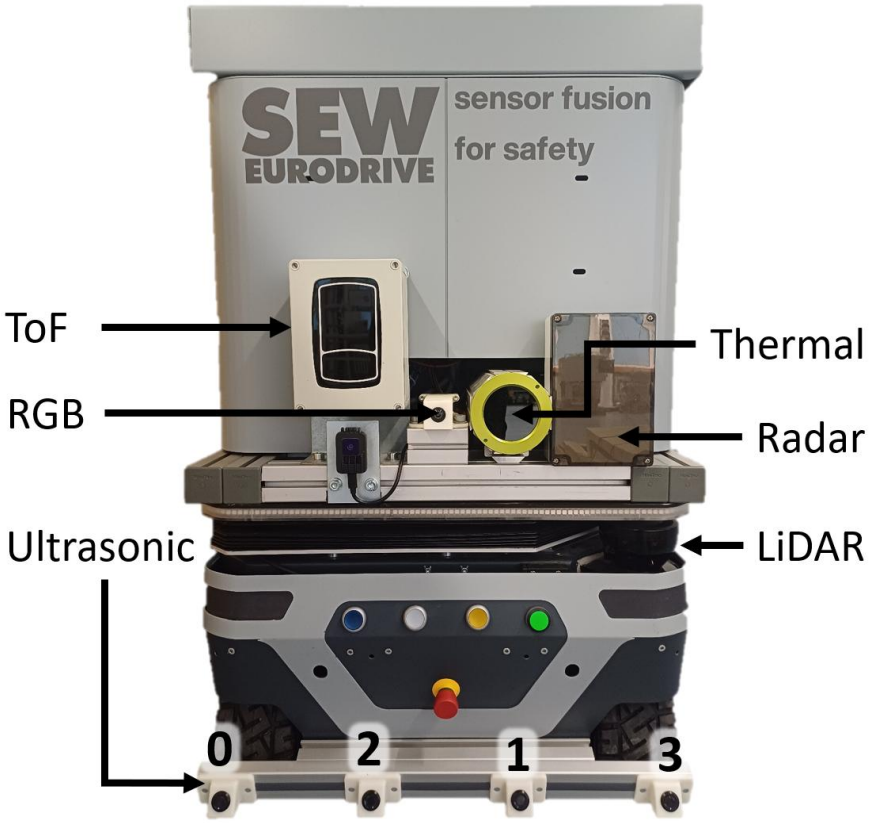
TEAM 4 / STATE OF THE ART

Sensor Interplay, Degradation and Fallback

Thermal, ToF depth and laser
for person detection on an industrial mobile robot

Supervisor: Myroslav Mishchuk
Team:
Oluwasesan Cornelius Akinbode (SUT)
Andrii Baran (LPNU)
Daniel Dąbrowski (SUT)



SEW-EURODRIVE dataset documentation
Image: CC BY 4.0

PART 1 / Q&A

Questions from the other teams

Seven questions from Teams 1, 2, 3, 5 and 6, one slide each.

Team 5	How do we tell complementarity, redundancy and conflict apart?	RQ1 · RQ2
Team 3	Genuine interactions or architecture artefacts? Value beyond ablation?	RQ1 · RQ2
Team 1	Do interaction scores save effort with only three streams?	RQ1
Team 6	Retrained ablation or inference-time removal?	RQ1 · RQ2
Team 2	Is a weak sensor still needed in difficult conditions?	RQ1 · RQ3
Team 1	Is the gate calibrated on clean or degraded validation data?	RQ3
Team 6	Should the gate face a degradation-trained baseline?	RQ3

How do we tell the three regimes apart?

TEAM 5 ASKED

"How will you objectively distinguish between sensor complementarity, redundancy, and conflict? What criteria will you use to determine which type of interaction is occurring?"

By measured fusion gain from retrained models, with thresholds fixed before testing. The interaction scores are then checked against these labels (H1).

Conflicting	Redundant	Complementary
gain < −δ → down-weight or gate	|gain| ≤ δ → prune or shrink	gain > +δ → keep both

−δ +δ

Fusion gain = fused model − best single-stream model, retrained, on the same test frames (the track's definition)

1. **Labels come from ablation:** δ is fixed in the pilot, before testing, so the scores cannot define their own ground truth.

2. **Only clear cases count:** the bootstrap interval over recordings must clear δ, else inconclusive; computed per condition and severity.

Genuine interactions or architecture artefacts?

TEAM 3 ASKED

"Since a sensor's apparent contribution depends on the fusion architecture, how will you separate genuine sensor interactions from architecture-specific artefacts, and what do XAI interaction measures add beyond full ablation (2^N-1), whether done by inference-time removal or by retraining?"

Scores explain a trained model, not the sensors, so we trust an interaction only where independent checks agree.

1. **Genuine:** it holds across both fusion designs, across seeds and against retrained subset models.

2. **Built-in control:** probability-averaging late fusion has zero pairwise interactions by construction. Non-zero means artefact.

3. **Beyond ablation:** one number per subset becomes signed single, pairwise and three-way effects, per frame, without labels.

Do scores save effort with three streams?

TEAM 1 ASKED

"With only three streams (thermal, ToF depth, laser) there are just 7 non-empty subsets to ablate. When would interaction scores save real effort compared with direct ablation, and does the claimed benefit (screening before costly ablation) hold at this scale?"

Not at N = 3, and that is why we test H1 here: the full ablation reference is still affordable.

1. **Validate first:** with 7 subsets we can check the scores against complete ground truth.
2. **Savings at scale:** retraining needs $2^N - 1$ models per design and seed: 7 now, 63 with all six SEW sensor types.
3. **Other gains:** per-condition and per-frame analysis, and label-free monitoring of the deployed model.

Retrained ablation or inference-time removal?

TEAM 6 ASKED

"Ablation is your ground truth, but retraining a model without a stream and removing that stream from the fused model at inference measure different things. Which one will you use, and how will it match the replacement strategy behind your interaction scores?"

Both, with separate roles. Inference-time removal uses exactly the same replacement as the scores.

1. **Retraining = ground truth for H1:** is the stream's information needed? No replacement involved.

2. **Inference-time removal = diagnostic:** does the deployed model rely on the stream? Same rule and reference frames as the scores.

3. **The gap is informative:** without retraining, a drop may reflect out-of-distribution inputs (Hooker et al., 2019).

Is a weak sensor still needed in hard conditions?

TEAM 2 ASKED

“Before deciding to remove a sensor that contributes little to average prediction accuracy, how will you check whether it remains valuable in difficult conditions, such as poor visibility or failure of another sensor?”

We never remove a stream on the average alone: it must be dispensable in every condition and failure case we can test.

1. **Per condition:** quality, missed people and interactions reported separately for night, rain, snow and indoor frames (SEW per-frame metadata).

2. **When another sensor fails:** the seven-subset ablation already shows each stream’s value without a partner, e.g. depth + laser vs. laser alone.

3. **Removal rule:** prune only if the preset quality-loss tolerance holds in the worst condition, not just on average.

Clean or degraded validation for the gate?

TEAM 1 ASKED

"The entropy gate in H3 is calibrated on validation data, while the model is trained on clean data and degraded only at evaluation. Is the validation set clean or degraded? If clean, why should the calibrated entropy stay reliable under degradation, given your own risk that a degraded branch stays confidently wrong?"

Clean, by design. We measure whether entropy stays reliable instead of assuming it.

1. **Why clean:** faults are unknown in deployment; tuning on the test corruption would leak it into the gate.
2. **Measured:** calibration can fail under shift (Ovadia et al., 2019), so we report it at every severity.
3. **Beyond entropy:** RQ3 also tests invalid-depth fraction and thermal–laser disagreement, plus a degraded-validation variant.

Should the gate face a degradation-trained baseline?

TEAM 6 ASKED

"Since models see degradation only at evaluation, would a fusion baseline trained with degradation augmentation be a fairer comparison for the entropy gate in H3?"

Yes, as an added baseline: we vary training and gating separately.

	Clean training	Augmented training
Fixed fusion	Current baseline	**New baseline**
Gated fusion	Current RQ3 gate	**Gate + augmentation**

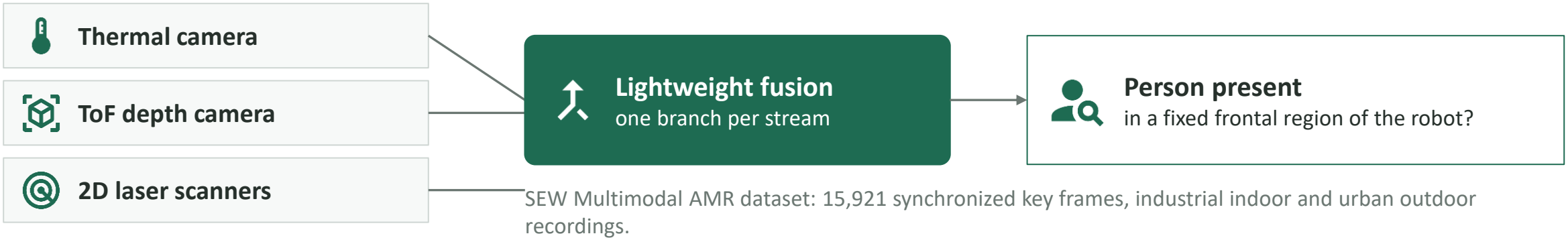
Shaded: new runs.

1. **Why separate:** training exposure can look like robustness (Aydın and Dizdaroğlu, 2026).
2. **Fair test:** also augment with a different degradation, not only the test one.
3. **When:** after the pilot, for the chosen degradation only.

The problem we are working on

Which sensor data helps us detect the person?

When a stream degrades, does it still earn its processing cost?



SEW Multimodal AMR dataset: 15,921 synchronized key frames, industrial indoor and urban outdoor recordings.

Complementary

Fused > best single stream

Action: Keep both; fuse adaptively

Redundant

Fused ≈ best single stream

Action: Prune, downsample or shrink

Conflicting

Fused < best single stream

Action: Down-weight, gate or flag

One hypothesis, three research questions

XAI interaction scores reveal whether sensor streams are complementary, redundant or conflicting under clean and degraded conditions. This knowledge can make a lightweight multimodal model more efficient or more robust without a significant loss in prediction quality.

RQ1 Interplay

Do interaction scores agree with ablation, the ground truth for which streams are needed?

H1. Stronger positive interactions go with larger non-additive gains in ablation.

Daniel Dąbrowski

RQ2 Degradation

As a stream degrades, do its interaction scores still reflect its benefit?

H2. Changes in the degraded stream's interaction scores track changes in its contribution.

Andrii Baran

RQ3 Fallback

Can a cheap run-time signal tell when to stop relying on a degraded stream?

H3. A calibrated-entropy gate loses less quality than fixed-weight fusion as degradation grows.

Oluwasesan Cornelius Akinbode

RQ1 and RQ2 test the interaction claim. RQ3 tests the practical fallback.

One shared experiment in four stages

Data

SEW thermal, ToF depth and 2D laser. Split by recording.

Training

Clean data. One lightweight branch per stream; intermediate and late fusion.

Degradation

One type, applied at evaluation only. Chosen after a pilot.

Evaluation

Quality, fusion gain, XAI fidelity and stability, latency, memory.

Today feeds this stage

1

Pilot

Check SEW samples, alignment and labels. Define the frontal region. Select models and one degradation.

2

Clean baselines

Recording-based splits. Compare all stream subsets and at least two fusion designs.

3

Degradation and fallback

Compare XAI with ablation across severities. Check stability and replacements. Tune the gate on validation.

4

Simplification

Test a smaller or removed branch where justified. Recheck quality, latency and memory.

Today: each member presents the state of the art, reusable resources, the open gap and a proposed test for their RQ.

Today’s agenda

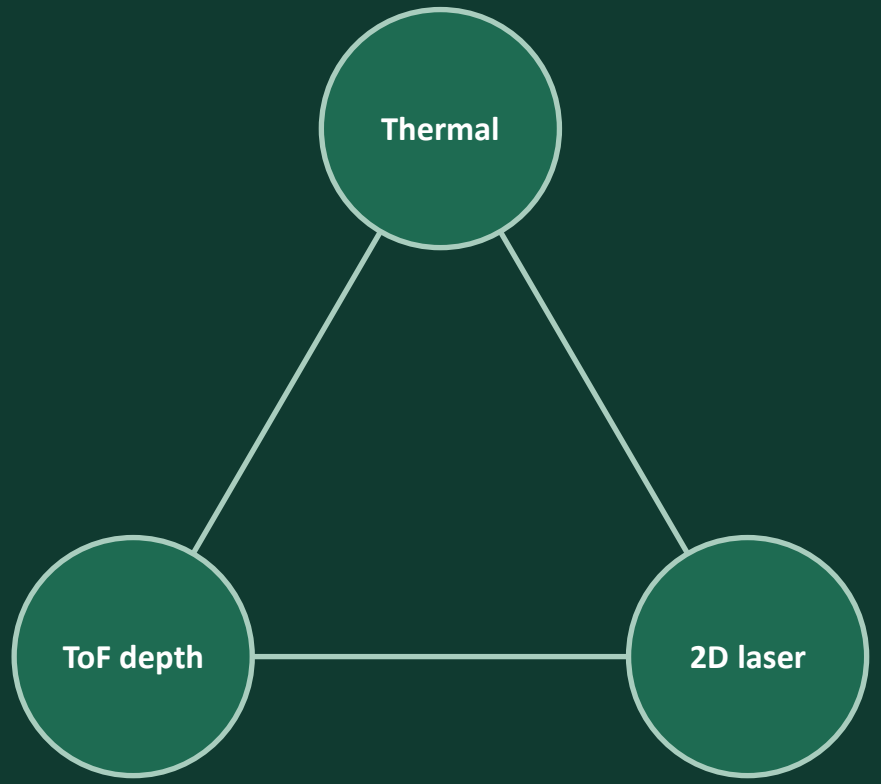
01	Q&A	Questions from Teams 1, 2, 3, 5 and 6, with our proposed answers	Whole team
02	Introduction	Problem, hypothesis and plan	Myroslav Mishchuk
03	RQ1 · Interplay	Interaction methods, code and a five-step experiment	Daniel Dąbrowski
04	RQ2 · Degradation	Degradation studies, explanations and benefit, a preliminary test	Andrii Baran
05	RQ3 · Fallback	Confidence under degradation, candidate signals, protocol	Oluwasesan Cornelius Akinbode

PART 3 / RQ1 · INTERPLAY

Do interaction scores agree with ablation?

H1. Stronger positive interactions go with larger non-additive gains in ablation. Weak or negative interactions go with redundancy or conflict.

Presented by
Daniel Dąbrowski (SUT)



Each stream is a player; interactions are the non-additive effects of using streams together.

Relevant papers and their evidence

Publication	What was established	Direct RQ1 question
Wenderoth et al. (2025) Measuring Cross-Modal Interactions in Multimodal Models	InterSHAP measures interactions between modalities independently of prediction accuracy. Synthetic experiments and comparisons with PID, EMAP and SHAPE validate its ability to detect learned interactions.	Do higher InterSHAP scores correspond to larger performance gains or losses measured through ablation?
Muschalik et al. (2024) shapiq	Different Shapley interaction methods can be evaluated using benchmark games with known interaction values. This validates interaction estimation rather than the predictive benefit of a modality	Which interaction method (SII, STII or FSII) best reflects the performance changes observed through ablation?
Liang et al. (2023) Quantifying & Modeling Multimodal Interactions ...	PID quantifies redundancy, uniqueness, and synergy, validated against synthetic ground truth and human annotations. Missing-modality experiments show that interaction scores correlate with performance drops.	Do PID-based interaction scores faithfully predict performance changes under modality ablation, and how stable are these explanations?
Tsai et al. (2023) Faith-Shap	Comparison of Different Methods for Calculating Interaction Scores	Lack of Ablation Studies Comparing Interaction Scores with Raw Ground Truth

Relevant papers and their evidence

Publication	What was established	Direct RQ1 question
Hessel & Lee (2020) EMAP	Removing cross-modal interactions from a model's predictions often causes little or no performance degradation. Therefore, multimodal models do not always need the interactions they have learned.	Do high interaction scores indicate interactions that are actually necessary for maintaining prediction accuracy?
Hu et al. (2022) SHAPE	SHAPE measures individual modality contributions and their cooperation. Experiments show that models sometimes rely mainly on one modality despite having access to complementary information.	Do SHAPE cooperation scores agree with performance gains from combining modalities, and how do they compare with InterSHAP?
Parcalabescu & Frank (2023) MM-SHAP	MM-SHAP measures the extent to which a model uses each modality independently of task performance. It demonstrates that modality usage and classification accuracy are not equivalent.	Ablation studies focus on measuring performance changes after removing information streams, rather than examining the relationship between SHAPE interaction scores and the impact of removing individual modalities.
Xin et al. (2025) I2MoE	Interaction-aware mixture-of-experts models improve multimodal prediction performance. Architectural ablations demonstrate the usefulness of explicitly modeling different interaction types.	Ablation studies focus on measuring performance changes after removing information streams.

Relevant papers and their evidence

Publication	What was established	Direct RQ1 question
Hedström et al. (2023) Quantus	Quantus provides metrics for evaluating explanation faithfulness, robustness and sensitivity. However, these metrics do not automatically establish agreement with modality-level performance ablation.	Are interaction scores that perform better on fidelity and stability metrics also more consistent with ablation results?
Hooker et al. (2019) ROAR	Remove And Retrain evaluates explanations by deleting important features, retraining the model and measuring the resulting accuracy degradation. Several attribution methods performed no better than random removal.	Can interaction scores identify sensor streams whose removal and retraining cause the greatest loss in predictive performance?

How interaction scores work

Each sensor is a player. Interaction = the non-additive effect of using streams together.

Method	How it works	What it tells us
InterSHAP (SII)	Exact Shapley interaction index over masked modality subsets; local output effects.	Measures joint model dependence; no labels needed.
shapiq (toolkit)	Computes SII, k-SII, Shapley-Taylor and Faith-Shap for a defined coalition game.	Enables controlled comparison of several indices.
Shapley-Taylor	Distributes joint effects up to a chosen order with an efficient decomposition.	Alternative split of lower- and higher-order effects.
Faith-Shap (FSII)	Weighted polynomial fit to coalition values, designed for faithful explanation.	Baseline for explanatory fit and index choice.

Code and integration points

Repository	What it provides	Role in SEW study
InterSHAP LauraWenderoth/InterSHAP	InterSHAP plus PID, EMAP and SHAPE; synthetic HD-XOR and multimodal scripts.	Reproduce toy benchmark; port scoring.
shapiq mmschlk/shapiq	ExactComputer for SII, k-SII, STII and FSII; callable coalition games.	Primary exact 3-player analysis.
SHAP shap/shap	General model explainers, masking infrastructure and attribution plots.	Cross-check masks and visualisations.

What our experiment must validate

A high interaction score does not automatically mean that removing a sensor will reduce held-out predictive quality.

Criterion	Proposed empirical test	Warning sign
Agreement	Compare signed interaction ranks with independent non-additive gains in BAcc and AUPRC.	Signs or rankings diverge.
Fidelity	Check model-output effects and, separately, held-out degradation when a stream is ablated.	Good output fit; weak utility.
Stability	Repeat across seeds, two masking strategies and mild corruption; bootstrap by recording.	Scores or rankings change.

Five steps to answer RQ1

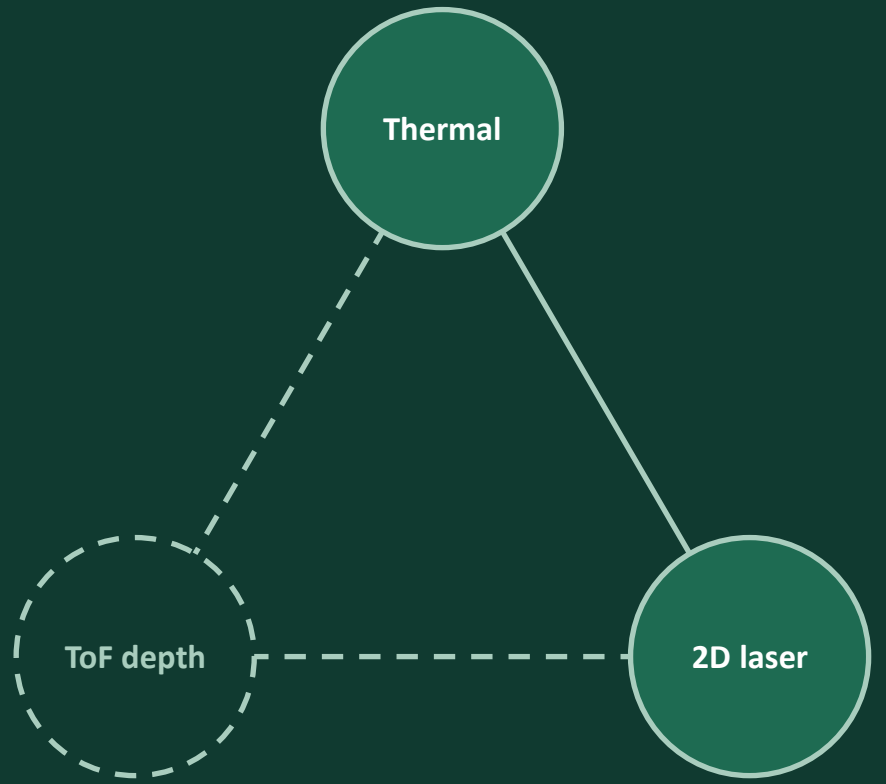
Step	Protocol and evaluation
1. Data and pilot	Verify access, timing and labels; define person in frontal ROI from boxes. Use Thermal, ToF and 2D Laser; split by recording.
2. Clean baselines	Train intermediate/late fusion. Report BAcc, AUPRC, F1; test 7 non-empty sensor subsets with masking and separately trained baselines.
3. Interaction scores	On a frozen model, score all 8 coalitions per frame; compute InterSHAP and exact SII, STII, FSII in shapiq.
4. Validation	Compare with fixed-model ablation and retrained-subset gain
5. Stability and choice	Repeat across seeds, two replacement rules and mild noise; evaluate rank stability, quality loss, latency and memory after pruning.

Support H1 only if agreement is repeatable and pruning meets a preset quality-loss tolerance.

PART 4 / RQ2 · DEGRADATION

Does a stream still pay off as it degrades?

H2. Changes in the degraded stream’s interaction scores track the direction of changes in its contribution to prediction quality.



Depth loses measurements step by step; thermal and laser stay unchanged.

Presented by
Andrii Baran (LPNU)

Depth benefit under degradation

H2. Changes in the degraded stream's interaction scores track the direction of changes in its contribution to prediction quality.

Person present in a fixed region in front of the robot?

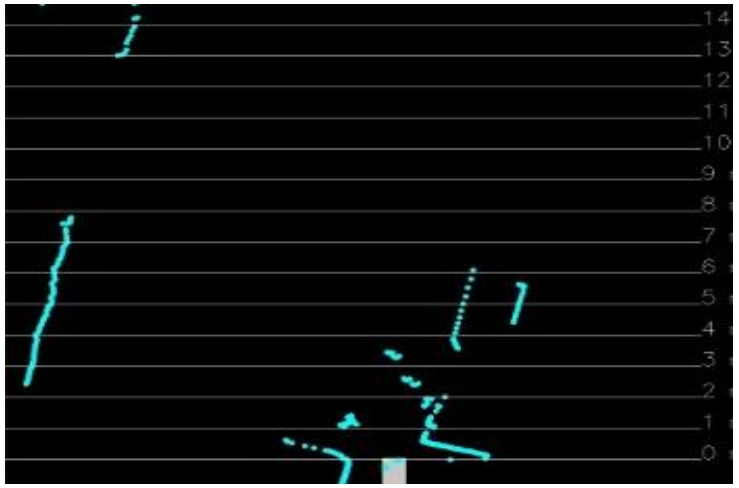
Thermal



ToF depth



2D laser from above



Studies of degraded sensor inputs

Study	What it shows	Use and limit for RQ2
MultiCorrupt IEEE IV 2024	Robustness depends on corruption type and fusion design.	Reuse failure types and severity levels. Adapt to ToF depth.
Liao et al. CVPRW 2025	Whole-input loss, partial loss and noise expose different weaknesses.	Keep failure types separate. Different task and sensor setup.
ADMN NeurIPS 2025	Adapts processing effort to input quality under a compute budget.	Useful motivation. Requires special training and a controller.

Explanations and sensor benefit

Study	What it shows	What this means for RQ2
Dr. SHAP-AV Interspeech 2026	Models shift toward video under noise but still rely on noisy audio.	A close example of following explanations under degradation.
InterSHAP AAAI 2025	Separates individual input effects from cross-modal interactions.	Provides interaction measures. Quality needs a separate check.
EMAP EMNLP 2020	Removing interactions can leave image–text performance almost unchanged.	Useful inputs do not always require strong interactions.

Resources for our study

Resource	What we can reuse / what needs work	Verdict
SEW dataset	Thermal, depth, laser and annotations. Define labels and recording groups.	Adapt
MultiCorrupt / Liao code	Use corruption examples. Adapt ToF formats and check reuse terms.	Adapt
shapiq	Compute interactions. Connect the model and define input replacement.	Adapt
ADMN / SWAN	Background on adapting computation. Their training procedures differ.	Cite only

The question still open for our study

Do depth interaction changes follow its measured prediction benefit?

What the literature provides

Corruption tests, explanation methods and examples of their limits.

Our comparison on SEW

Same frames and depth-loss levels. Measure benefit and interactions separately.

A preliminary test on SEW

01 Prepare the task

Person present in a fixed front region. Keep whole recordings together.

02 Reduce depth quality

Remove depth measurements in steps. Keep thermal and laser unchanged.

03 Measure both quantities

Prediction benefit against a thermal–laser model. Depth interaction scores.

04 Compare the changes

Do they follow the same direction? Does the replacement rule matter?

Contribution and team decisions

Compared with Dr. SHAP-AV, our study tests depth interactions against measured prediction benefit on SEW.

Decisions for the team

01 Shared recordings, task and model setup

02 Interaction definition with RQ1

03 Degradation settings with RQ3, then a small pilot

RQ3

When should a robot stop trusting a sensor?

Can a cheap run-time signal tell when to stop relying on a degraded stream?

Proposed sub-hypothesis

A single-pass quality score can identify harmful ToF input and support a useful fallback to thermal and laser.

Oluwasesan Cornelius Akinbode
Team 4 LM-RtA 2026



Confidence needs testing under degradation

Work	What it establishes	Implication for RQ3
Guo et al. ICML 2017 [3]	Temperature scaling improves probability calibration on held-out data.	Calibrate each branch before comparing confidence.
Ovadia et al. NeurIPS 2019 [4]	Post-hoc calibration can fail under distribution shift.	Evaluate calibration again at every corruption severity.
Aydın and Dizdaroğlu IDAP 2026 [8]	Sensor-fault comparisons can confound architecture with fault exposure during training.	Give every baseline the same clean-only training protocol.

A confident prediction can still be wrong. A useful gate needs evidence beyond confidence alone.

[3] Guo et al. 2017 [4] Ovadia et al. 2019 [8] Aydın and Dizdaroğlu 2026

Two starting points for adaptive fusion

Approach	Mechanism and evidence	Fit to our study
QMF Zhang et al. ICML 2023 [1]	Energy-based uncertainty weights the modalities. Its theoretical advantage depends on how weights align with prediction error.	Adapt the late-fusion idea. The experiments use RGB–depth and image–text classification, rather than our three robot streams.
Entropy-steered fusion Bijelic et al. CVPR 2020 [2]	Measurement entropy steers feature fusion. The study trains on clean data and evaluates adverse weather.	Adapt the quality-control principle. Measurement entropy differs from predictive class entropy, and the sensor architecture needs changing.

Neither result guarantees that our proposed cheap fallback will work on SEW.

[1] Zhang et al. 2023, QMF [2] Bijelic et al. 2020

Recent work adds stronger quality estimation

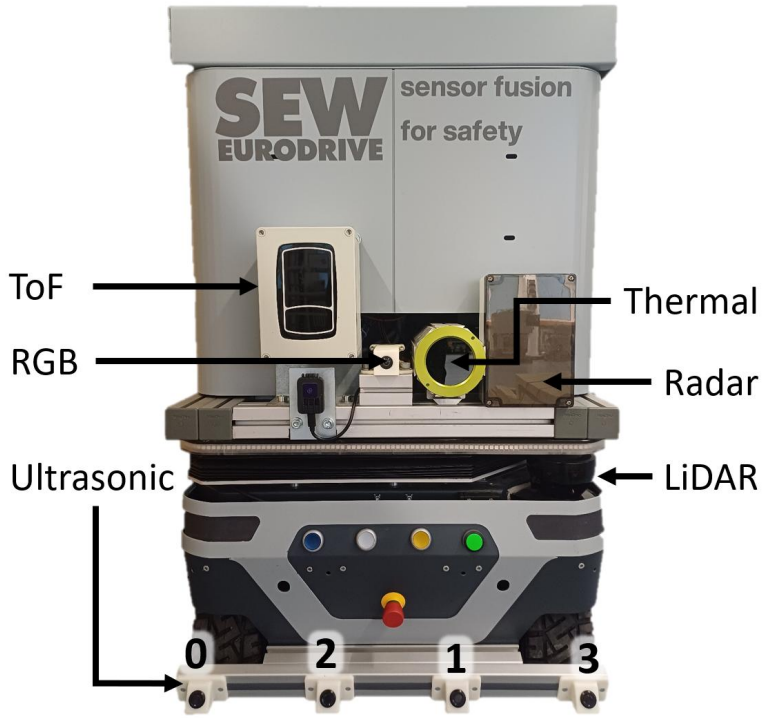
Work	Extension	Resource verdict
PDF ICML 2024 [5]	Combines modality and collaboration confidence with relative calibration.	Adapt if time permits. Public code targets image–text tasks.
ADMN NeurIPS 2025 [6]	Reallocates active layers across streams using input quality and compute budgets.	Cite only for the pilot. Requires layer-adaptive training and a controller.
UDML CVPR Findings 2026 [7]	Learns noise-aware uncertainty and corrects modality bias using noise injection and modality dropout during training.	Cite only in the clean-only pilot. Its fault-exposed training requires a separate comparison.

These methods motivate testing simpler scores under a fixed training protocol.

[5] Cao et al. 2024, PDF [6] Wu et al. 2025, ADMN [7] Wei et al. 2026, UDML

Resources for the SEW pilot

Resource	Verdict and use
Official SEW [9]	Adapt. Use thermal, ToF and laser data, labels and calibration.
Voxel51 conversion [10]	Use as-is for inspection. Test-only export. Keep it out of training and threshold selection.
PyTorch and TorchMetrics [12,13]	Use as-is as tools. Add our branch models, subset fusion and gate.
MultiCorrupt [11]	Cite only. Transfer severity controls and reimplement corruption for ToF.



SEW sensor bracket [9]

Data audit first: pin the release, frame IDs, labels and coordinate convention before combining exports.

[9] SEW-EURODRIVE dataset documentation [10] Voxel51 SEW conversion [12] TorchMetrics [13] PyTorch benchmarking

The gap is cheap detection of harmful input

What remains open for our configuration

Does a runtime score predict when excluding ToF improves the decision?

Noise detection and utility are different

A stream can look degraded and still add useful evidence. It can also look confident while making fusion worse.

Offline utility = loss without ToF − loss with ToF

Negative utility means ToF harms this prediction. Labels define this reference offline. The runtime gate never sees those labels.

Candidate signals and a fixed fallback rule

Signal	What we compute	When it is available
Invalid-depth fraction	Fraction of invalid ToF pixels in the fixed observation region.	Before the ToF encoder.
Predictive entropy	Uncertainty of the calibrated ToF class probability.	After the ToF encoder.
Disagreement	Difference between ToF and the thermal–laser prediction.	After the branches run.

Gate: exclude ToF when its score crosses a frozen threshold and the remaining streams agree confidently.

Compare each signal separately. Set all thresholds using clean validation only. A combined score is an optional extension.

A controlled clean-training experiment

1 Audit the task	Pilot person presence in a fixed common field of view. Verify negative examples and coordinate alignment.
2 Separate recordings	Hold out clean training sequences for calibration and policy selection. Reserve labeled official validation sequences for evaluation.
3 Fit and freeze	Train the branches on clean inputs. Freeze temperatures, fusion weights and gate thresholds before corruption tests.
4 Apply one corruption	Candidate for RQ2 agreement: ToF pixel dropout at 0%, 10%, 30% and 50%. Use the same frames and seeds for every method.
5 Evaluate all subsets	Three streams give seven nonempty subsets. Define missing-stream fusion explicitly. Use the empty subset only as an abstention/prior reference.

Proposed protocol. Corruption and sub-hypothesis need team agreement

Evidence that would support or reject the gate

Question	Measurement
Does the score detect harmful ToF input?	AUROC and average precision against negative offline utility. Compare average precision with harmful-case prevalence.
Does fallback improve predictions?	Macro-F1 and person recall by severity. Compare fixed fusion, every subset and gated fusion. Report the false-positive rate.
Do probabilities remain meaningful?	Negative log-likelihood, Brier score and expected calibration error at each severity.
Is the gate cheap in practice?	Batch-one end-to-end p50/p95 latency, gate overhead and memory on the same hardware.

Proposed targets: positive degraded-data gain, at most 1 percentage point clean macro-F1 loss, and at most 5% p95 overhead.

Proposed acceptance criteria. No experimental results are claimed

Proposed contribution and team handoff

Novelty for this study

Unlike QMF and UDML, we evaluate fixed, inexpensive fallback rules on SEW thermal, ToF and laser streams against ablation-defined utility under matched clean-only training, with measured runtime cost.

This is a proposed contribution for our configuration, rather than a claim that uncertainty-based fusion is new.

Owner	Connection to the shared study
Daniel RQ1	Compare interaction explanations with the subset-ablation reference.
Andrii RQ2	Agree on one corruption, its severity levels and missing-stream representation.
Cornelius RQ3	Test the score, calibration, fallback decision and measured overhead.

Discussion: agree on the hypothesis, data release, corruption and pilot budget

BACKUP

Backup slides

- References for the introduction and Q&A
- RQ2 method details B1–B3 (hidden in the slide show)
- RQ3 literature references

References for the introduction and Q&A

SEW-EURODRIVE. SEW Multimodal AMR Dataset 2026, documentation. GitHub.

Hooker, S., Erhan, D., Kindermans, P.-J., Kim, B. (2019). A Benchmark for Interpretability Methods in Deep Neural Networks. NeurIPS 32, 9737–9748.

Ovadia, Y., et al. (2019). Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift. NeurIPS 32.

Aydın, H., Dizdaroğlu, B. (2026). Per-Modality Uncertainty in Multimodal Sensor Fusion: Evidential Architecture versus Training-Time Fault Exposure. IDAP 2026, 1–9.

Other sources: Team 4 Joint Proposal (1 Nov 2026), the preliminary proposal slides and the RQ1–RQ3 member slides in this deck.

References for the literature review

[1] Zhang et al. 2023	Provable Dynamic Fusion for Low-Quality Multimodal Data. ICML.
[2] Bijelic et al. 2020	Seeing Through Fog Without Seeing Fog. CVPR.
[3] Guo et al. 2017	On Calibration of Modern Neural Networks. ICML.
[4] Ovadia et al. 2019	Can You Trust Your Model's Uncertainty? NeurIPS.
[5] Cao et al. 2024	Predictive Dynamic Fusion. ICML.
[6] Wu et al. 2025	ADMN: A Layer-Wise Adaptive Multimodal Network. NeurIPS main conference.
[7] Wei et al. 2026	Unbiased Dynamic Multimodal Fusion. CVPR Findings.
[8] Aydın and Dizdaroğlu 2026	Per-Modality Uncertainty in Multimodal Sensor Fusion. IDAP.

Titles are clickable. Full source URLs also appear in speaker notes