WHEN CROSS-MODAL EVIDENCE HELPS: MODALITY-LOCAL VERSUS CROSS-MODAL REPRESENTATIONS FOR RELIABILITY INFERENCE IN STATE ESTIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning-based state estimation increasingly relies on multimodal sensing under changing sensing conditions, yet it remains unclear when reliability should be inferred from a sensor's own history and when cross-modal evidence is necessary. We introduce a Local–Cross framework that separates modality-local temporal representations from cross-modal relational representations while keeping the prediction target and model capacity matched. Using UAV optical flow, IMU, and GPS state estimation, we study two estimator-facing reliability quantities: covariance for stochastic uncertainty and consistency for compatibility with the nominal sensing model. We find that covariance prediction consistently favors modality-local representations, whereas consistency prediction consistently benefits from cross-modal representations, both when each degradation occurs alone and when the two occur simultaneously. When integrated into a classical ESKF, modality-local covariance prediction combined with cross-modal consistency prediction improves state estimation under degraded sensing while preserving nominal behavior. These results suggest that cross-modal evidence should be used selectively according to the reliability quantity being inferred rather than by default.

1 INTRODUCTION

Reliable state estimation is fundamental to robotic autonomy, supporting downstream planning, control, and decision-making (Talbot et al., 2025; Barfoot, 2017). Two recurring aspects of estimator reliability are stochastic uncertainty and measurement consistency. Covariance characterizes stochastic uncertainty in process or measurement models (Kalman, 1960), whereas measurement consistency reflects whether an available measurement remains compatible with the nominal measurement model. Under changing sensing conditions, learning-based methods have been used to predict such quantities and adapt state estimators accordingly.

A basic question is what information should be used to infer them. Modality-local temporal information captures changes within an individual sensing stream, whereas cross-modal information captures relationships among sensing modalities. Covariance may be recoverable from a sensor's own history, while consistency may require comparisons with other sensors. Existing methods either compress different degradation effects into one prediction, use separate pathways for different reliability quantities, or introduce cross-modal inputs for consistency prediction (Fig. 1(a)–(b)). The use of cross-modal inputs, however, does not show that they improve on modality-local inputs. Prior work has not compared the two representations for the same target with matched model capacity.

We study this question in unmanned aerial vehicle (UAV) state estimation with optical flow (OF), an inertial measurement unit (IMU), and the Global Positioning System (GPS), and explicitly separate modality-local and cross-modal representations (Fig. 1(c)). For each reliability target, a *Local* predictor uses only modality-local temporal representations, whereas a *Cross-Modal* predictor also uses relational representations. The two variants are compared at matched model capacities under stochastic covariance variation (SCV), sensing consistency disruption (SCD), and their simultaneous occurrence as compound degradation (CD). Their covariance and consistency predictions are

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.

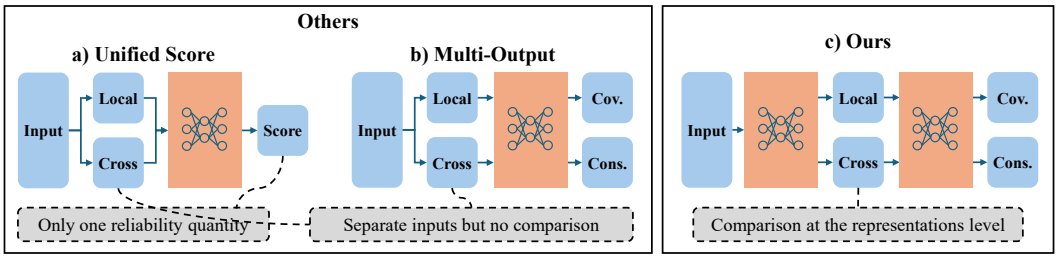


Figure 1: Representative design patterns for learned sensing-degradation prediction.

then used in the same non-learned error-state Kalman filter (ESKF). Predicted covariance specifies process and measurement uncertainty; predicted consistency modulates measurement covariance.

The preferred representation differs by task. Local representations give lower covariance error under isolated SCV and compound degradation. Cross-Modal representations give better consistency discrimination under isolated SCD and compound degradation at both tested capacity tiers. In the ESKF, Local covariance prediction combined with Cross-Modal consistency prediction also improves estimation under SCD and compound degradation without degrading nominal accuracy.

Our contributions are as follows:

1. A Local–Cross framework that compares modality-local temporal and cross-modal relational representations for the same prediction target at matched capacity.
2. Evidence that the value of cross-modal information depends on the reliability target. Local representations improve covariance prediction under both isolated SCV and compound degradation, whereas Cross-Modal representations improve consistency discrimination under both isolated SCD and compound degradation across the two matched-capacity tiers.
3. A downstream evaluation showing that continuous Cross-Modal consistency predictions improve state estimation under isolated SCD and compound degradation.

2 Related Work

Learning Reliability Quantities for State Estimation. Learning-based sensor fusion increasingly augments classical multisensor estimation under changing sensing conditions (Fayyad et al., 2020; Zhuang et al., 2023). One line of work learns data-conditioned process or measurement uncertainty, including covariance and Kalman-filter noise parameters (Liu et al., 2018; Wu et al., 2021; Yan et al., 2021; Or & Klein, 2022; Brossard et al., 2020; Liu et al., 2020). Another addresses measurements that violate the nominal sensing model through learned non-line-of-sight (NLOS) detection and error prediction, measurement-validity estimation, or cross-modal fault detection (Zhang et al., 2023a; Lin et al., 2025; Cao et al., 2026). These studies predict different reliability quantities, but do not compare the information needed to infer the same target.

Information Sources for Reliability Inference. Reliability can be inferred from the target stream itself or from its relation to other modalities. Multimodal learning provides representation, alignment, and fusion mechanisms for the latter (Ngiam et al., 2011; Baltrušaitis et al., 2019; Guo et al., 2019). Learned selection and uncertainty mechanisms can also adapt modality contributions when inputs degrade (Chen et al., 2019; Cho et al., 2025), and cross-modal relationships can reveal sensor faults. In most prior comparisons, however, the available modalities, fusion architecture, and learning objective change together. The contribution of cross-modal information alone is therefore unclear.

Multimodal Uncertainty Estimation. Deep uncertainty estimation commonly distinguishes data-dependent aleatoric uncertainty from model-dependent epistemic uncertainty, using Bayesian approximations, heteroscedastic likelihoods, or evidential objectives (Kendall & Gal, 2017; Amini et al., 2020). Multimodal extensions estimate modality-specific and global uncertainty to weight, discount, or fuse heterogeneous evidence when modalities are noisy or conflicting (Ma et al., 2021; Zhang et al., 2023b; Gao et al., 2024; Bezirganyan et al., 2025). Fusion behavior also changes un-

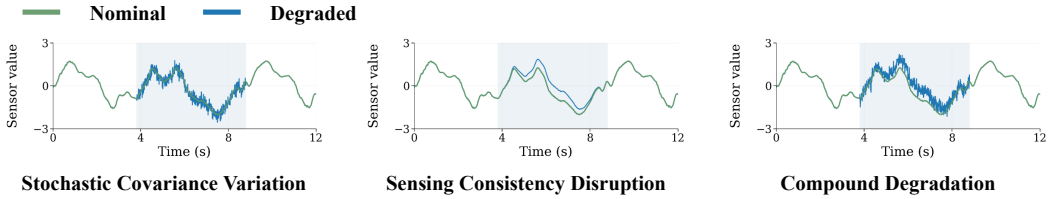


Figure 2: Illustration of nominal sensing, stochastic covariance variation, sensing consistency disruption, and compound degradation.

der missing modalities, and no single strategy is uniformly preferable across datasets (Ma et al., 2022). That literature mainly uses uncertainty to improve a downstream prediction. Here, covariance and measurement consistency are the prediction targets. We vary only whether the predictor receives modality-local or cross-modal evidence while holding the target, data, optimization, and model capacity fixed.

3 Analysis of Sensing Degradation

Sensing degradation can affect multimodal state estimation through both the stochastic behavior of individual sensing modalities and the relationships among modalities. We distinguish three recurring cases: stochastic covariance variation (SCV), sensing consistency disruption (SCD), and their simultaneous occurrence as compound degradation (CD), as illustrated in Fig. 2.

Stochastic Covariance Variation (SCV). Sensor observations contain stochastic errors whose dispersion can change with sensing conditions even when the nominal process or measurement model remains applicable (Mehra, 1970). Such changes alter the uncertainty represented by the process and measurement covariances, but do not by themselves imply that an observation has become incompatible with the nominal sensing model. In our OF/IMU/GPS setting, SCV affects the covariance roles associated with the IMU accelerometer and gyroscope, OF velocity, GPS position, and GPS velocity.

Sensing Consistency Disruption (SCD). Multiple sensing modalities observe different aspects of the same system motion and are therefore coupled through the underlying state and nominal measurement models. A measurement can remain locally smooth or statistically plausible while becoming inconsistent with the motion supported by other sensing modalities. We refer to such a violation of the nominal sensing relationship as SCD. Examples include temporally bounded freezing, coherent bias or drift, coherent but incorrect measurements, and outlier corruption.

Compound Degradation (CD). Changes in noise statistics and incompatible or outlying measurements can also occur simultaneously (Li et al., 2016; Luo & Wang, 2017; Zhu et al., 2022). In this case, one or more stochastic uncertainty changes while one or more measurements also become inconsistent with the nominal sensing relationship. The two effects therefore need to be distinguished: increased stochastic uncertainty does not necessarily imply inconsistency, and inconsistency cannot in general be represented solely by increasing the nominal covariance.

We predict two quantities: covariance for stochastic uncertainty and measurement consistency for compatibility with the nominal sensing model. Here, *measurement consistency* refers to the compatibility score $a_{t,j}$; *estimator consistency* in Sec. 5.4 refers to covariance calibration. We compare modality-local and cross-modal representations for both targets.

4 Method

Throughout this section, bold lowercase symbols denote vectors, bold uppercase symbols denote matrices or feature arrays, and ordinary italic symbols denote scalars; the superscript GT denotes ground truth. A *prediction role* is one scalar output associated with a particular sensor quantity, such as IMU accelerometer process noise or GPS position measurement noise. We use $C_{t,j}$ as a generic

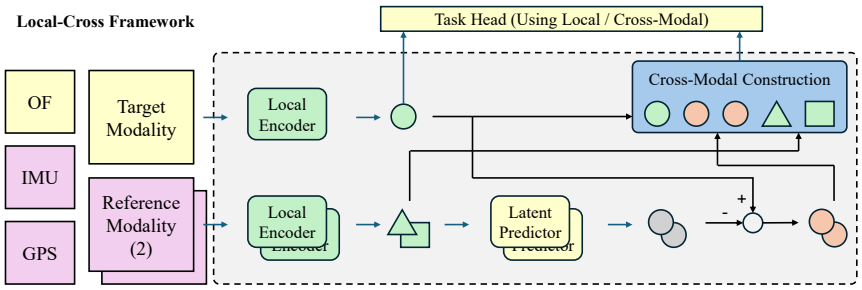


Figure 3: Overview of the Local–Cross reliability prediction framework. One target–reference modality pair is shown for illustration.

covariance target for role j; $Q_{t,j}$ and $R_{t,j}$ distinguish process- and measurement-covariance terms, whereas $\mathbf{Q}_t$ and $\mathbf{R}_t$ denote the assembled covariance matrices.

4.1 Framework Overview

Figure 3 shows the Local–Cross framework. Modality-specific causal encoders map OF, IMU, and GPS histories to local representations. Directed latent predictors combine them into cross-modal representations. Either representation can be passed to the same task head, allowing a matched comparison. Both variants are evaluated with the same non-learned ESKF (Solà, 2017); Sec. 5.1 lists the estimator configurations.

4.2 Modality-Specific Local Encoder

Constructing Local Representations. Each modality $m \in \{\mathrm{OF}, \mathrm{IMU}, \mathrm{GPS}\}$ is processed by an independent causal temporal encoder with the same architecture but separate parameters. Given a causal history $\mathbf{X}_t^m$ of length $W = 32$, the encoder follows a temporal convolutional network design based on causal dilated convolutions (Yu & Koltun, 2016; Bai et al., 2018), with one causal convolution followed by three residual blocks (He et al., 2016):

$$\mathbf{H}_m^{(0)} = \sigma\big(\mathrm{CConv}_{3,1}^m(\mathbf{X}_t^m)\big)\,, \tag{1}$$

$$\mathbf{H}_m^{(\ell)} = \sigma\Big(\mathbf{H}_m^{(\ell-1)} + \mathrm{CConv}_{3,d_\ell}^m\Big(\mathbf{H}_m^{(\ell-1)}\Big)\Big)\,, \quad d_\ell \in \{2,4,8\}, \tag{2}$$

$$\mathbf{h}_t^m = \mathbf{H}_m^{(3)}[-1], \tag{3}$$

where $\mathrm{CConv}_{3,d}^m$ denotes a modality-specific causal one-dimensional convolution with kernel size 3 and dilation d, and $\sigma(\cdot)$ denotes the sigmoid linear unit (SiLU) activation. The final-step feature $\mathbf{h}_t^m$ is the local latent of modality m and is used directly for prediction or subsequently for cross-modal construction.

4.3 Latent Predictor

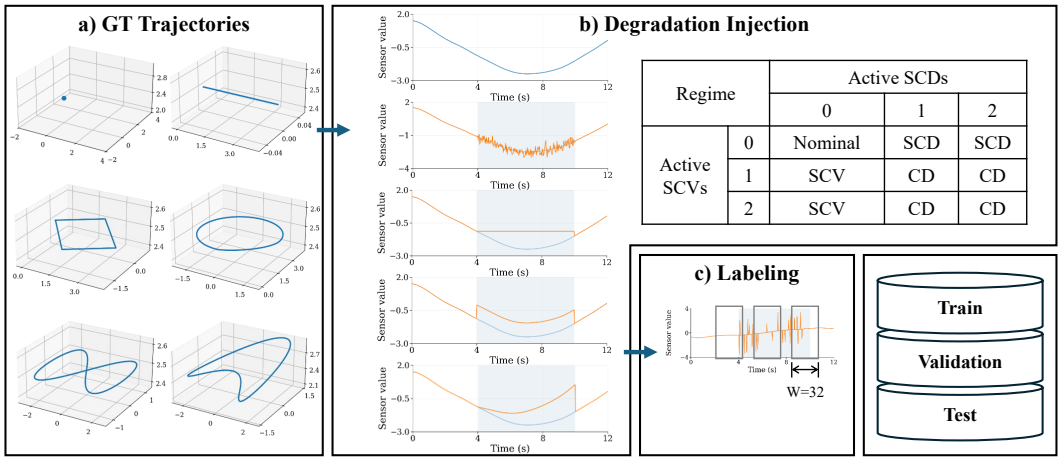
Cross-Modal Latent Prediction. A directed latent predictor maps the local latent of a reference modality n toward the latent space of a target modality m:

$$\mathbf{u}_t^{m\leftarrow n} = \sigma(\mathbf{A}_1^{n\to m}\mathbf{h}_t^n + \mathbf{b}_1^{n\to m})\,, \tag{4}$$

$$\widehat{\mathbf{h}}_t^{m\leftarrow n} = \mathbf{A}_2^{n\to m}\mathbf{u}_t^{m\leftarrow n} + \mathbf{b}_2^{n\to m}, \tag{5}$$

$$\mathbf{e}_t^{m\leftarrow n} = \mathbf{h}_t^m - \widehat{\mathbf{h}}_t^{m\leftarrow n}, \tag{6}$$

where $\mathbf{A}_1^{n\to m}$, $\mathbf{A}_2^{n\to m}$, $\mathbf{b}_1^{n\to m}$, and $\mathbf{b}_2^{n\to m}$ are the parameters of the two-layer predictor $\mathcal{P}_{n\to m}$. The predictors receive no reconstruction loss and are trained end-to-end with the reliability objective. We therefore treat $\mathbf{e}_t^{m\leftarrow n}$ as a task-conditioned relational residual, not as a reconstruction error. Separate predictors are used for the six ordered modality pairs among OF, IMU, and GPS.



Regime		Active SCDs		
		0	1	2
Active SCVs	0	Nominal	SCD	SCD
	1	SCV	CD	CD
	2	SCV	CD	CD

Figure 4: Data generation pipeline for covariance and consistency prediction. "Active SCDs" counts corruption processes; one process can affect one or more consistency-prediction roles.

4.4 Cross-Modal Representation Construction

Composing Cross-Modal Representations. For a target modality m, pairwise relational residuals from the two reference modalities provide complementary cross-modal relations. Let n_1 and n_2 denote the other two modalities. We construct

$$\mathbf{r}_t^m = \mathrm{Concat}(\mathbf{h}_t^m, \mathbf{h}_t^{n_1}, \mathbf{h}_t^{n_2}, \mathbf{e}_t^{m\leftarrow n_1}, \mathbf{e}_t^{m\leftarrow n_2}),$$

where $\mathrm{Concat}(\cdot)$ denotes feature-wise concatenation. The cross-modal latent $\mathbf{r}_t^m$ contains the target-local latent, both reference latents, and two target-directed relational residuals. Thus, Cross-Modal adds cross-modal evidence without removing the target-local evidence: the comparison is between local-only and local-plus-cross-modal representations.

4.5 Dataset Generation

Data Generation. Covariance and consistency supervision are generated from UAV trajectories without changing the underlying vehicle motion (Fig. 4). We generate synchronized OF, IMU, and GPS measurements at 50 Hz over six trajectory families and extract causal windows of length $W = 32$. The benchmark crosses zero, one, or two active SCV roles with zero, one, or two active SCD corruption processes in a 3×3 design. An SCD process can affect one or more of the OF-velocity, GPS-position, and GPS-velocity consistency roles; for example, GPS freezing affects both GPS roles. The four resulting regimes are nominal sensing, isolated SCV, isolated SCD, and compound degradation.

For covariance prediction, zero, one, or multiple roles are selected for SCV among OF velocity, the IMU accelerometer, the IMU gyroscope, GPS position, and GPS velocity. For role j,

$$k_j = 10^{u_j}, \qquad u_j \sim \mathcal{U}(0, 2),$$

with target

$$C_{t,j}^{\mathrm{GT}} = k_j C_j^{\mathrm{nom}}.$$

We use the corresponding covariance to generate the noisy observation and retain it as the regression target.

For consistency prediction, temporally bounded structured corruptions include signal freezing, position or velocity bias and drift, coherent but incorrect OF velocity, time-varying outliers, and outlier bursts. For $j \in \{\text{OF velocity}, \text{GPS position}, \text{GPS velocity}\}$,

$$a_{t,j}^{\mathrm{GT}} = 1 - \mathbb{I}[t \in \mathcal{D}_j],$$

where $\mathcal{D}_j$ contains the indices at which the injected corruption disrupts the nominal sensing relationship of role j, and $\mathbb{I}[\cdot]$ is the indicator function. SCV alone leaves $a_{t,j}^{\mathrm{GT}} = 1$; compound-degradation samples jointly contain SCV and SCD.

4.6 TASK HEAD

Task-Specific Prediction. Each prediction task can operate on either a modality-local or cross-modal representation:

$$\mathbf{z}_t^m = \begin{cases} \mathbf{h}_t^m, & \text{Local,} \\ \mathbf{r}_t^m, & \text{Cross-Modal.} \end{cases}$$

The task head contains three linear layers with SiLU activations after the first two:

$$\mathbf{v}_t^m = \sigma(\mathbf{B}_1^m \mathbf{z}_t^m + \mathbf{c}_1^m)\,, \tag{7}$$

$$\mathbf{w}_t^m = \sigma(\mathbf{B}_2^m \mathbf{v}_t^m + \mathbf{c}_2^m)\,, \tag{8}$$

$$\mathbf{o}_t^m = \mathbf{B}_3^m \mathbf{w}_t^m + \mathbf{c}_3^m, \tag{9}$$

where $\mathbf{o}_t^m$ is the raw output vector and $o_{t,j}$ is its component for prediction role j.

For covariance prediction, the OF, IMU, and GPS heads output one, two, and two terms, respectively:

$$\widehat{R}_{\mathrm{OF}}, \qquad (\widehat{Q}_{\mathrm{acc}}, \widehat{Q}_{\mathrm{gyro}}), \qquad (\widehat{R}_{\mathrm{GPS,pos}}, \widehat{R}_{\mathrm{GPS,vel}}).$$

Each raw output is mapped to

$$\widehat{C}_{t,j} = C_j^{\mathrm{nom}} \left[1 + (M_{\max} - 1)\,\mathrm{sigmoid}(o_{t,j})\right], \qquad M_{\max} = 100.$$

The OF and GPS consistency heads instead produce

$$\widehat{a}_{t,j} = \mathrm{sigmoid}(o_{t,j}).$$

Estimator Modulation. The network predicts the assembled process covariance matrix $\widehat{\mathbf{Q}}_t$, the role-specific measurement covariance terms $\widehat{R}_{t,j}$, and the consistency scores $\widehat{a}_{t,j}$. The process covariance is used directly, whereas each measurement covariance term is modulated by its corresponding consistency score:

$$\mathbf{Q}_t^{\mathrm{ESKF}} = \widehat{\mathbf{Q}}_t, \qquad \bar{a}_{t,j} = \max(\widehat{a}_{t,j}, a_{\min}), \qquad R_{t,j}^{\mathrm{ESKF}} = \frac{\widehat{R}_{t,j}}{\bar{a}_{t,j}},$$

where $a_{\min} = 1/30$, so consistency can inflate $\widehat{R}_{t,j}$ by at most $30\times$. A lower consistency score gives the measurement less weight in the ESKF update.

5 EXPERIMENTS AND RESULTS

5.1 EXPERIMENTAL SETUP AND EVALUATION PROTOCOL

Experimental Setup. We evaluate UAV OF/IMU/GPS state estimation under four sensing conditions: nominal sensing, isolated SCV, isolated SCD, and compound degradation. All learned models operate causally on sensor histories of length $W = 32$ and provide covariance and/or consistency predictions to the same non-learned ESKF.

Training objectives, optimization settings, matched-capacity dimensions, and parameter counts are provided in Appendix B.

Evaluation Protocol. The final downstream benchmark contains 12 unseen clean trajectories from six UAV trajectory families, with two realizations per family. Each trajectory is crossed with the 3×3 SCV–SCD design, and two independent degradation realizations are generated for every trajectory–cell combination, yielding $12 \times 9 \times 2 = 216$ evaluation cases. For each case, we use an antithetic noise pair (Hammersley & Morton, 1956), giving 432 evaluation runs. We first average each antithetic pair, aggregate the cells and two degradation realizations belonging to the same clean trajectory, and then use 5,000 bootstrap resamples (Efron & Tibshirani, 1986) over 12 clean-trajectory clusters with six-family stratification. Thus, the clean trajectory, rather than an individual degradation realization or timestep, is the unit of dependence. We compare the five estimator configurations in Table 1. The matched Local–Cross comparisons use the same paired, six-family-stratified cluster bootstrap. Consistency prediction is evaluated from the continuous output using the area under the receiver operating characteristic curve (AUROC) and the area under the precision–recall curve

Table 1: Downstream estimator configurations

Config.	Covariance	Consistency	Measurement Handling	Role
F0	Fixed nominal	–	Standard ESKF update	Baseline
F1	Local representation	–	Standard ESKF update	Covariance only
F2	Local representation	Local representation	Consistency-based covariance modulation	Local consistency
F3	Local representation	Cross-Modal representation	Consistency-based covariance modulation	**Final**
F4	Local representation	–	99.9% χ^2 innovation gate	Gating baseline

(AUPRC) (Fawcett, 2006; Davis & Goadrich, 2006; Saito & Rehmsmeier, 2015). Because inconsistency is treated as the positive class, we use $1 - \widehat{a}_{t,j}$ as the ranking score. Covariance prediction is evaluated by the mean absolute error (MAE) between predicted and ground-truth log-covariance terms during the active-degradation phase; the same definition is used for isolated SCV and compound degradation in Fig. 5(a).

Downstream estimator calibration is assessed using the normalized innovation squared (NIS) and normalized estimation error squared (NEES), standard consistency diagnostics for stochastic state estimators (Bar-Shalom et al., 2001; Ivanov et al., 2014; Piché, 2016). NIS normalizes the measurement innovation by its predicted innovation covariance, whereas NEES normalizes the state-estimation error by the reported state covariance. Under a statistically consistent Gaussian filter, these quadratic forms follow chi-squared distributions with degrees of freedom equal to the corresponding innovation or state dimension. We therefore divide each statistic by its dimension, giving an expected value of one: values substantially above one indicate that the filter is overconfident relative to its realized errors, while values substantially below one indicate conservative uncertainty estimates. NEES requires ground-truth state and evaluates state-space consistency directly; NIS can be computed from innovations without access to the true state and evaluates consistency in measurement space.

5.2 Downstream Performance and Component Attribution

Downstream Performance and Component Attribution. Table 2 shows that covariance adaptation primarily benefits SCV, whereas consistency modulation provides the larger gains when the nominal sensing relationship is disrupted. Replacing Local with Cross-Modal consistency yields the final improvement under SCD and compound degradation.

5.3 Local–Cross Representation Comparison

Matched Local–Cross Comparison. Fig. 5 compares the two representations with the same prediction target, training protocol, and evaluation set. Parameter counts differ by less than 1% within each capacity tier. Cross-Modal has higher covariance error but better consistency discrimination in both degradation regimes and at both capacities. Figure 5 shows the paired effects for readability; Tables 3 and 4 report the corresponding Local and Cross-Modal absolute MAE, AUROC, and AUPRC values together with all paired differences and confidence intervals.

5.4 Estimator Consistency and Nominal Behavior

Under nominal sensing, F3 has normalized OF, GPS position, and GPS velocity NIS of 0.975, 0.939, and 0.922, respectively, and velocity NEES of 0.992. Position NEES remains moderately overconfident at 1.556. The degraded-condition analysis uses paired contrasts over the active-degradation interval.

Table 2: Downstream performance and primary component attribution. Panel (a) reports position/velocity MAE (m / m/s; lower is better). Panel (b) reports paired mean differences over complete trajectories after antithetic-pair averaging; negative values favor the first method.

(a) Downstream performance

Method	Nominal	SCV	SCD	CD	Overall
F0	0.0172 / 0.0583	0.0238 / 0.0860	0.3421 / 0.2080	0.3406 / 0.2155	0.2346 / 0.1676
F1	0.0175 / 0.0586	0.0199 / 0.0651	0.3497 / 0.1717	0.3337 / 0.1656	0.2324 / 0.1327
F2	0.0175 / 0.0586	0.0202 / 0.0651	0.2222 / 0.1120	0.2313 / 0.1169	0.1586 / 0.0978
F3	0.0175 / 0.0586	0.0202 / 0.0652	0.2106 / 0.1032	0.2093 / 0.1056	**0.1462 / 0.0909**
F4	0.0175 / 0.0587	0.0199 / 0.0651	1.5922 / 0.1410	1.8705 / 0.1607	1.1915 / 0.1237

(b) Primary component attribution

Contrast	Regime	Δ Pos. MAE	Δ Vel. MAE
F1–F0	SCV	-0.0039	-0.0209
F2–F1	SCD	-0.1275	-0.0597
F2–F1	CD	-0.1024	-0.0487
F3–F2	SCD	-0.0116	-0.0088
F3–F2	CD	-0.0219	-0.0113

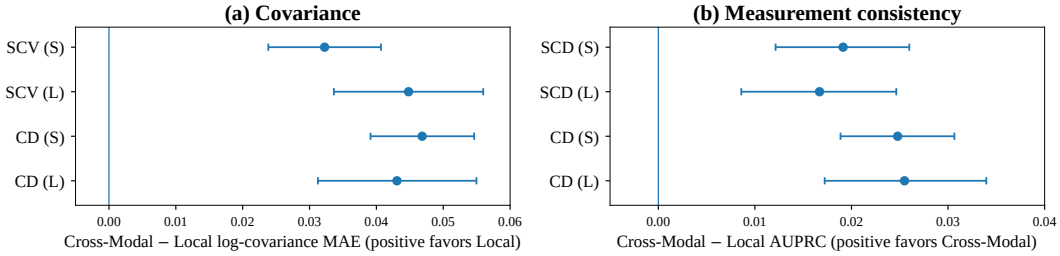


Figure 5: Matched Local–Cross controls shown as paired mean differences with 95% six-family-stratified trajectory-cluster bootstrap confidence intervals. (a) Cross-Modal−Local active-degradation log-covariance MAE under isolated SCV and compound degradation; positive values favor Local because lower MAE is better. (b) Cross-Modal−Local AUPRC for measurement-consistency prediction under isolated SCD and compound degradation; positive values favor Cross-Modal. S/L denote the Small/Large matched-capacity tiers. All eight intervals exclude zero.

Table 3: Complete matched-capacity covariance results. Lower values are better.

Tier	Regime	Local MAE	Cross MAE	Δ	95% CI
Small	SCV	0.0720	0.1043	0.0323	[0.0238, 0.0407]
Small	CD	0.1442	0.1910	0.0468	[0.0391, 0.0546]
Large	SCV	0.0679	0.1127	0.0448	[0.0336, 0.0560]
Large	CD	0.1491	0.1922	0.0431	[0.0313, 0.0550]

Under SCD and compound degradation, F3 reduces position and velocity calibration error relative to F2 (Table 5). The compound-degradation OF NIS contrast slightly favors F2 because its normalized value is already closer to one. Position NEES remains overconfident, so neither configuration is statistically consistent.

Hard innovation gating is less robust under compound degradation: the trajectory-wide normalized position NEES is 1531.2 for F4 versus 181.6 for F3, and velocity NEES is 11.27 versus 1.97. Here, continuous consistency modulation is more effective than hard rejection.

Table 4: Complete matched-capacity consistency-prediction results. Higher values are better.

Tier	Regime	Metric	Local	Cross	Δ	95% CI
Small	SCD	AUROC	0.9800	0.9863	0.0063	[0.0021, 0.0106]
Small	SCD	AUPRC	0.9590	0.9781	0.0191	[0.0121, 0.0260]
Small	CD	AUROC	0.9733	0.9925	0.0192	[0.0124, 0.0259]
Small	CD	AUPRC	0.9607	0.9855	0.0248	[0.0189, 0.0307]
Large	SCD	AUROC	0.9805	0.9882	0.0077	[0.0053, 0.0101]
Large	SCD	AUPRC	0.9631	0.9797	0.0167	[0.0086, 0.0246]
Large	CD	AUROC	0.9749	0.9923	0.0174	[0.0138, 0.0210]
Large	CD	AUPRC	0.9591	0.9846	0.0255	[0.0172, 0.0340]

Table 5: Estimator-consistency contrasts under degradation. The middle column gives trajectory-wide dimension-normalized values (expected value: 1). For the active-degradation phase, $\Delta_{\mathrm{cal}} = |\log s_{\mathrm{cand}}| - |\log s_{\mathrm{ref}}|$, where s is the dimension-normalized NIS or NEES; negative values indicate that the candidate is closer to one. The 95% confidence intervals use the same six-family-stratified trajectory-cluster bootstrap.

Regime	Contrast / metric	Overall Ref. $\rightarrow$ Cand.	Δ_{cal}	95% CI
SCV	F1–F0, OF NIS	$2.796 \rightarrow \mathbf{0.984}$	$\mathbf{-1.783}$	$[-2.032, -1.431]$
SCD	F3–F2, position NEES	$433.2 \rightarrow \mathbf{320.5}$	$\mathbf{-0.315}$	$[-0.419, -0.267]$
SCD	F3–F2, velocity NEES	$7.75 \rightarrow \mathbf{5.68}$	$\mathbf{-0.371}$	$[-0.694, -0.232]$
CD	F3–F2, OF NIS	$1.003 \rightarrow 0.943$	$+0.227$	$[+0.010, +0.347]$
CD	F3–F2, position NEES	$542.5 \rightarrow \mathbf{181.6}$	$\mathbf{-1.076}$	$[-1.374, -0.754]$
CD	F3–F2, velocity NEES	$4.17 \rightarrow \mathbf{1.97}$	$\mathbf{-0.966}$	$[-1.204, -0.655]$

6 Conclusion

We compared modality-local and cross-modal representations for learned covariance and consistency prediction. On this benchmark, Local gives lower covariance error, while Cross-Modal gives better consistency discrimination across both tested degradation regimes and model capacities. Using Local covariance and Cross-Modal consistency in the ESKF improves estimation under degraded sensing without reducing nominal accuracy. Cross-modal inputs are therefore not uniformly beneficial: their value depends on the reliability target.

AI use statement

In this work, we have not used generative AI tools for tasks with required or recommended disclosure.

Reproducibility statement

The experimental protocol is described in Sec. 5.1. Appendices A, B, and C document the data generation and splits, training configuration, and estimator configuration, respectively. The accompanying reproducibility package provides the implementation, checkpoints, numerical result sources, software requirements, and an end-to-end execution script corresponding to these experiments.

References

Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. Deep evidential regression. In *Advances in Neural Information Processing Systems*, volume 33, pp. 14927–14937, 2020.

Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2): 423–443, 2019. doi: 10.1109/TPAMI.2018.2798607.

Yaakov Bar-Shalom, X.-Rong Li, and Thiagalingam Kirubarajan. *Estimation with Applications to Tracking and Navigation: Theory, Algorithms and Software*. John Wiley & Sons, 2001. doi: 10.1002/0471221279.

Timothy D. Barfoot. *State Estimation for Robotics*. Cambridge University Press, 2017. doi: 10.1017/9781316671528.

Grigor Bezirganyan, Sana Sellami, Laure Berti-Equille, and Sébastien Fournier. Multimodal learning with uncertainty quantification based on discounted belief fusion. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 3142–3150, 2025.

Martin Brossard, Axel Barrau, and Silvère Bonnabel. AI-IMU dead-reckoning. *IEEE Transactions on Intelligent Vehicles*, 5(4):585–595, 2020. doi: 10.1109/TIV.2020.2980758.

Shiteng Cao, Qiyuan Liu, and Zhiheng Li. Leveraging cross-modal consistency for sensor fault detection in autonomous driving. In *2026 IEEE Intelligent Vehicles Symposium (IV)*, pp. 452–457. IEEE, 2026. doi: 10.1109/IV66570.2026.11623995.

Changhao Chen, Stefano Rosa, Yishu Miao, Chris Xiaoxuan Lu, Wei Wu, Andrew Markham, and Niki Trigoni. Selective sensor fusion for neural visual-inertial odometry. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10542–10551, 2019. doi: 10.1109/CVPR.2019.01079.

Minkyoung Cho, Yulong Cao, Jiachen Sun, Qingzhao Zhang, Marco Pavone, Jeong Joon Park, Heng Yang, and Zhuoqing Mao. Cocoon: Robust multi-modal perception with uncertainty-aware sensor fusion. In *The Thirteenth International Conference on Learning Representations*, 2025.

Jesse Davis and Mark Goadrich. The relationship between precision-recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning*, pp. 233–240, 2006. doi: 10.1145/1143844.1143874.

Bradley Efron and Robert Tibshirani. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, 1(1):54–75, 1986. doi: 10.1214/ss/1177013815.

Tom Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006. doi: 10.1016/j.patrec.2005.10.010.

Jamil Fayyad, Mohammad A. Jaradat, Dominique Gruyer, and Homayoun Najjaran. Deep learning sensor fusion for autonomous vehicle perception and localization: A review. *Sensors*, 20(15): 4220, 2020. doi: 10.3390/s20154220.

Zixian Gao, Xun Jiang, Xing Xu, Fumin Shen, Yujie Li, and Heng Tao Shen. Embracing unimodal aleatoric uncertainty for robust multimodal fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26876–26885, 2024.

Wenzhong Guo, Jianwen Wang, and Shiping Wang. Deep multimodal representation learning: A survey. *IEEE Access*, 7:63373–63394, 2019. doi: 10.1109/ACCESS.2019.2916887.

J. M. Hammersley and K. W. Morton. A new Monte Carlo technique: Antithetic variates. *Mathematical Proceedings of the Cambridge Philosophical Society*, 52(3):449–475, 1956. doi: 10.1017/S0305004100031455.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Pavel Ivanov, Simo Ali-Löyttÿ, and Robert Piché. Evaluating the consistency of estimation. In *Proceedings of the 2014 International Conference on Localization and GNSS*, pp. 1–5, 2014. doi: 10.1109/ICL-GNSS.2014.6934171.

R. E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(1):35–45, 1960. doi: 10.1115/1.3662552.

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, pp. 5574–5584, 2017.

Kailong Li, Lubin Chang, and Baiqing Hu. A variational Bayesian-based unscented Kalman filter with both adaptivity and robustness. *IEEE Sensors Journal*, 16(18):6966–6976, 2016. doi: 10.1109/JSEN.2016.2591260.

Changhong Lin, Jiarong Lin, Zhiqiang Sui, Xiaozhi Qu, Rui Wang, Kehua Sheng, and Bo Zhang. An end-to-end learning-based multi-sensor fusion for autonomous vehicle localization. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 7460–7466. IEEE, 2025. doi: 10.1109/ICRA55743.2025.11128443.

Katherine Liu, Kyel Ok, William Vega-Brown, and Nicholas Roy. Deep inference for covariance estimation: Learning Gaussian noise models for state estimation. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1436–1443. IEEE, 2018. doi: 10.1109/ICRA.2018.8461047.

Wenxin Liu, David Caruso, Eddy Ilg, Jing Dong, Anastasios I. Mourikis, Kostas Daniilidis, Vijay Kumar, and Jakob Engel. TLIO: Tight learned inertial odometry. *IEEE Robotics and Automation Letters*, 5(4):5653–5660, 2020. doi: 10.1109/LRA.2020.3007421.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Xiaowen Luo and Haitao Wang. Robust adaptive Kalman filtering—a method based on quasi-accurate detection and plant noise variance–covariance matrix tuning. *The Journal of Navigation*, 70(1):137–148, 2017. doi: 10.1017/S0373463316000564.

Huan Ma, Zongbo Han, Changqing Zhang, Huazhu Fu, Joey Tianyi Zhou, and Qinghua Hu. Trustworthy multimodal regression with mixture of normal-inverse gamma distributions. In *Advances in Neural Information Processing Systems*, volume 34, pp. 6881–6893, 2021.

Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. Are multimodal transformers robust to missing modality? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18177–18186, 2022.

R. K. Mehra. On the identification of variances and adaptive Kalman filtering. *IEEE Transactions on Automatic Control*, 15(2):175–184, 1970. doi: 10.1109/TAC.1970.1099422.

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. Multimodal deep learning. In *Proceedings of the 28th International Conference on Machine Learning (ICML)*, pp. 689–696, 2011.

Barak Or and Itzik Klein. A hybrid model and learning-based adaptive navigation filter. *IEEE Transactions on Instrumentation and Measurement*, 71:1–11, 2022. doi: 10.1109/TIM.2022.3197775.

Robert Piché. Online tests of Kalman filter consistency. *International Journal of Adaptive Control and Signal Processing*, 30(1):115–124, 2016. doi: 10.1002/acs.2571.

Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 10(3):e0118432, 2015. doi: 10.1371/journal.pone.0118432.

Joan Solà. Quaternion kinematics for the error-state Kalman filter. *arXiv preprint arXiv:1711.02508*, 2017.

William Talbot, Julian Nubert, Turcan Tuna, Cesar Cadena, Frederike Dümbgen, Jesús Tordesillas, Timothy D. Barfoot, and Marco Hutter. Continuous-time state estimation methods in robotics: A survey. *IEEE Transactions on Robotics*, 41:4975–4999, 2025. doi: 10.1109/TRO.2025.3593079.

Fan Wu, Haiyong Luo, Hongwei Jia, Fang Zhao, Yimin Xiao, and Xile Gao. Predicting the noise covariance with a multitask learning model for Kalman filter-based GNSS/INS integrated navigation. *IEEE Transactions on Instrumentation and Measurement*, 70:1–13, 2021. doi: 10.1109/TIM.2020.3024357.

Fei Yan, Sheng Li, Enze Zhang, Jian Guo, and Qingwei Chen. An adaptive nonlinear filter for integrated navigation systems using deep neural networks. *Neurocomputing*, 446:130–144, 2021. doi: 10.1016/j.neucom.2021.03.046.

Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. In *International Conference on Learning Representations*, 2016.

Haoming Zhang, Zhanxin Wang, and Heike Vallery. Learning-based NLOS detection and uncertainty prediction of GNSS observations with transformer-enhanced LSTM network. In *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*, pp. 910–917. IEEE, 2023a. doi: 10.1109/ITSC57777.2023.10422672.

Qingyang Zhang, Haitao Wu, Changqing Zhang, Qinghua Hu, Huazhu Fu, Joey Tianyi Zhou, and Xi Peng. Provable dynamic fusion for low-quality multimodal data. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 41753–41769, 2023b.

Hao Zhu, Guorui Zhang, Yongfu Li, and Henry Leung. An adaptive Kalman filter with inaccurate noise covariances in the presence of outliers. *IEEE Transactions on Automatic Control*, 67(1): 374–381, 2022. doi: 10.1109/TAC.2021.3056343.

Yuan Zhuang, Xiao Sun, You Li, Jianzhu Huai, Luchi Hua, Xiansheng Yang, Xiaoxiang Cao, Peng Zhang, Yue Cao, Longning Qi, Jun Yang, Nashwa El-Bendary, Naser El-Sheimy, John S. Thompson, and Ruizhi Chen. Multi-sensor integrated navigation/positioning systems using data fusion: From analytics-based to learning-based approaches. *Information Fusion*, 95:62–90, 2023. doi: 10.1016/j.inffus.2023.01.025.

A Data Generation and Splits

Clean sensor streams are collected in PX4 SITL with Gazebo at 50 Hz for 40 s, yielding 2,000 synchronized samples per flight. The six disjoint trajectory families are hover, point-to-point translation, multi-waypoint flight, orbit, planar lemniscate, and coupled three-dimensional motion. The train, validation, and final splits contain five, one, and two independently collected realizations per family, respectively (30/6/12 clean trajectories). No clean trajectory is shared across splits.

Each clean trajectory is crossed with the 3×3 design formed by zero, one, or two active SCV roles and zero, one, or two active SCD processes. Two independently sampled degradation realizations are generated for each cell. Events begin between steps 420 and 719 and last 520–899 steps. SCV affects one or two of R_{OF}, Q_{acc}, Q_{gyro}, $R_{\mathrm{GPS,pos}}$, and $R_{\mathrm{GPS,vel}}$ with a multiplier sampled log-uniformly from $[1, 100]$. SCD processes comprise GPS freeze, GPS position/velocity bias, GPS position/velocity drift, coherent incorrect OF velocity, dynamic OF outliers, and OF outlier bursts. The final split therefore has $12 \times 9 \times 2 = 216$ logical cases. Each is evaluated with positive and negative antithetic noise, producing 432 runs; antithetic pairs are averaged before trajectory-level statistics. All windows from a trajectory remain in the same split.

B Training and Capacity Details

For each task, the modality encoders, directed latent predictors (Cross-Modal only), and task heads are trained end-to-end with AdamW (Loshchilov & Hutter, 2019) at a learning rate of 3×10^{-4}; the

Table 6: Training hyperparameters.

Setting	Covariance	Consistency
Objective	log-covariance MSE	binary cross-entropy
Optimizer	AdamW	AdamW
Learning rate	3×10^{-4}	3×10^{-4}
Weight decay	0	10^{-4}
AdamW (β_1, β_2)	$(0.9, 0.999)$	$(0.9, 0.999)$
AdamW ϵ	10^{-8}	10^{-8}
Batch size	512	256
Updates	20,000	12,000
Gradient-norm clip	10	5
Validation interval	200	500
Selection metric	log-covariance RMSE	macro-AUPRC
Training seed	94051	73141
Validation-sampling seed	94052	73142
Batch-sampling seed	94053	73143
History length	$W = 32$	

latent predictors receive no separate reconstruction objective. Table 6 gives the complete optimization settings. Training samples a degradation cell uniformly, then a sequence and causal window within that cell. Model selection uses 32,768 fixed validation windows.

The matched-capacity covariance tiers are Local $h = 64$ (139,525 parameters) versus Cross-Modal $h = 49$ (140,782) for Small, and Local $h = 84$ (238,565) versus Cross-Modal $h = 64$ (238,597) for Large. For consistency, the corresponding tiers are Local $h = 64$ (131,075) versus Cross-Modal $h = 50$ (131,303), and Local $h = 82$ (213,695) versus Cross-Modal $h = 64$ (213,763).

C ESKF CONFIGURATION

The non-learned ESKF uses the nominal state $(\mathbf{p}, \mathbf{v}, \mathbf{q}, \mathbf{b}_a, \mathbf{b}_g)$ and a 15-dimensional error state. IMU propagation runs at $\Delta t = 0.02$ s with gravity -9.81 m/s^2. The nominal accelerometer and gyroscope noise standard deviations are 0.12 m/s^2 and 0.015 rad/s; their bias random-walk standard deviations are 0.004 and 0.0005, respectively. Nominal GPS position and velocity variances are 0.0025 m^2 and 0.01 (m/s)^2, and the planar OF velocity variance is 0.01 (m/s)^2.

The filter is initialized after 25 IMU samples. Position and velocity are initialized from the contemporaneous GPS observation, while attitude and its covariance are obtained from the mean measured gravity direction and its uncertainty. The initial covariance uses $0.0025\mathbf{I}_3$ for position, $0.01\mathbf{I}_3$ for velocity, the gravity-derived covariance for attitude, and $0.01\mathbf{I}_3$ for each bias block.

Learned process-covariance predictions directly supply the absolute accelerometer and gyroscope process covariances. Learned measurement-covariance predictions directly supply the absolute OF and GPS observation variances. Consistency subsequently inflates each predicted measurement variance as $R_{t,j}^{\mathrm{ESKF}} = \widehat{R}_{t,j} / \max(\widehat{a}_{t,j}, 1/30)$. All configurations use the same propagation and measurement models, Joseph-form covariance update, initialization, and sensor timing. F2 and F3 apply continuous consistency modulation, whereas F4 instead applies a 99.9% chi-squared innovation gate using the nominal measurement covariance for the gating decision.