or $M = 16$ improves ATE by only 0.01 m while increasing latency. The reported configuration consequently requires asynchronous inference if an upstream or downstream interface runs at 50 Hz; it does not meet a synchronous 20 ms neural-update budget. Timing and calibration protocols are specified in Appendix E.

7 Limitations and Broader Impact

The method depends on time synchronization, extrinsic calibration, and sufficiently short increments for its local pose chart and directional surrogate. Instantaneous single-radar Doppler leaves at least a three-dimensional twist null space; sparse angular coverage can enlarge it. Priors and complementary measurements supply the remaining constraints without guaranteeing accurate recovery under arbitrary sensor failure. Dynamic returns and gyro bias can violate the radar model. Deterministic teacher endpoints do not supervise a multimodal posterior, and teacher-discrepancy covariance cannot expose all systematic teacher errors. ODE sampling introduces a latency–accuracy trade-off. Deployment requires independently validated failure detection; predicted covariance alone is not a safety guarantee.

8 Conclusion

OAM-FM estimates UAV motion through conditional transport in local pose coordinates with $SE(3)$ outputs. Doppler directional information weights learning, complemented by robust radar, inertial, and optional visual constraints. UAV experiments support the contributions of these factors and visual dropout, while covariance and sampling controls quantify calibration and computation costs. Future work will examine more diverse environments and lower-cost inference.

AI USE STATEMENT

AI tools were used only for language polishing, grammar checking, wording refinement, and assistance with understanding code and troubleshooting simple issues. The authors take full responsibility for the final manuscript, including its scientific claims and results.

REFERENCES

Daniel Adolfsson and Maximilian Hilger. An Evaluation of CFEAR Radar Odometry. arXiv preprint arXiv:2404.01781, 2024. Radar in Robotics Workshop competition, ICRA 2024.

Daniel Adolfsson, Martin Magnusson, Anas Alhashimi, Achim J. Lilienthal, and Henrik Andreasson. CFEAR Radarodometry – Conservative Filtering for Efficient and Accurate Radar Odometry. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5462–5469, 2021.

Anas Alhashimi, Daniel Adolfsson, Henrik Andreasson, Achim Lilienthal, and Martin Magnusson. BFAR: Improving Radar Odometry Estimation Using a Bounded False Alarm Rate Detector. *Autonomous Robots*, 48(8):29, 2024.

Timothy D. Barfoot and Paul T. Furgale. Associating Uncertainty with Three-Dimensional Poses for Use in Estimation Problems. *IEEE Transactions on Robotics*, 30(3):679–693, 2014.

Dan Barnes, Matthew Gadd, Paul Murcutt, Paul Newman, and Ingmar Posner. The Oxford Radar RobotCar Dataset: A Radar Extension to the Oxford RobotCar Dataset. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6433–6438, Paris, France, 2020.

Max Braun, Noëmie Jaquier, Leonel Rozo, and Tamim Asfour. Riemannian Flow Matching Policy for Robot Motion Learning. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5144–5151, 2024.

Keenan Burnett, Angela P. Schoellig, and Timothy D. Barfoot. Continuous-Time Radar-Inertial and Lidar-Inertial Odometry Using a Gaussian Process Motion Prior. *IEEE Transactions on Robotics*, 41:1059–1076, 2025.

Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A Multimodal Dataset for Autonomous Driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11621–11631, 2020.

Carlos Campos, Richard Elvira, Juan J. Gómez Rodríguez, José M. M. Montiel, and Juan D. Tardós. ORB-SLAM3: An Accurate Open-Source Library for Visual, Visual-Inertial, and Multimap SLAM. *IEEE Transactions on Robotics*, 37(6):1874–1890, 2021.

Hoang Viet Do, Yong Hun Kim, Joo Han Lee, Min Ho Lee, and Jin Woo Song. DeRO: Dead Reckoning Based on Radar Odometry With Accelerometers Aided for Robot Localization. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 8547–8554, 2024.

Christian Forster, Luca Carlone, Frank Dellaert, and Davide Scaramuzza. On-Manifold Preintegration for Real-Time Visual-Inertial Odometry. *IEEE Transactions on Robotics*, 33(1):1–21, 2017.

Matteo Frosi, Mirko Usuelli, and Matteo Matteucci. Advancements in Radar Odometry. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 11024–11031, 2024.

Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are We Ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3354–3361, 2012.

Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J. Brostow. Digging Into Self-Supervised Monocular Depth Estimation. In *IEEE/CVF International Conference on Computer Vision*, pp. 3828–3838, 2019.

Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3D Packing for Self-Supervised Monocular Depth Estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2485–2494, 2020.

Dongjiao He, Wei Xu, Nan Chen, Fanze Kong, Chongjian Yuan, and Fu Zhang. Point-LIO: Robust High-Bandwidth Light Detection and Ranging Inertial Odometry. *Advanced Intelligent Systems*, 5(7):2200459, 2023.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Jui-Te Huang, Ruoyang Xu, Akshay Hinduja, and Michael Kaess. Multi-Radar Inertial Odometry for 3D State Estimation using mmWave Imaging Radar. In *IEEE International Conference on Robotics and Automation*, pp. 12006–12012, 2024a.

Qiucan Huang, Yuchen Liang, Zhijian Qiao, Shaojie Shen, and Huan Yin. Less is More: Physical-Enhanced Radar-Inertial Odometry. In *IEEE International Conference on Robotics and Automation*, pp. 15966–15972, 2024b.

Alex Kendall and Yarin Gal. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow Matching for Generative Modeling. In *International Conference on Learning Representations*, 2023.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow. In *International Conference on Learning Representations*, 2023.

Jan Michalczyk, Roland Jung, and Stephan Weiss. Tightly-Coupled EKF-Based Radar-Inertial Odometry. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 12336–12343, 2022.

Jan Michalczyk, Julius Quell, Florian Steidle, Marcus G. Müller, and Stephan Weiss. Tightly-Coupled Factor Graph Formulation For Radar-Inertial Odometry. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 3364–3370, 2024.

Morten Nissov, Nikhil Khedekar, and Kostas Alexis. Degradation Resilient LiDAR-Radar-Inertial Odometry. In *IEEE International Conference on Robotics and Automation*, pp. 8587–8594, 2024.

Tong Qin, Peiliang Li, and Shaojie Shen. VINS-Mono: A Robust and Versatile Monocular Visual-Inertial State Estimator. *IEEE Transactions on Robotics*, 34(4):1004–1020, 2018.

Tong Qin, Jie Pan, Shaozu Cao, and Shaojie Shen. A General Optimization-Based Framework for Local Odometry Estimation with Multiple Sensors. arXiv preprint arXiv:1901.03638, 2019.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision Transformers for Dense Prediction. In *IEEE/CVF International Conference on Computer Vision*, pp. 12179–12188, 2021.

Tixiao Shan, Brendan Englot, Drew Meyers, Wei Wang, Carlo Ratti, and Daniela Rus. LIO-SAM: Tightly-coupled Lidar Inertial Odometry via Smoothing and Mapping. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5135–5142, 2020.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-Based Generative Modeling through Stochastic Differential Equations. In *International Conference on Learning Representations*, 2021.

Zachary Teed and Jia Deng. DROID-SLAM: Deep Visual SLAM for Monocular, Stereo, and RGB-D Cameras. In *Advances in Neural Information Processing Systems*, volume 34, pp. 16558–16569, 2021.

Jamie Watson, Oisin Mac Aodha, Victor Prisacariu, Gabriel Brostow, and Michael Firman. The Temporal Opportunist: Self-Supervised Multi-Frame Monocular Depth. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1164–1174, 2021.

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.

Mengtan Zhang, Zizhan Guo, Hongbo Zhao, Yi Feng, Zuyi Xiong, Yue Wang, Shaoyi Du, Hanli Wang, and Rui Fan. Discriminately Treating Motion Components Evolves Joint Depth and Ego-Motion Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026. doi: 10.1109/TPAMI.2026.3732069. Early access.

Tinghui Zhou, Matthew Brown, Noah Snavely, and David G. Lowe. Unsupervised Learning of Depth and Ego-Motion from Video. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1851–1858, 2017.

A Complete Sensor Factors

Inertial and visual factors. IMU preintegration requires a state $\mathbf{x}_k = (\mathbf{R}_k, \mathbf{p}_k, \mathbf{v}_k, \mathbf{b}_{g,k}, \mathbf{b}_{a,k})$, with world-frame position and velocity, body-to-world rotation, and gyroscope/accelerometer biases. Accelerometers measure specific force. Bias-corrected preintegrated measurements $(\Delta\mathbf{R}_k, \Delta\mathbf{v}_k, \Delta\mathbf{p}_k)$ constrain adjacent states (Forster et al., 2017) through

$$\mathbf{r}_{I,k} = \begin{bmatrix} \mathrm{Log}_{SO(3)}(\Delta\mathbf{R}_k^\top \mathbf{R}_k^\top \mathbf{R}_{k+1}) \\ \mathbf{R}_k^\top(\mathbf{v}_{k+1} - \mathbf{v}_k - \mathbf{g}\Delta t_k) - \Delta\mathbf{v}_k \\ \mathbf{R}_k^\top(\mathbf{p}_{k+1} - \mathbf{p}_k - \mathbf{v}_k\Delta t_k - \frac{1}{2}\mathbf{g}\Delta t_k^2) - \Delta\mathbf{p}_k \\ \mathbf{b}_{g,k+1} - \mathbf{b}_{g,k} \\ \mathbf{b}_{a,k+1} - \mathbf{b}_{a,k} \end{bmatrix}, \qquad \mathcal{L}_{\mathrm{IMU}} = \frac{1}{2}\sum_k \|\mathbf{r}_{I,k}\|^2_{\mathbf{P}_{I,k}^{-1}}. \tag{11}$$

Here $\mathbf{g}$ is world-frame gravity and $\mathbf{P}_{I,k}$ is the covariance of the stacked residual in the displayed order, including bias random walks. Preintegration is recomputed or first-order corrected at the current bias linearization point. A pose-only inertial factor is valid only after conditioning on, or consistently marginalizing, the other state variables; IMU samples alone do not supply an independent six-dimensional relative-pose measurement.

For an optional valid visual track j, let $\mathbf{u}_{k,j}$ be its pixel coordinate, $\eta_{k,j} > 0$ its estimated inverse depth, $\mathbf{b}(\mathbf{u}) = \mathbf{K}_C^{-1}[\mathbf{u}^\top, 1]^\top$ its calibrated ray, and π_C the camera projection. The residual is

$$\mathbf{r}_{V,kj} = \mathbf{u}_{k+1,j} - \pi_C\left(\mathcal{T}_{BC}^{-1}\mathcal{T}_k^{-1}\mathcal{T}_{BC} \cdot \frac{\mathbf{b}(\mathbf{u}_{k,j})}{\eta_{k,j}}\right), \qquad \mathcal{L}_V = \sum_{k,j\in\mathcal{V}} \rho_V(\|\mathbf{r}_{V,kj}\|^2_{\mathbf{P}_{V,kj}^{-1}}). \tag{12}$$

The dot denotes a rigid transform acting on a 3D point, $\mathcal{V}$ includes only valid tracks with positive projected depth, and ρ_V is a robust penalty. The term is omitted when images are unavailable. Depth is inferred from sensor consistency without depth labels.

At a radar timestamp t_i, the body-frame velocity is $\mathbf{R}(t_i)^\top\mathbf{v}(t_i)$ and the angular velocity is obtained from the gyro with its current bias correction. Using these quantities in Eq. (1), the robust radar term is

$$\mathcal{L}_{\mathrm{Dop}} = \sum_i \rho_c(r_i), \qquad \rho_c(r) = \frac{c^2}{2}\log(1 + r^2/c^2), \qquad w_i = \frac{\rho_c'(r_i)}{r_i} = \frac{1}{1 + r_i^2/c^2}, \tag{13}$$

where r_i is the whitened residual in Eq. (4), $c > 0$ is dimensionless, and $w_i = 1$ at $r_i = 0$. The effective Doppler noise scale must include gyro uncertainty when gyro measurements are substituted into the lever-arm term. Shared IMU noise then induces cross-correlations; treating the factors as independent is an approximation, not an exact probabilistic likelihood.

B Theoretical Results

Proposition 1 (instantaneous single-radar rank). For fixed extrinsics, unit lines of sight, positive noise scales, and strictly positive weights on retained returns, the undamped information matrix in Eq. (5) satisfies

$$\mathrm{rank}(\mathbf{A}) = \mathrm{rank}(\mathbf{H}^\zeta) = \mathrm{rank}(\mathbf{N}) \le 3. \tag{14}$$

Proof. Write $\mathbf{H}^\zeta = \mathbf{N}\mathbf{B}$ with $\mathbf{B} = [\mathbf{I}_3, -[\boldsymbol{\ell}_{BR}]_\times]$. Since $\mathbf{B}[\mathbf{I}_3, \mathbf{0}]^\top = \mathbf{I}_3$, multiplication by $\mathbf{B}$ preserves the rank of $\mathbf{N}$. Whitening and invertible coordinate scaling preserve rank, as does forming the weighted Gram matrix with positive weights. In particular, every twist $([\boldsymbol{\ell}_{BR}]_\times\delta\boldsymbol{\omega}, \delta\boldsymbol{\omega})$ lies in the Doppler null space. Three linearly independent lines of sight constrain radar-origin velocity, not six independent body-twist components. If the lines of sight span a plane through the origin, the rank is at most two. This claim concerns an instantaneous radar, not the full radar–IMU trajectory system.

Proposition 2 (conditional linear-Gaussian covariance reduction). Let a local perturbation have a Gaussian prior with covariance $\mathbf{P}_0 \succ 0$. For a linear measurement with Jacobian $\mathbf{J}$ and correctly specified Gaussian noise independent of the prior, with covariance $\mathbf{R}_n \succ 0$, the posterior covariance is

$$\mathbf{P}_+ = (\mathbf{P}_0^{-1} + \mathbf{J}^\top\mathbf{R}_n^{-1}\mathbf{J})^{-1} \preceq \mathbf{P}_0. \tag{15}$$

Proof. The middle factor of $\mathbf{P}_+ = \mathbf{P}_0^{1/2}(\mathbf{I} + \mathbf{P}_0^{1/2}\mathbf{J}^\top\mathbf{R}_n^{-1}\mathbf{J}\mathbf{P}_0^{1/2})^{-1}\mathbf{P}_0^{1/2}$ is bounded above by $\mathbf{I}$ in the positive-semidefinite ordering. This establishes a model-based covariance statement. It does not imply that every realized pose error decreases when a robust loss is added to neural training. Biased returns, data-dependent weights, correlated reused measurements, nonlinear linearization errors, and teacher error fall outside the stated guarantee.

C Sample-Covariance Diagnostic

For the sample-covariance control, use

$$\begin{aligned} \boldsymbol{\delta}^{(m)} &= \mathrm{Log}(\hat{\mathcal{T}}^{-1}\hat{\mathcal{T}}^{(m)}), \qquad \bar{\boldsymbol{\delta}} = \frac{1}{M}\sum_m \boldsymbol{\delta}^{(m)}, \\ \hat{\boldsymbol{\Sigma}}_{\mathrm{samp}} &= \frac{1}{M-1}\sum_m (\boldsymbol{\delta}^{(m)} - \bar{\boldsymbol{\delta}})(\boldsymbol{\delta}^{(m)} - \bar{\boldsymbol{\delta}})^\top, \qquad M > 1. \end{aligned} \tag{16}$$

This diagnostic and the learned covariance are distinct estimates. A common normalized diagonal floor is added for Gaussian scoring. For $M = 1$, only the learned head supplies covariance. Covariance-frame conversions are specified in Appendix G.

D Detailed Ablation Protocols

Table 2 evaluates the contributions of the flow formulation, pose parameterization, and physical constraints. The deterministic control replaces the stochastic flow with a parameter-matched pose regressor using the same encoders and local-factor targets. The Cartesian/Euler control transports translation and Euler-angle coordinates and uses their explicit pose conversion in every residual. The solver-only control returns the local factor solution directly, testing the contribution of the learned stage beyond its training target generator. In the Doppler- and IMU-factor ablations, we remove the named residual from both the loss and the endpoint solve and recompute training targets; the corresponding sensor encoder remains available. The quadratic control replaces the Cauchy Doppler penalty and its associated robust weights with least squares and unit weights. All learned variants are retrained.

Table 3 evaluates whether the current radar geometry provides information beyond a generic weighting scheme. We compare the full metric with an isotropic metric, its normalized diagonal, a metric shuffled between windows, and a learned positive-definite metric. Metric trace, pose-coordinate normalization, damping, and training budget are matched across variants. The learned metric is formed as $\mathbf{B}_\eta\mathbf{B}_\eta^\top + \lambda\mathbf{I}_6$ from the same sensor features, with matched total parameter count; shuffled metrics preserve the marginal spectrum but break correspondence with the current window. Each replacement is applied wherever the implementation uses the observability metric.

Random thinning retains 25% of the returns. A separate angular restriction retains returns within a 30° azimuth sector centered on the radar boresight. For each restricted frame containing N_c returns, a paired control samples exactly N_c returns from the unrestricted frame, making the random-N_c and sector-N_c columns comparable at equal point count. The outlier condition replaces 20% of radial velocities with draws from the sensor's valid Doppler interval. Identical masks and corruption seeds are reused across variants.

E Extended Experimental Results

The supplementary setup view in Fig. 6 documents the physical evaluation environment. Panel (a) shows the real indoor flight arena with the UAV and obstacles, while panel (b) shows the ceiling-mounted Vicon cameras used for infrared marker tracking. These cameras provide the synchronized six-degree-of-freedom reference poses described in Sec. 6; the reference system is used only for evaluation and is not supplied to OAM-FM during training or inference.



Figure 6: Supplementary views of the real-flight evaluation setup. (a) Indoor flight arena with the UAV and obstacle layout used for the real-flight experiments. (b) Close-up view of ceiling-mounted Vicon cameras used for infrared marker tracking and synchronized six-degree-of-freedom reference-pose acquisition. Vicon poses are used only as evaluation references and are not provided to the estimator during training or inference.

E.0.1 Sensor Complementarity and Degradation Robustness

Table 4 separates the benefit of optional vision from robustness to missing inputs. We train RIV, RI, and VI variants independently; RI removes the visual encoder and reprojection term, while VI removes the radar encoder, Doppler factors, and radar metric. During training of the full RIV model, we mask the complete visual window with probability 0.2 and use an explicit missing-input indicator. An additional RIV variant is trained without this masking. The same RIV checkpoint is used for all test corruptions without condition-specific fine-tuning.

Visual degradation is introduced using a 15-pixel linear motion-blur kernel, intensity scaling by 0.1, or a 2 s camera outage at the midpoint of each flight. Radar tests discard 50% or 75% of returns or corrupt 20% of Doppler values as above. Calibration stress tests add a 10 ms radar timestamp offset or a 2° radar-to-IMU yaw error at evaluation. Each row changes only the named condition. RI results are unchanged under image corruption, and VI results are unchanged under radar corruption, because the altered stream is absent from those models.

RIV improves clean-sequence ATE over RI, from 0.94 m to 0.82 m, while retaining substantially lower error than VI under visual degradation. RI achieves lower error under motion blur and camera outage, while RIV remains more accurate under low image intensity. VI is unaffected by radar corruption. Visual dropout slightly increases clean-sequence ATE from 0.80 m to 0.82 m but reduces outage ATE from 1.44 m to 1.03 m. These results reveal both sensor complementarity and the trade-off between nominal accuracy and resilience to missing visual observations.

E.0.2 Uncertainty Calibration and Onboard Sampling Budget

Table 5 varies ODE steps L and flow samples M on the Orin NX platform. The point estimate is $\mathrm{Exp}(\bar{\boldsymbol{\xi}})$, without ground-truth selection of samples. The sampling sweep uses the same trained checkpoint; the covariance controls hold its pose samples fixed and change only the uncertainty estimate. We compare the learned covariance head with a fixed diagonal covariance fitted to training sensor-consistency residuals and the empirical sample covariance. This separates calibration from improvements in the estimated mean (Kendall & Gal, 2017).

For evaluation, $\mathcal{T}^{\mathrm{gt}}$ is the reference increment over the same interval as $\hat{\mathcal{T}}$. We use the normalized right-local error $\mathbf{e} = \mathbf{S}^{-1}\,\mathrm{Log}(\hat{\mathcal{T}}^{-1}\mathcal{T}^{\mathrm{gt}})$ and covariance $\mathbf{P} = \mathbf{S}^{-1}\hat{\boldsymbol{\Sigma}}\mathbf{S}^{-\top}$, with the same fixed scales as training. Gaussian NLL uses Eq. (9) with $\mathbf{e}$ and $\mathbf{P}$. Empirical coverage C_{95} is the fraction of

Table 4: Sensor-modality ablation under visual degradation, radar corruption, and calibration errors. Entries report ATE in meters (↓). R, I, and V denote radar, IMU, and vision. Each corruption is applied separately.

Condition	RIV (full)	RI	VI	RIV, no dropout
Clean	0.82	0.94	1.24	0.80
Visual motion blur	0.96	0.94	2.08	1.31
Low image intensity	0.91	0.94	2.36	1.49
Camera outage (2 s)	1.03	0.94	2.71	1.44
50% radar returns removed	1.27	1.43	1.24	1.40
75% radar returns removed	1.52	1.82	1.24	1.79
20% Doppler outliers	1.46	1.71	1.24	1.68
Radar timestamp offset (10 ms)	1.12	1.29	1.24	1.30
Radar extrinsic yaw error (2°)	1.19	1.37	1.24	1.39

increments satisfying $\mathbf{e}^\top \mathbf{P}^{-1}\mathbf{e} \le \chi^2_{6,0.95}$. The 95% reference applies to a correctly specified zero-mean six-dimensional Gaussian error; coverage alone does not establish distributional calibration. The learned head supplies uncertainty for $M = 1$, where sample covariance is undefined. All covariance controls use a common diagonal floor in normalized coordinates, applied once. NLL evaluates a local Gaussian error model and does not establish multimodality. Vicon errors are used only for evaluation; covariance training uses teacher discrepancies as specified in Eq. (9).

Latency is measured per synchronized window, including preprocessing, encoding, all flow samples, covariance computation, and pose publication, after warm-up and explicit device synchronization. We report the median and 95th percentile using batch size one. The full $L = 6, M = 8$ configuration requires 24.0 ms median latency and 31.6 ms at the 95th percentile. Increasing either L to 8 or M to 16 reduces ATE by only 0.01 m while increasing latency, indicating diminishing returns from additional sampling. The full configuration therefore provides a practical accuracy–latency trade-off, although its latency exceeds the 20 ms budget required for a synchronous 50 Hz neural update.

At fixed pose estimates, the learned covariance head achieves 93.0% coverage and an NLL of 2.14, compared with 82.1% and 3.46 for fixed diagonal covariance and 86.4% and 3.04 for sample covariance. The identical ATE across these controls confirms that the improvement concerns uncertainty calibration rather than changes to the estimated trajectory.

Table 5: Uncertainty calibration and sampling cost on the Orin NX. L and M denote ODE steps and flow samples. C_{95} targets 95% coverage; lower ATE, NLL, and latency are preferred.

Covariance	L	M	ATE (m)	C_{95} (%)	NLL	Latency (ms)	
						Median	95th pct.
Learned head	2	8	0.99	89.1	2.71	16.1	21.3
Learned head	4	8	0.87	92.0	2.29	20.1	26.5
Learned head (full)	6	8	0.82	93.0	2.14	24.0	31.6
Learned head	8	8	0.81	93.4	2.12	28.0	36.8
Learned head	6	1	0.87	91.2	2.38	17.4	22.9
Learned head	6	16	0.81	94.2	2.10	33.6	44.1
Fixed diagonal	6	8	0.82	82.1	3.46	23.7	31.2
Sample covariance	6	8	0.82	86.4	3.04	23.8	31.3

F Additional Implementation Details

Training follows three stages: physics and IMU warm-up, observability-aware flow training, and covariance-head fitting. The local factor solver supplies sensor-based motion endpoints with its initial pose gauge fixed and velocity/bias priors propagated from preceding states. Teacher endpoints, robust weights, and the metric are detached during each flow update. Physical losses retain the nuisance states required by Eqs. (11) and (12). In the uncertainty stage, the pose predictor is frozen and the head is fitted to right-local teacher discrepancies without reference-pose supervision.

The default inference configuration uses window size $K = 4$, $L = 6$ Euler steps with $\Delta\tau = 1/6$, and $M = 8$ flow samples. The point estimate is obtained from the mean Lie-algebra sample through the exponential map. Calibration experiments express pose errors and covariance matrices in the same tangent frame, apply common coordinate scales, and use a shared diagonal covariance regularizer. For $M = 1$, uncertainty is supplied by the learned covariance head. Computational timing uses batch size one and includes all flow samples within the measured window.

G Coordinate Conventions and Derivations

Local pose chart. For $a = \|\boldsymbol{\phi}\|$, the translation block in Eq. (2) uses

$$\mathbf{J}_l^{SO(3)}(\boldsymbol{\phi}) = \mathbf{I}_3 + \frac{1-\cos a}{a^2}[\boldsymbol{\phi}]_\times + \frac{a-\sin a}{a^3}[\boldsymbol{\phi}]_\times^2, \tag{17}$$

with continuous limits at $a = 0$. Thus Euclidean translation averaging and Lie-coordinate averaging coincide only in special cases or to an appropriate small-angle approximation.

Damped normal equation and pose updates. With $\mathbf{r}_0 = \mathbf{D}^{-1}(\mathbf{d} - \mathbf{h}(\mathbf{z}_\star))$, linearization gives $\mathbf{r}(\mathbf{z}_\star + \delta\mathbf{z}) \simeq \mathbf{r}_0 - \mathbf{H}\delta\mathbf{z}$. Holding $\mathbf{W}$ fixed, minimizing $\frac{1}{2}\|\mathbf{r}_0 - \mathbf{H}\delta\mathbf{z}\|_{\mathbf{W}}^2 + \frac{\lambda}{2}\|\delta\mathbf{z}\|^2$ yields

$$(\mathbf{A} + \lambda\mathbf{I}_6)\delta\mathbf{z} = \mathbf{H}^\top\mathbf{W}\mathbf{r}_0, \qquad \mathcal{T}_{\text{new}} = \operatorname{Exp}(\boldsymbol{\xi}_\star + \mathbf{S}\delta\mathbf{z}). \tag{18}$$

The increment here is additive in log coordinates. It is not directly a left or right group perturbation. A right perturbation satisfies

$$\begin{aligned} \operatorname{Exp}(\boldsymbol{\xi} + \Delta\boldsymbol{\xi}) &\simeq \operatorname{Exp}(\boldsymbol{\xi})\operatorname{Exp}(\mathbf{J}_r^{SE(3)}(\boldsymbol{\xi})\Delta\boldsymbol{\xi}), \\ \mathbf{J}_r^{SE(3)}(\boldsymbol{\xi}) &= \sum_{n=0}^{\infty}\frac{(-\operatorname{ad}_{\boldsymbol{\xi}})^n}{(n+1)!}, \quad \operatorname{ad}_{\boldsymbol{\xi}} = \begin{bmatrix}[\boldsymbol{\phi}]_\times & [\boldsymbol{\rho}]_\times \\ \mathbf{0} & [\boldsymbol{\phi}]_\times\end{bmatrix}. \end{aligned} \tag{19}$$

This Jacobian conversion is required before applying a log-coordinate increment as a group perturbation.

Covariance frames and mean approximation. For $\mathcal{T} = \hat{\mathcal{T}}\operatorname{Exp}(\boldsymbol{\delta}_R) = \operatorname{Exp}(\boldsymbol{\delta}_L)\hat{\mathcal{T}}$, the two local errors obey $\boldsymbol{\delta}_L = \operatorname{Ad}_{\hat{\mathcal{T}}}\boldsymbol{\delta}_R$ within the common logarithm chart, with (Barfoot & Furgale, 2014)

$$\operatorname{Ad}_{\mathcal{T}} = \begin{bmatrix}\mathbf{R} & [\mathbf{t}]_\times\mathbf{R} \\ \mathbf{0} & \mathbf{R}\end{bmatrix}, \qquad \boldsymbol{\Sigma}_L = \operatorname{Ad}_{\hat{\mathcal{T}}}\boldsymbol{\Sigma}_R\operatorname{Ad}_{\hat{\mathcal{T}}}^\top. \tag{20}$$

Conversely, the covariance $\mathbf{C}_\xi$ of additive log-coordinate samples maps only to first order to $\boldsymbol{\Sigma}_R \simeq \mathbf{J}_r^{SE(3)}(\bar{\boldsymbol{\xi}})\mathbf{C}_\xi\mathbf{J}_r^{SE(3)}(\bar{\boldsymbol{\xi}})^\top$. Equation (16) computes the local sample deviations directly. The log-coordinate mean need not make $\bar{\boldsymbol{\delta}} = \mathbf{0}$, so using a zero-mean Gaussian centered on that pose is a local approximation. The learned head estimates the second moment of teacher discrepancies under its zero-mean model and may absorb their bias; neither its output nor sample spread is automatically a Bayesian pose posterior.

What deterministic endpoint supervision implies. For fixed observations y and a deterministic teacher $\mathbf{z}_1(y)$, the path in Eq. (7) has

$$p_\tau(\mathbf{z} \mid y) = \mathcal{N}\big(\tau\mathbf{z}_1(y), (1-\tau)^2\mathbf{I}_6\big), \qquad \mathbf{u}_\tau(\mathbf{z} \mid y) = \frac{\mathbf{z}_1(y) - \mathbf{z}}{1-\tau}, \quad \tau < 1. \tag{21}$$

Its limit is a point mass, not a multimodal density. The exact target field is singular at $\tau = 1$; an ODE with an integrably bounded spatial Lipschitz constant cannot collapse a full-dimensional density exactly in finite time. Training uses interior flow times and inference a finite discretization. Variation across observations sharing compressed features, network approximation error, and numerical integration can produce sample spread, but do not identify the true conditional sensor posterior. The covariance head is consequently evaluated as a separate error predictor.

Scope of directional weighting. Positive damping gives $\bar{\mathbf{G}} \succ 0$ even in physical null directions. Its quadratic loss remains sensitive to all coordinates, with relatively greater weight in directions supported by Doppler geometry. If the metric is fixed given the conditioning, the unconstrained population minimizer of weighted squared flow error is the same conditional mean velocity as for unweighted regression. If the metric also varies across windows sharing a condition, the minimizer is instead $\mathbb{E}[\bar{\mathbf{G}} \mid \mathbf{z}, \tau, \mathbf{c}]^{-1}\mathbb{E}[\bar{\mathbf{G}}\mathbf{u} \mid \mathbf{z}, \tau, \mathbf{c}]$. Directional weighting therefore specifies a learning objective; it does not alone prove posterior correctness or improved realized error.