LEAR: EVIDENCE-AWARE LOCAL LEARNING WITH MISSING MODALITIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning with missing modalities requires a single model to predict from different observed subsets. Under block-local training, each module must learn from the available evidence without receiving gradients from later blocks. We introduce Layer-wise Evidence-Aware Representation Learning (LEAR), extending Forward-Forward-style local learning to prediction over every nonempty modality subset. LEAR turns the predictions of observed modality heads into a detached evidence reference for each sample. Within each block, this reference drives confidence-weighted distillation and coverage-bounded routing; across blocks, normalized detached representations preserve gradient locality. Evidence-Guided Inference (EGI) further combines fusion and available-evidence scores at prediction time. Compared with matched Local CE, LEAR improves mean-subset Macro-F1 by 0.91 percentage points on WISDM. On CREMA-D, the actor-averaged mean-subset Macro-F1 gain is 9.79 points, including a 3.08-point contribution from evidence-aware training beyond EGI alone. Improvements hold across subject-disjoint folds, modality cardinalities, and 18 of 19 held-out CREMA-D actors. These results establish available evidence as an effective learning signal for coordinating gradient-isolated local modules under changing multimodal inputs.

1 INTRODUCTION

Learning with missing modalities requires a model to make predictions from different observed subsets. Sensor failures, acquisition constraints, and sample-specific availability can change which inputs are accessible. A model that operates across these conditions must use the evidence present in each sample. We study this requirement under block-local optimization, where gradients from later blocks cannot update earlier ones. The central question is how locally trained modules can coordinate their learning when their available evidence changes.

Incomplete multimodal learning addresses missing inputs through robust fusion, modality reconstruction, knowledge distillation, and missingness-aware training (Ma et al., 2022; Mckinzie et al., 2023; Wang et al., 2024; Hu et al., 2024; Chen et al., 2024). Recent methods such as MARS and CoMiss further develop specialized architectures for arbitrary missing patterns (Baek et al., 2026; Mena et al., 2026). These approaches establish missingness as a training problem as well as an inference problem. Under the additional constraint of gradient isolation, each block needs a learning reference that reflects the modalities it can actually observe.

Local learning provides the setting for this coordination problem. Greedy layerwise training and gradient-isolated representation learning assign objectives to individual modules (Belilovsky et al., 2019; 2020; Löwe et al., 2019). Forward-Forward (FF) develops local goodness objectives, with later extensions improving competition, depth, and training mechanisms (Hinton, 2022; Papachristodoulou et al., 2024; Sun et al., 2025; Yao et al., 2026; Wu et al., 2024). Robustness across changing modality subsets adds a distinct requirement to these local objectives. An objective designed around fixed evidence need not remain aligned with the information available at deployment. We refer to this coordination problem as evidence mismatch.

We formulate the task using a single model that predicts for every nonempty modality subset. Training constructs availability states by masking aligned multimodal samples. Each block must therefore coordinate several views of the same prediction problem without relying on gradients from subsequent blocks. This formulation separates the availability of evidence from the locality of optimization.

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.

It also makes the learning target important: a reference constructed from the observed subset can guide a module using the information accessible in that state. Figure 1 illustrates the transition from fixed-evidence local learning to evidence-aware coordination.

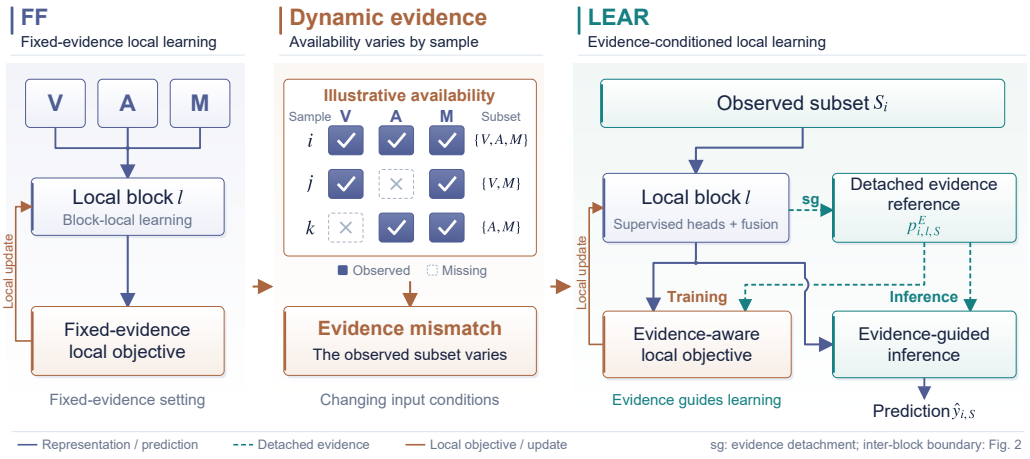

Figure 1: **From fixed-evidence Forward-Forward learning to evidence-aware local learning.** Fixed-evidence local learning encounters an evidence mismatch when the available subset S_i changes. LEAR turns the observed subset into a detached reference that coordinates block-local objectives and evidence-guided inference. V/A/M illustrate modality streams.

We introduce Layer-wise Evidence-Aware Representation Learning (LEAR) to provide this reference. For each sample and observed subset, LEAR aggregates the predictions of supervised modality heads and detaches the resulting evidence reference. The modality heads retain direct label supervision, while their detached predictions guide fusion learning. Confidence-weighted distillation limits the influence of uncertain references on incomplete-input states. Coverage-bounded routing allocates additional learning weight to subsets whose fusion predictions underperform their available evidence. A positive coverage floor retains objective weight for every enumerated subset. These operations adapt the local objective to the evidence state rather than requiring an unavailable full-input teacher.

Across blocks, normalized representations are passed through stop-gradient boundaries. LEAR thus preserves the inter-block locality principle of FF while using supervised cross-entropy and distillation within each block. We further introduce Evidence-Guided Inference (EGI), which combines fusion scores with observed-modality evidence before aggregation across blocks. Figure 2 shows these training and prediction paths. Matched Local CE, Local CE with EGI, and global counterparts distinguish the contribution of evidence-aware training from inference-time guidance and inter-block gradient propagation.

We evaluate LEAR on multi-sensor and audio-visual classification tasks. On WISDM, LEAR improves mean-subset Macro-F1 by 0.91 percentage points over matched Local CE, with positive gains in every subject-disjoint fold and modality cardinality. On CREMA-D, the actor-averaged mean-subset Macro-F1 gain reaches 9.79 points. Evidence-aware training contributes 3.08 points beyond EGI alone under the same inference rule. The overall improvement appears for 18 of 19 held-out actors. Additional PAMAP2 experiments extend the evaluation to another multi-sensor dataset, while MARS and CoMiss provide task-level comparisons with specialized incomplete-multimodal methods.

Our contributions are threefold:

- We formulate evidence mismatch as a coordination problem for block-local learning across nonempty modality subsets.
- We introduce LEAR, combining detached evidence references, confidence-weighted distillation, and coverage-bounded routing while preserving gradient locality.

- We introduce EGI and use matched evaluations to distinguish inference-time guidance from the additional contribution of evidence-aware training.

The resulting framework treats available evidence as a learning signal for coordinating local modules under changing multimodal inputs.

2 Related Work

Forward-Forward and layer-local learning. Greedy layerwise supervision scales local objectives to image classification, and decoupled greedy learning separates module updates while preserving local training targets (Belilovsky et al., 2019; 2020). Gradient-isolated representation learning provides another route to removing end-to-end credit assignment (Löwe et al., 2019). These precedents distinguish the broader local-learning constraint from a particular goodness loss. End-to-end backpropagation has dominated modern deep learning by enabling global credit assignment across the entire network. The Forward-Forward (FF) algorithm introduced an alternative learning paradigm by replacing global credit assignment with independently optimized layer-wise local objectives, allowing each layer to learn without relying on gradient propagation from other modules (Hinton, 2022). Subsequent studies have improved the effectiveness of FF-based local learning through stronger competitive objectives, deeper architectures, and enhanced training mechanisms (Papachristodoulou et al., 2024; Sun et al., 2025; Yao et al., 2026; Wu et al., 2024). These advances demonstrate the potential of local learning as an alternative to globally coupled optimization. These methods develop local objectives and architectures for their evaluated input settings; robustness across all multimodal availability subsets is a separate requirement. LEAR studies this requirement by introducing evidence-aware local objectives that adapt to dynamically changing evidence availability.

Incomplete multimodal learning. Incomplete multimodal learning investigates how models can maintain reliable predictions when some modalities are unavailable during inference. Existing approaches explore robust multimodal fusion, modality reconstruction, knowledge distillation, and missingness-aware optimization strategies (Ma et al., 2022; Mckinzie et al., 2023; Wang et al., 2024; Hu et al., 2024; Chen et al., 2024). MARS and CoMiss are recent strong baselines for this task, using specialized global architectures to improve robustness under arbitrary missing patterns (Baek et al., 2026; Mena et al., 2026). MARS uses privileged residuals to guide expert specialization during training, while CoMiss promotes inter-modal collaboration at feature and decision levels. LEAR advances this line of work under the additional constraint of block-local optimization: its local objectives adapt to changing evidence while preserving inter-block gradient isolation.

Evidence-aware learning and adaptive inference. Adaptive fusion and routing mechanisms have demonstrated that the reliability and availability of input information can guide prediction under changing environments. Existing methods mainly utilize available evidence to adjust inference strategies, such as modality selection, confidence weighting, or adaptive fusion. LEAR extends this principle beyond inference by treating available evidence as a reference for representation learning. This shift enables local modules to learn representations that better match the information accessible during deployment, providing an evidence-aware learning mechanism for dynamic multimodal environments.

3 Problem Formulation

A sample is $(x_i^1, \ldots, x_i^M, y_i)$, with $y_i \in \{1, \ldots, C\}$. At test time, a nonempty subset S is observed. Training constructs these availability states by masking aligned multimodal samples. We define

$$\mathcal{A} = \{S \subseteq \{1, \ldots, M\} : S \neq \emptyset\}, \qquad K = |\mathcal{A}| = 2^M - 1. \tag{1}$$

One model predicts for every $S \in \mathcal{A}$. Let $F_1(S)$ denote Macro-F1 over the evaluation set for state S. Mean-subset F1 is $K^{-1} \sum_S F_1(S)$, worst-subset F1 is $\min_S F_1(S)$, and full-input F1 is $F_1(S_{\text{full}})$. We also report

$$\textsc{AllSubsetsCorrect} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}[\forall S \in \mathcal{A},\ f(x_{i,S}) = y_i]\,. \tag{2}$$

The network contains L blocks. Block l produces modality representations $h_{i,l}^m$. It passes

$$\tilde{h}_{i,l}^m = \mathrm{sg}\big(\mathrm{Norm}(h_{i,l}^m)\big) \tag{3}$$

to the next block. A loss above this boundary cannot update block l. Each block must therefore coordinate its own availability states.

4 From Forward-Forward to Layer-wise Evidence-Aware Learning

LEAR extends the Forward-Forward paradigm from fixed-evidence local learning to evidence-conditioned local learning. For each available modality state, LEAR constructs a detached evidence reference from observed modalities and uses this reference to guide local optimization and inference. Unlike conventional local learning approaches that rely only on internal representations, LEAR allows each block to adapt its objective according to the evidence accessible to the current sample while preserving gradient isolation across blocks. An overview of the framework is shown in Figure 2.

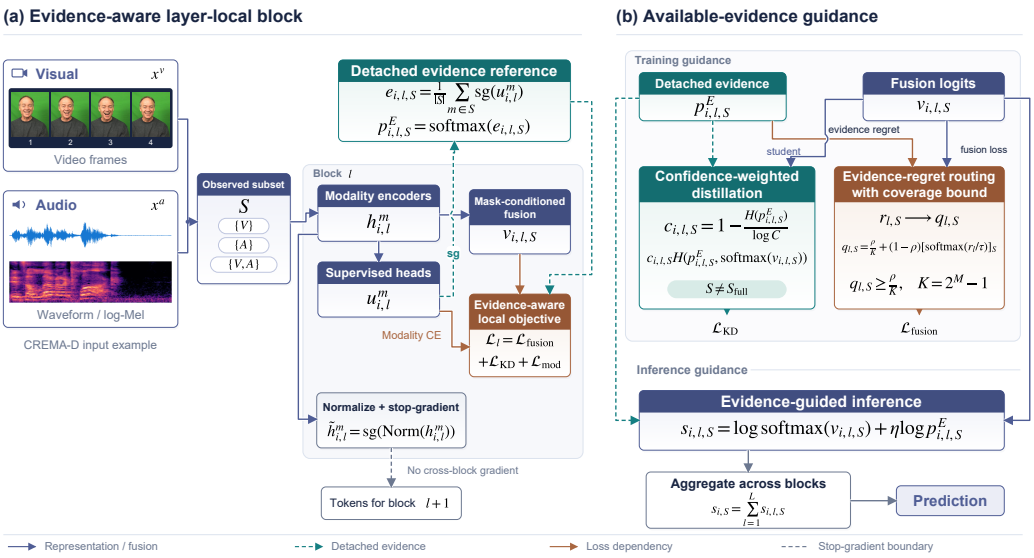

Figure 2: **Overview of LEAR.** (a) For each observed subset S, a layer-local block forms a detached reference from supervised modality heads, trains a mask-conditioned fusion prediction with the evidence-aware objective, and passes normalized detached tokens to the next block. (b) Available evidence weights distillation and evidence-regret routing during training, then guides inference by combining the detached reference with fusion logits before aggregation across blocks. Blue, green, and orange arrows denote representation and fusion flow, detached evidence, and local updates; the dashed gray path marks the stop-gradient boundary. Figure A4 gives the complete mechanism and loss-level dependencies.

4.1 Available-evidence reference

A supervised head for modality m outputs logits $u_{i,l}^m \in \mathbb{R}^C$. For an available modality subset S, LEAR aggregates the supervised predictions from observed modalities to construct an evidence reference:

$$e_{i,l,S} = \frac{1}{|S|} \sum_{m \in S} \mathrm{sg}(u_{i,l}^m), \qquad p_{i,l,S}^E = \mathrm{softmax}(e_{i,l,S}). \tag{4}$$

Only observed modalities contribute to this reference. The stop-gradient operation prevents the evidence reference from receiving updates from the fusion objective, turning $p_{i,l,S}^E$ into a detached learning target.

A shared mask-conditioned fusion head produces logits $v_{i,l,S}$ for each evidence state. Here B is the minibatch size, sg denotes stop-gradient, and Norm normalizes the inter-block representation. The corresponding fusion loss is

$$\ell^F_{i,l,S} = \mathrm{CE}(v_{i,l,S}, y_i). \tag{5}$$

For routing, we evaluate the evidence reference against the label:

$$\ell^E_{i,l,S} = \mathrm{CE}(e_{i,l,S}, y_i). \tag{6}$$

This detached loss is a diagnostic reference, not a trainable auxiliary CE term. Direct label supervision reaches the modality heads through $\mathcal{L}^{\mathrm{mod}}_l$. The loss difference measures fusion performance relative to available unimodal predictions.

4.2 Coverage-bounded routing

Different modality subsets may have different learning difficulties. To adapt training emphasis while preventing rare evidence states from being ignored, LEAR estimates the evidence regret of each state:

$$r_{l,S} = \frac{1}{B}\sum_{i=1}^{B}\left(\ell^F_{i,l,S} - \mathrm{sg}[\ell^E_{i,l,S}]\right). \tag{7}$$

A positive value indicates that the fusion prediction underperforms the evidence already provided by observed modalities.

The routing distribution combines adaptive emphasis with uniform state coverage:

$$q_{l,S} = \frac{\rho}{K} + (1-\rho)\frac{\exp(r_{l,S}/\tau)}{\sum_{T\in\mathcal{A}}\exp(r_{l,T}/\tau)}. \tag{8}$$

We set $\rho = 0.5$ and $\tau = 0.25$ in all experiments. Therefore, every evidence state receives at least $0.5/K$ objective weight while difficult states obtain additional optimization focus.

The fusion objective of block l is

$$\mathcal{L}^{\mathrm{fusion}}_l = \sum_{S\in\mathcal{A}} q_{l,S}\frac{1}{B}\sum_i \ell^F_{i,l,S}. \tag{9}$$

4.3 Confidence-weighted evidence distillation

Although the evidence reference provides useful supervision, unreliable evidence should have limited influence. LEAR therefore introduces confidence weighting according to prediction uncertainty:

$$c_{i,l,S} = 1 - \frac{H(p^E_{i,l,S})}{\log C}. \tag{10}$$

When the evidence distribution has high entropy, the confidence approaches zero.

The evidence distillation objective is defined as:

$$\mathcal{L}^{\mathrm{KD}}_l = \frac{1}{B(K-1)}\sum_i \sum_{S\neq S_{\mathrm{full}}} c_{i,l,S} H\left(p^E_{i,l,S}, \mathrm{softmax}(v_{i,l,S})\right). \tag{11}$$

The full-input state retains direct fusion supervision but receives no KD term. Incomplete states use their own observed-modality references; the full-input prediction is not their teacher. Here $H(p)$ is entropy and $H(p,q) = -\sum_c p_c \log q_c$ is cross-entropy.

Each modality head also receives direct label supervision, $\mathcal{L}_l^{\text{mod}} = (BM)^{-1} \sum_{i,m} \text{CE}(u_{i,l}^m, y_i)$. The final block objective combines fusion learning, modality supervision, and evidence distillation:

$$\mathcal{L}_l = \mathcal{L}_l^{\text{fusion}} + \mathcal{L}_l^{\text{KD}} + \mathcal{L}_l^{\text{mod}}. \tag{12}$$

4.4 Evidence-guided inference and gradient locality

During inference, LEAR combines fused prediction with available evidence. For evidence state S, block l produces:

$$s_{i,l,S} = \log \text{softmax}(v_{i,l,S}) + \eta \log p_{i,l,S}^E, \qquad \eta = 2. \tag{13}$$

We refer to this mechanism as *evidence-guided inference (EGI)*. The evidence term provides score guidance according to currently available modalities rather than performing probability calibration. We sum scores across blocks, $s_{i,S} = \sum_{l=1}^{L} s_{i,l,S}$, and predict $\arg\max_c s_{i,S,c}$.

LEAR inherits the inter-block locality principle of FF while using supervised CE and distillation within each block. This formulation combines ordinary block-local optimization with an evidence signal that adapts to the current modality subset. The representation passed between adjacent blocks is detached:

$$\tilde{h}_{i,l}^m = \text{sg}\left(\text{Norm}(h_{i,l}^m)\right). \tag{14}$$

Therefore, optimization within one block cannot update preceding blocks. Gradient analysis confirms zero cross-block gradients in the local formulation, while removing the detachment operation restores the expected gradient propagation. Algorithm 1 in Appendix B.1 summarizes the complete block-local training procedure.

5 Experiments

5.1 Experimental setup and evaluation protocols

We evaluate whether evidence-aware learning improves Forward-Forward style local optimization under incomplete evidence conditions. Our experiments are designed to answer three questions: (1) Can LEAR improve local learning when evidence availability changes across samples? (2) Does the proposed objective conditioned on evidence generalize to heterogeneous multimodal environments? (3) Which components contribute to the observed improvements?

WISDM (Weiss, 2019) serves as the primary controlled evaluation for verifying whether evidence-aware learning improves local optimization. Four inertial streams generate 15 nonempty evidence states. The dataset contains $15,803$ aligned ten-second windows collected from 50 complete subjects. We use five fixed subject-disjoint folds with $30/10/10$ train/validation/test subjects. Test subjects are never observed during training. For each fold, three random seeds are averaged, and the five folds are treated as the statistical evaluation units.

CREMA-D (Cao et al., 2014) evaluates LEAR in a heterogeneous audio-visual environment. The dataset contains audio and visual modalities collected from 91 actors. Actors are split into $63/9/19$ train/validation/test groups. The evaluation considers audio-only (A), visual-only (V), and complete audio-visual (AV) states. All matched methods use identical encoders, hidden dimensions, optimizers, batch sizes, and checkpoint selection criteria. The split was fixed before the final experiments.

We compare LEAR with several categories of baselines. Matched Local CE preserves the LEAR architecture but removes evidence-aware training and inference. Matched Global controls remove inter-block gradient isolation to evaluate the effect of locality. WISDM additionally includes conventional backpropagation subset averaging, Group DRO (Sagawa et al., 2020), and modality dropout. MARS (Baek et al., 2026) and CoMiss (Mena et al., 2026) serve as recent strong baselines for incomplete multimodal learning, providing current task-level comparisons against specialized global multimodal architectures. All adaptation details are provided in Appendix H; dataset and matched-model protocols appear in Appendices D and G.

Table 1: **WISDM comparison under subject-disjoint evaluation.** Results average five held-out-subject folds and three seeds per fold (%). ↑ indicates higher is better; All-corr. is all-subsets-correct accuracy. The shaded row is LEAR; bold marks the best result in each column. †Recent strong baseline using its native validation selector.

Method	Mean ↑	Worst ↑	Full ↑	All-corr. ↑
Matched block-local models				
Local CE	56.35	24.56	69.01	8.24
Local LEAR (ours)	57.27	25.21	69.50	8.81
Matched end-to-end controls				
Global CE	56.65	24.32	69.69	8.05
Global LEAR	57.44	**25.31**	**69.72**	**8.87**
Recent strong incomplete-multimodal baseline				
CoMiss†	**58.28**	24.54	–	–

Table 2: **CREMA-D comparison under actor-disjoint evaluation.** Results average three seeds (%); A/V/AV denote audio-only, visual-only, and complete inputs. The shaded row is LEAR; bold marks the best result in each column. †Recent strong baseline adapted to the same protocol.

Method	Mean ↑	Worst ↑	A ↑	V ↑	AV ↑
Matched block-local models					
Local CE	44.67	23.59	52.55	23.59	57.87
Local CE+EGI	51.09	32.12	56.35	32.12	64.80
Local LEAR (ours)	**54.33**	**42.41**	53.66	**42.41**	66.91
Matched end-to-end controls					
Global CE	53.55	37.54	**55.53**	37.54	67.58
Global CE+EGI	53.28	40.70	52.53	40.70	66.62
Global LEAR	54.31	42.26	52.58	42.26	**68.10**
Recent strong incomplete-multimodal baseline					
MARS†	45.69	30.62	53.42	30.62	53.02

5.2 Controlled evaluation of evidence-aware local learning on WISDM

Table 1 reports results under the controlled subject-disjoint evaluation. The comparison between Local CE and LEAR measures the combined training and inference change. CE+EGI controls isolate the additional training contribution; local/global controls isolate inter-block detachment.

LEAR improves Local CE by $+0.91\,pp$ in mean-subset Macro-F1 and by $+0.65\,pp$ on the worst subset. The gain appears in every subject-disjoint fold, showing that evidence-aware coordination improves both average performance and the most difficult availability states.

The improvement appears across different evidence cardinalities. The gains are $+0.88$, $+1.04$, $+0.87$, and $+0.48\,pp$ when one, two, three, and four modalities are available, respectively. Figure A1 shows that the mean gain remains positive at each cardinality.

Local LEAR reaches 57.27 mean-subset F1, close to the 57.44 achieved by its globally optimized counterpart, while preserving block-local gradients. The recent strong baseline CoMiss reaches 58.28 with its specialized architecture and native validation selector, defining the leading task-level reference on WISDM.

5.3 Generalization to heterogeneous multimodal environments

CREMA-D provides a more challenging setting where available evidence differs substantially between audio and visual modalities. Table 2 summarizes the results.

The heterogeneous audio-visual setting reveals a larger benefit of evidence-aware local learning. Local LEAR improves the pooled mean-subset score by 9.66 points and the actor-averaged score by 9.79 points over Local CE. The actor-level gain appears for 18 of 19 held-out actors.

EGI alone contributes $+6.71\,pp$, showing that available evidence is a strong inference signal. Evidence-aware training adds a further $+3.08\,pp$ under the same EGI rule, separating the representation-learning contribution from inference-time adaptation.

Local LEAR and Global LEAR achieve nearly identical actor-averaged performance, with a $+0.21\,pp$ difference in favor of the local model. Actor-level effects are detailed in Figure A2.

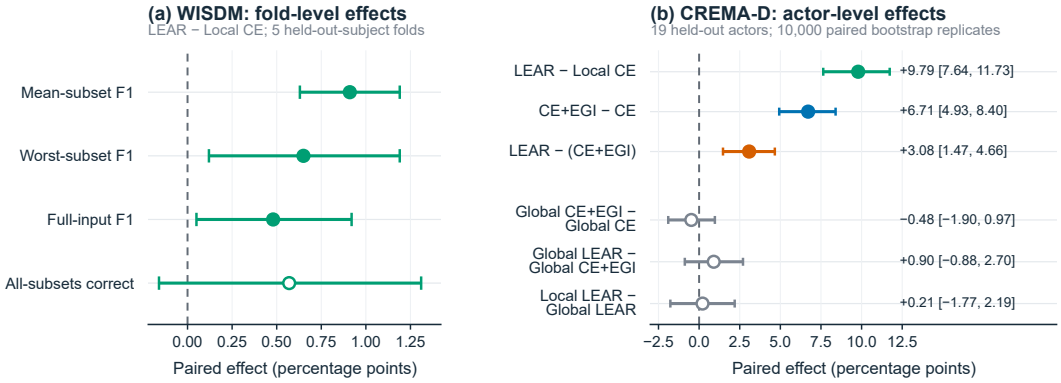

Figure 3: **Paired effects of evidence-aware learning.** (a) Differences between LEAR and Local CE across WISDM endpoints. (b) CREMA-D contrasts averaged across actors, separating EGI used alone at inference from the additional LEAR training effect. Markers show mean paired effects and lines show 95% intervals; filled markers indicate intervals above zero.

5.4 Ablation Studies and Analysis

We analyze the contribution of individual LEAR components on WISDM. EGI alone improves Local CE by $+0.60\,pp$. Available-evidence knowledge distillation increases the gain to $+0.82\,pp$, compared with $+0.54\,pp$ from a full-input teacher.

Under matched training protocols, subset-matched available evidence provides a stronger distillation target than the full-input teacher. The validation specificity study in Figure A3 further compares alternative score sources. The complete LEAR objective achieves $+0.91\,pp$ over Local CE. Coverage-bounded routing adds $+0.09\,pp$ beyond evidence distillation while ensuring that every enumerated state retains positive objective weight. The validation grid in Appendix I shows stable performance across the tested coverage and temperature settings.

5.5 Additional Cross-Dataset Generalization

Additional evaluations provide insight into the empirical scope of LEAR. On PAMAP2, LEAR improves every reported endpoint for all four held-out subjects, extending the observed benefits to a second multi-sensor activity dataset.

6 Discussion

The matched experiments reveal a consistent pattern: available evidence becomes valuable when it is explicitly incorporated into the learning process of local modules. Rather than treating missing modalities only as an inference-time challenge, LEAR aligns local objectives with the evidence that each block can actually access. WISDM demonstrates that evidence-guided inference already provides stable improvements under controlled evidence variation, while CREMA-D further reveals additional gains from incorporating evidence into local training. This difference is consistent with the stronger modality imbalance in CREMA-D, where visual-only evidence provides a more challenging learning condition.

The component results make this two-stage role of evidence explicit. On WISDM, EGI contributes $+0.60\,pp$, evidence distillation raises the gain to $+0.82\,pp$, and coverage-bounded routing brings the complete improvement to $+0.91\,pp$. On CREMA-D, EGI contributes $+6.71\,pp$ and evidence-aware

training adds a further $+3.08\,pp$. The same reference thus serves as a decision-time cue and as a target for shaping representations. Its effect grows when the observed modalities differ more sharply in reliability, matching the contrast between the two datasets.

The gains also extend beyond a favorable average. LEAR improves every WISDM fold, every evaluated modality cardinality, and 18 of 19 held-out CREMA-D actors. This consistency shows that evidence conditioning changes how the model handles the full set of observed-input states rather than concentrating improvement in one subset or subject group. The paired analyses in Figure 3 and the actor-level results in Figure A2 further connect the aggregate gains to the underlying evaluation units.

The local/global comparisons isolate the role of gradient locality. Across WISDM and CREMA-D, block-local LEAR closely matches the corresponding globally optimized model. Evidence-aware supervision therefore recovers the coordination needed by local modules without cross-block credit assignment.

This result suggests a broader design principle for local learning. A block does not need access to hidden modalities or gradients from later blocks to respond to changing inputs. It needs a target that summarizes the evidence currently available and remains stable during its update. LEAR supplies that target through detached modality-head predictions, then allocates fusion supervision with coverage-bounded routing. The construction keeps optimization local while allowing each block to adapt its objective to the sample-specific evidence state.

The comparisons against recent strong task baselines position LEAR within current incomplete multimodal learning. CoMiss achieves the strongest WISDM mean, while LEAR exceeds the adapted MARS baseline on CREMA-D. LEAR adds a distinct capability to this competitive landscape: it equips gradient-isolated local modules with a learning signal derived from the evidence they can observe.

7 Conclusion

We introduced LEAR, an evidence-conditioned extension of Forward-Forward local learning for dynamic multimodal environments. LEAR incorporates available evidence into supervised block-local objectives while preserving cross-block gradient isolation. The proposed framework constructs detached evidence references to guide local optimization and evidence-guided inference without requiring hidden modalities.

Experiments on WISDM demonstrate consistent improvements over matched Local CE under subject-disjoint evaluation, supporting the combined evidence-aware training and inference procedure. CREMA-D further shows that evidence-aware training provides additional improvements beyond inference-only adaptation in heterogeneous multimodal environments. Together with ablation studies and external comparisons, these results demonstrate that available evidence can serve as a principled mechanism for coordinating local learning under changing information conditions.

Reproducibility Statement

The supplement records configurations, data splits, protocol adaptations, fold- and actor-level statistics, and figure source data. The total block objective is the unscaled sum of the three component losses, as specified in both the method and appendix. WISDM uses train-only normalization and fold-level inference. CREMA-D uses a locked actor-disjoint split and actor-cluster bootstrap evaluation. All reported results were generated from the recorded experimental configurations.

AI Use Statement

Generative AI tools were used only as auxiliary tools for manuscript preparation and technical assistance. They supported English writing refinement, LaTeX formatting, code organization and debugging, and figure preparation.

REFERENCES

Seunghun Baek, Jihwan Park, Jaeyoon Sim, Minjae Jeong, Hoseok Lee, and Won Hwa Kim. Residual-guided expert specialization for incomplete multimodal learning. *arXiv preprint arXiv:2606.30355*, 2026. URL https://arxiv.org/abs/2606.30355.

Eugene Belilovsky, Michael Eickenberg, and Edouard Oyallon. Greedy layerwise learning can scale to imagenet. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 583–593, 2019.

Eugene Belilovsky, Michael Eickenberg, and Edouard Oyallon. Decoupled greedy learning of cnns. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 736–745, 2020.

Houwei Cao, David G. Cooper, Michael K. Keutmann, Ruben C. Gur, Ani Nenkova, and Ragini Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on Affective Computing*, 5(4):377–390, 2014. doi: 10.1109/TAFFC.2014.2336244.

Mengxi Chen, Fei Zhang, Zihua Zhao, Jiangchao Yao, Ya Zhang, and Yanfeng Wang. Probabilistic conformal distillation for enhancing missing modality robustness. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Geoffrey Hinton. The forward-forward algorithm: Some preliminary investigations. *arXiv preprint arXiv:2212.13345*, 2022. URL https://arxiv.org/abs/2212.13345.

Lianyu Hu, Tongkai Shi, Wei Feng, Fanhua Shang, and Liang Wan. Deep correlated prompting for visual recognition with missing modalities. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Sindy Löwe, Peter O'Connor, and Bastiaan S. Veeling. Putting an end to end-to-end: Gradient-isolated learning of representations. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

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*, pp. 18177–18186, 2022.

Brandon Mckinzie, Vaishaal Shankar, Joseph Yitan Cheng, Yinfei Yang, Jonathon Shlens, and Alexander T. Toshev. Robustness in multimodal learning under train-test modality mismatch. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 24291–24303, 2023.

Francisco Mena, Dino Ienco, Roberto Interdonato, Cassio F. Dantas, and Simon Besnard. Co-learning for missing arbitrary modalities in multi-modal classification. *arXiv preprint arXiv:2607.24683*, 2026. URL https://arxiv.org/abs/2607.24683.

Andreas Papachristodoulou, Christos Kyrkou, Stelios Timotheou, and Theocharis Theocharides. Convolutional channel-wise competitive learning for the forward-forward algorithm. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024. URL https://arxiv.org/abs/2312.12668.

Attila Reiss. Pamap2 physical activity monitoring. UCI Machine Learning Repository, 2012.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations*, 2020.

Liang Sun, Yang Zhang, Weizhao He, Jiajun Wen, Linlin Shen, and Weicheng Xie. DeeperForward: Enhanced forward-forward training for deeper and better performance. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=kOYnXVQCtA.

Hao Wang, Shengda Luo, Guosheng Hu, and Jianguo Zhang. Gradient-guided modality decoupling for missing-modality robustness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 15483–15491, 2024. doi: 10.1609/aaai.v38i14.29474.

Gary M. Weiss. Wisdm smartphone and smartwatch activity and biometrics dataset. UCI Machine Learning Repository, 2019.

Yujie Wu, Siyuan Xu, Jibin Wu, Lei Deng, Mingkun Xu, Qinghao Wen, and Guoqi Li. Distance-forward learning: Enhancing the forward-forward algorithm towards high-performance on-chip learning. *arXiv preprint arXiv:2408.14925*, 2024. URL https://arxiv.org/abs/2408.14925.

Jie-En Yao, Hong-En Chen, and C.-C. Jay Kuo. HCL-FF: Hierarchical and contrastive learning for forward-forward algorithm. *arXiv preprint arXiv:2605.24797*, 2026. URL https://arxiv.org/abs/2605.24797.

A Appendix Overview and Evaluation Protocols

The appendix provides the complete LEAR formulation, implementation details, dataset construction, seed-level and fold-level results, component analyses, strong baseline adaptations, additional dataset results, and efficiency measurements. Table A1 summarizes the role of each experimental suite.

Table A1: **Appendix evaluation map.** Datasets, availability states, and the role of each study.

Study	Inputs / states	Evaluation role
WISDM local suite	4 inertial / 15 states	Primary layer-local evaluation
WISDM global suite	same	Matched gradient-locality control
CREMA-D matched suite	audio, visual / 3 states	Heterogeneous multimodal evaluation
MARS	audio, visual / 3 states	Recent strong CREMA-D baseline
CoMiss	4 inertial / 15 states	Recent strong WISDM baseline
PAMAP2	3 inertial / 7 states	Additional cross-dataset evaluation
Specificity / sensitivity	WISDM validation	Mechanism diagnostics
Efficiency	WISDM	Training and inference cost

For WISDM, three seeds are averaged within each held-out subject fold before summaries across folds. For CREMA-D, results are reported across three seeds on a fixed actor-disjoint split. External methods are used as performance references because their architectures and learning mechanisms differ from the matched CE/LEAR family.

B Complete Method Specification

For block l, modality m produces supervised logits $u^m_{i,l}$. For an available modality subset S, the evidence logits and evidence distribution are

$$e_{i,l,S} = \frac{1}{|S|} \sum_{m \in S} \mathrm{sg}\big(u^m_{i,l}\big)\,, \qquad p^E_{i,l,S} = \mathrm{softmax}(e_{i,l,S}). \tag{15}$$

Fusion logits $v_{i,l,S}$ use

$$\ell^F_{i,l,S} = \mathrm{CE}(v_{i,l,S}, y_i), \qquad \ell^E_{i,l,S} = \mathrm{CE}(e_{i,l,S}, y_i). \tag{16}$$

The minibatch evidence regret is

$$r_{l,S} = \frac{1}{B} \sum_i \big(\ell^F_{i,l,S} - \mathrm{sg}[\ell^E_{i,l,S}]\big)\,. \tag{17}$$

The routing distribution is

$$q_{l,S} = \frac{\rho}{K} + (1-\rho)\,\mathrm{softmax}(r_l/\tau)_S, \tag{18}$$

with $\rho = 0.5$ and $\tau = 0.25$. Thus $q_{l,S} \geq 0.5/K$ and $\sum_S q_{l,S} = 1$. The routed fusion term is

$$\mathcal{L}^{\mathrm{fusion}}_l = \sum_{S \in \mathcal{A}} q_{l,S} \frac{1}{B} \sum_i \ell^F_{i,l,S}. \tag{19}$$

The evidence confidence is

$$c_{i,l,S} = 1 - \frac{H(p^E_{i,l,S})}{\log C}. \tag{20}$$

The confidence-weighted evidence distillation term excludes the full-input state:

$$\mathcal{L}^{\mathrm{KD}}_l = \frac{1}{B(K-1)} \sum_i \sum_{S \neq S_{\mathrm{full}}} c_{i,l,S}\, H\big(p^E_{i,l,S}, \mathrm{softmax}(v_{i,l,S})\big)\,. \tag{21}$$

Each modality head receives direct supervision,

$$\mathcal{L}_l^{\mathrm{mod}} = \frac{1}{BM} \sum_{i,m} \mathrm{CE}(u_{i,l}^m, y_i). \tag{22}$$

The block objective combines fusion learning, modality supervision, and evidence distillation:

$$\mathcal{L}_l = \mathcal{L}_l^{\mathrm{fusion}} + \mathcal{L}_l^{\mathrm{KD}} + \mathcal{L}_l^{\mathrm{mod}}. \tag{23}$$

The representation passed to the next block is normalized and detached. During inference,

$$s_{i,l,S} = \log \mathrm{softmax}(v_{i,l,S}) + 2 \log p_{i,l,S}^E, \qquad s_{i,S} = \sum_{l=1}^{L} s_{i,l,S}. \tag{24}$$

All real-task experiments use three blocks and enumerate every nonempty availability subset.

B.1 Training Procedure

A training step consists of five operations: (1) encode the observed modalities in the current block; (2) produce supervised modality logits; (3) enumerate admissible subsets and evaluate their fusion paths; (4) detach available-evidence logits before routing and distillation; and (5) optimize the current block locally, then normalize and detach its representations before passing them upward. Reverse-mode differentiation is therefore standard within a block, while gradients do not propagate across block boundaries.

Algorithm 1 Layer-wise Evidence-Aware Representation Learning (LEAR)

Require: Minibatch $\{(x_i^1, \ldots, x_i^M, y_i)\}_{i=1}^B$; L blocks; C classes
Require: Nonempty modality subsets $\mathcal{A}$; routing parameters ρ, τ
$K \leftarrow |\mathcal{A}|$; $S_{\mathrm{full}} \leftarrow \{1, \ldots, M\}$
$\tilde{h}_{i,0}^m \leftarrow x_i^m$ for all samples i and modalities m
for $l = 1, \ldots, L$ **do**
Local representations and modality supervision
  for $m = 1, \ldots, M$ **do**
    $h_{i,l}^m \leftarrow \mathrm{Encoder}_l^m(\tilde{h}_{i,l-1}^m)$ for all i
    $u_{i,l}^m \leftarrow \mathrm{Head}_l^m(h_{i,l}^m)$ for all i
  end for
  $\mathcal{L}_l^{\mathrm{mod}} \leftarrow \frac{1}{BM} \sum_{i=1}^B \sum_{m=1}^M \mathrm{CE}(u_{i,l}^m, y_i)$
Subset-specific evidence and fusion predictions
  for $S \in \mathcal{A}$ **do**
    $e_{i,l,S} \leftarrow \frac{1}{|S|} \sum_{m \in S} \mathrm{sg}(u_{i,l}^m)$; $\quad p_{i,l,S}^E \leftarrow \mathrm{softmax}(e_{i,l,S})$ for all i
    $v_{i,l,S} \leftarrow \mathrm{Fusion}_l(\{h_{i,l}^m : m \in S\}, S)$ for all i
    $\ell_{i,l,S}^F \leftarrow \mathrm{CE}(v_{i,l,S}, y_i)$; $\quad \ell_{i,l,S}^E \leftarrow \mathrm{CE}(e_{i,l,S}, y_i)$ for all i
    $r_{l,S} \leftarrow \frac{1}{B} \sum_{i=1}^B \big(\ell_{i,l,S}^F - \mathrm{sg}(\ell_{i,l,S}^E)\big)$
  end for
Coverage-bounded routing
  $q_{l,S} \leftarrow \frac{\rho}{K} + (1-\rho) \frac{\exp(r_{l,S}/\tau)}{\sum_{T \in \mathcal{A}} \exp(r_{l,T}/\tau)}$ for all $S \in \mathcal{A}$
  $\mathcal{L}_l^{\mathrm{fusion}} \leftarrow \sum_{S \in \mathcal{A}} q_{l,S} \frac{1}{B} \sum_{i=1}^B \ell_{i,l,S}^F$
Confidence-weighted evidence distillation
  for $S \in \mathcal{A} \setminus \{S_{\mathrm{full}}\}$ **do**
    $c_{i,l,S} \leftarrow 1 - \frac{H(p_{i,l,S}^E)}{\log C}$ for all i
  end for
  $\mathcal{L}_l^{\mathrm{KD}} \leftarrow \frac{1}{B(K-1)} \sum_{i=1}^B \sum_{S \in \mathcal{A} \setminus \{S_{\mathrm{full}}\}} c_{i,l,S}\, H\big(p_{i,l,S}^E, \mathrm{softmax}(v_{i,l,S})\big)$
Block-local optimization and gradient isolation
  $\mathcal{L}_l \leftarrow \mathcal{L}_l^{\mathrm{fusion}} + \mathcal{L}_l^{\mathrm{KD}} + \mathcal{L}_l^{\mathrm{mod}}$
  Update only block-l parameters θ_l using $\nabla_{\theta_l} \mathcal{L}_l$
  $\tilde{h}_{i,l}^m \leftarrow \mathrm{sg}\big(\mathrm{Norm}(h_{i,l}^m)\big)$ for all i, m
end for

B.2 Useful Properties

The coverage floor prevents any availability state from losing all training mass. Uniform evidence contributes no KD signal because $c = 0$. The full-input state receives direct fusion supervision but no evidence-KD term. Every evidence reference is constructed only from observed modalities.

C Gradient Locality

The local/global comparison changes only the inter-block detach operation. The local implementation passes $\mathrm{sg}(\mathrm{Norm}(h_l))$ to the next block, whereas the matched global implementation passes the corresponding representation without stop-gradient. Gradient checks confirm that upper-block losses do not update lower blocks in the local model and that removing the detach restores cross-block gradients.

D WISDM Data Construction and Availability States

WISDM (Weiss, 2019) uses the official Smartphone and Smartwatch Activity and Biometrics dataset. The prepared benchmark contains 15,803 aligned ten-second windows from 50 complete subjects and 18 activities. Four streams are treated as modalities: phone accelerometer (PA), phone gyroscope (PG), watch accelerometer (WA), and watch gyroscope (WG). Their 15 nonempty subsets are listed in Table A2.

Table A2: **WISDM availability space.** All 15 nonempty subsets of the four inertial streams.

Cardinality	States
1	PA; PG; WA; WG
2	PA+PG; PA+WA; PA+WG; PG+WA; PG+WG; WA+WG
3	PA+PG+WA; PA+PG+WG; PA+WA+WG; PG+WA+WG
4	PA+PG+WA+WG

Alignment uses subject, activity, and within-activity window index. Normalization statistics are fitted on training subjects only. Five fixed folds use 30 training, 10 validation, and 10 test subjects. Test subjects are mutually disjoint across folds, and the three seeds within a fold share the same split and preprocessing.

E WISDM Results by Seed and Fold

Table A3 reports the complete Local CE and LEAR seed grid. After averaging seeds, LEAR improves Local CE in every held-out-subject fold.

Table A3: **WISDM results by fold and seed.** Local CE and LEAR results are reported in percent.

Fold	Seed	CE mean	LEAR mean	Δ	CE full	LEAR full
0	0	62.98	63.85	0.87	75.46	75.77
0	1	62.28	63.15	0.87	74.96	75.80
0	2	62.80	63.27	0.47	77.13	76.30
1	0	55.85	56.15	0.30	68.86	68.22
1	1	55.78	56.61	0.83	68.19	69.59
1	2	56.08	56.84	0.76	68.31	68.54
2	0	52.67	53.87	1.20	65.60	65.38
2	1	52.23	53.32	1.08	65.27	65.80
2	2	53.49	54.72	1.23	66.35	67.02
3	0	55.21	56.48	1.27	66.65	68.54
3	1	55.29	56.07	0.78	68.29	67.57
3	2	55.60	56.71	1.10	67.32	68.10
4	0	55.35	56.17	0.82	67.49	68.31
4	1	55.13	56.23	1.10	67.44	69.22
4	2	54.55	55.58	1.03	67.85	68.27

Table A4: **WISDM effects across folds.** Subset average Macro-F1 after averaging the three seeds within each fold.

Fold	Local CE	LEAR	CoMiss	LEAR−CE	CoMiss−LEAR
0	62.69	63.43	64.74	0.74	1.32
1	55.90	56.53	57.41	0.63	0.88
2	52.80	53.97	54.74	1.17	0.77
3	55.37	56.42	57.05	1.05	0.64
4	55.01	55.99	57.45	0.98	1.46

F WISDM Baselines, Cardinality, and Component Analysis

Table A5 summarizes the complete WISDM baseline comparison using fold means after averaging seeds.

Table A5: **WISDM baseline comparison.** Aggregate results in percent; ↑ indicates higher is better.

Method	Params	Mean ↑	Worst ↑	Full ↑	All-corr. ↑
Local CE	412,558	56.35	24.56	69.01	8.24
LEAR (ours)	412,558	57.27	25.21	69.50	8.81
BP subset mean	444,306	54.56	24.19	65.36	7.46
BP Group DRO	444,306	48.98	24.36	57.16	6.09
BP modality dropout	444,306	53.71	22.96	65.69	6.26
Global CE	412,558	56.65	24.32	69.69	8.05
Global LEAR	412,558	57.44	25.31	69.72	8.87

The gains of LEAR over Local CE are +0.88, +1.04, +0.87, and +0.48 percentage points when one, two, three, and four modalities are available, respectively. Figure A1 shows positive gains at every availability cardinality.

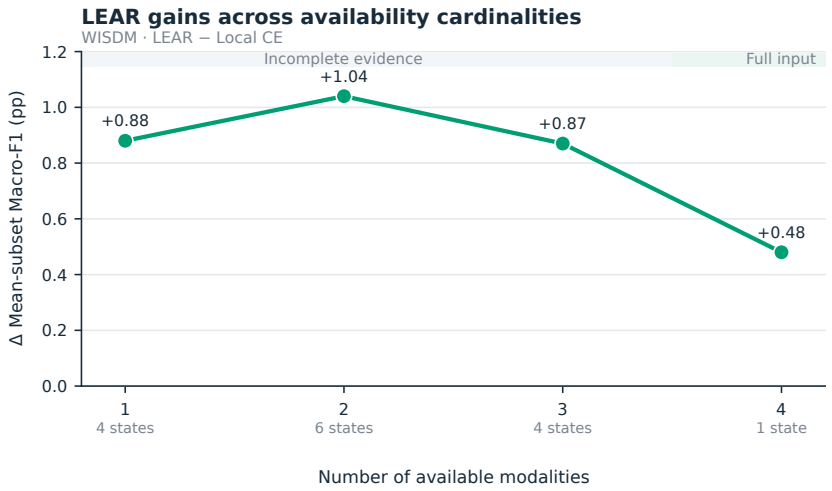

Figure A1: **WISDM gains across availability cardinalities.** Differences in subset average Macro-F1 between LEAR and Local CE after averaging seeds and folds. Cardinalities 1–4 contain 4, 6, 4, and 1 availability states.

Table A6: **WISDM component analysis.** Matched local variants in percent; ↑ indicates higher is better.

Method	Mean ↑	Worst ↑	Full ↑	All-corr. ↑	Δ mean
CE	56.35	24.56	69.01	8.24	0.00
CE + EGI	56.95	24.81	69.84	8.53	+0.60
Full-teacher KD + EGI	56.89	25.08	69.69	8.59	+0.54
Available-evidence KD + EGI	57.18	24.99	69.55	8.58	+0.82
Availability routing + EGI	56.85	25.10	69.62	8.60	+0.49
LEAR (ours)	**57.27**	**25.21**	69.50	**8.81**	**+0.91**

EGI provides the first improvement over Local CE, and available-evidence KD raises the gain further. The complete objective adds coverage-bounded routing and reaches the best mean-subset result among the matched local variants. Direct mean contrasts are +0.09 points for LEAR over available-evidence KD+EGI, +0.42 for LEAR over routing+EGI, +0.23 for available-evidence KD+EGI over CE+EGI, and -0.06 for full-teacher KD+EGI over CE+EGI.

G CREMA-D Protocol and Results

CREMA-D (Cao et al., 2014) contains 7,442 clips from 91 actors. Actor identities follow a fixed, reproducible permutation established before the final experiments, producing 63 training actors, 9 validation actors, and 19 test actors. Labels are ANG, DIS, FEA, HAP, NEU, and SAD. Audio uses the official AudioWAV files. Visual input uses one-frame-per-second extraction following the MARS pipeline.

The six matched methods share the same encoders, width 128, three-block topology, fusion head, and modality heads. They use AdamW with learning rate 10^{-3}, weight decay 10^{-4}, batch size 64, a maximum of 50 epochs, and patience 8. Checkpoints are selected by validation mean-subset Macro-F1. Seeds are 0, 1, and 2.

The validation actors are 1015, 1079, 1020, 1030, 1050, 1089, 1090, 1070, and 1021. The test actors are 1086, 1073, 1026, 1038, 1047, 1040, 1066, 1059, 1013, 1071, 1037, 1022, 1084, 1010, 1088, 1068, 1065, 1048, and 1045. The remaining 63 actors form the training split.

Table A7: **CREMA-D results by seed.** A/V/AV denote audio, visual, and complete inputs; values are percentages.

Method	Seed	Mean ↑	Worst ↑	A ↑	V ↑	AV ↑	All-corr. ↑
Local CE	0	37.51	13.27	49.43	13.27	49.82	9.63
Local CE	1	39.81	13.71	52.27	13.71	53.45	11.87
Local CE	2	56.70	43.80	55.97	43.80	70.34	27.92
Local CE+EGI	0	40.57	4.86	58.43	4.86	58.43	10.08
Local CE+EGI	1	54.74	45.04	54.18	45.04	64.99	24.58
Local CE+EGI	2	57.96	46.47	56.43	46.47	70.99	25.61
Local LEAR	0	53.08	42.55	54.64	42.55	62.05	23.68
Local LEAR	1	52.23	39.03	51.21	39.03	66.46	23.56
Local LEAR	2	57.67	45.65	55.13	45.65	72.22	27.41
Global CE	0	48.73	24.09	56.66	24.09	65.43	16.75
Global CE	1	54.66	42.94	55.53	42.94	65.51	26.64
Global CE	2	57.26	45.59	54.40	45.59	71.78	23.04
Global CE+EGI	0	56.05	43.46	55.28	43.46	69.40	25.48
Global CE+EGI	1	53.44	40.30	52.64	40.30	67.39	24.33
Global CE+EGI	2	50.36	38.35	49.66	38.35	63.06	21.18
Global LEAR	0	55.03	41.61	52.37	41.61	71.12	23.43
Global LEAR	1	55.76	42.22	57.70	42.22	67.36	29.65
Global LEAR	2	52.15	42.96	47.68	42.96	65.82	18.87

Table A8: **CREMA-D paired effects by actor.** Effects after averaging three seeds, in percentage points.

Actor	LEAR−CE	CE+EGI−CE	LEAR−(CE+EGI)	Local LEAR−Global LEAR
1010	+17.19	+10.65	+6.54	+0.22
1013	+15.66	+8.15	+7.51	-0.39
1022	+1.99	+5.43	-3.43	-9.97
1026	+12.53	+12.01	+0.52	+2.89
1037	-0.13	-1.51	+1.38	-5.29
1038	+9.49	+3.38	+6.11	+9.50
1040	+15.24	+12.38	+2.86	+3.17
1045	+9.54	+9.28	+0.26	-4.67
1047	+10.67	+10.69	-0.01	+5.67
1048	+10.48	+0.46	+10.03	-2.50
1059	+10.62	+7.85	+2.77	+2.92
1065	+12.22	+9.83	+2.39	+1.31
1066	+10.60	+4.44	+6.17	+5.86
1068	+12.17	+10.68	+1.48	+1.79
1071	+0.53	+4.60	-4.07	-1.14
1073	+9.51	+4.62	+4.89	-0.66
1084	+8.19	+6.27	+1.92	-4.50
1086	+13.05	+7.32	+5.73	+1.42
1088	+6.39	+0.96	+5.43	-1.59

At the actor level, Local LEAR exceeds Local CE for 18 of 19 held-out actors. CE+EGI also improves on 18 actors, and the additional LEAR-over-EGI training increment is positive for 16 actors. Averaged over actors, the main local effects are +9.79 points for Local LEAR over Local CE, +6.71 for Local CE+EGI over Local CE, and +3.08 for Local LEAR over Local CE+EGI. The mean Local LEAR versus Global LEAR difference is +0.21 points. Figure A2 shows the three local contrasts for each actor.

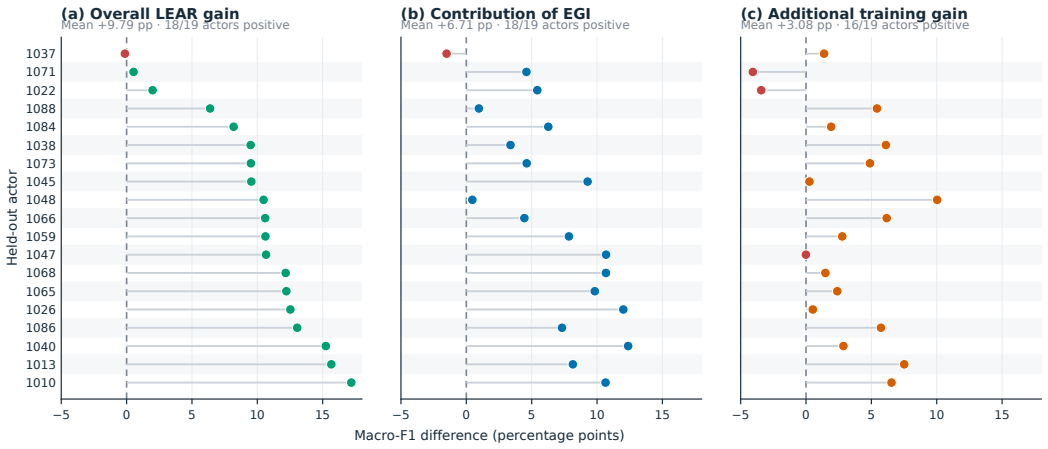

Figure A2: **CREMA-D effects for individual held out actors.** Differences in subset average Macro-F1 are computed within each actor after averaging three seeds. Actors share the same order across panels, sorted by the gain of LEAR over CE. (a) Total LEAR gain. (b) Gain from EGI at inference. (c) Additional training gain with the same EGI rule. Red markers indicate negative effects.

H Recent Strong Baseline Adaptations

H.1 MARS on CREMA-D

MARS (Baek et al., 2026) is evaluated on the same 63/9/19 actor split and the same A/V/AV state family. Its residual-guided routing, expert specialization, distillation, noise regularization, and adaptive modality sampling are retained. We adapt data paths, actor-split handling, media loading, evaluation logging, and checkpoint selection to the CREMA-D protocol used here. Validation mean-subset Macro-F1 is used for checkpoint selection. Each run uses batch size 64, learning rate 10^{-3}, and at most 100 epochs.

Table A9: **MARS on the CREMA-D protocol.** Results for this recent strong baseline are reported in percent.

Seed	Mean	A	V	AV
0	42.27	52.10	28.65	46.04
1	48.14	54.60	31.16	58.66
2	46.66	53.57	32.03	54.37
Mean	45.69	53.42	30.62	53.02

H.2 CoMiss on WISDM

The CoMiss (Mena et al., 2026) adaptation uses the public Co4Miss implementation. Its collaborative learning loss and missing view enumeration, fusion mechanism, and decision distillation are retained. The adapter supplies four 91-dimensional WISDM views and 18 classes, preserves the five subject-disjoint folds and train-only normalization, and evaluates all 15 nonempty availability states. The original implementation's validation criterion for checkpoint selection and the same training budget are used across all folds and seeds.

Table A10: **CoMiss on the WISDM protocol.** Fold and seed results for this recent strong baseline, in percent.

Fold	Seed	Mean-subset F1	Worst-state F1
0	0	64.41	25.15
0	1	65.00	24.71
0	2	64.82	24.73
1	0	57.42	24.41
1	1	57.57	23.99
1	2	57.23	23.76
2	0	54.66	23.98
2	1	54.71	24.61
2	2	54.86	23.63
3	0	56.99	27.26
3	1	56.82	27.22
3	2	57.34	27.11
4	0	57.37	21.91
4	1	57.05	23.10
4	2	57.94	22.60

Across all 15 runs, CoMiss averages 58.28 mean-subset F1 with a standard deviation of 3.51 and a mean worst-state F1 of 24.54. After seed averaging, the differences between CoMiss and LEAR across folds are 1.32, 0.88, 0.77, 0.64, and 1.46 points, with a mean difference of +1.01 points.

I Evidence Specificity and Routing Sensitivity

The evidence specificity study uses WISDM validation data and compares evidence matched to each subset reference with several alternative score sources.

Figure A3 joins the specificity controls and the validation routing grid.

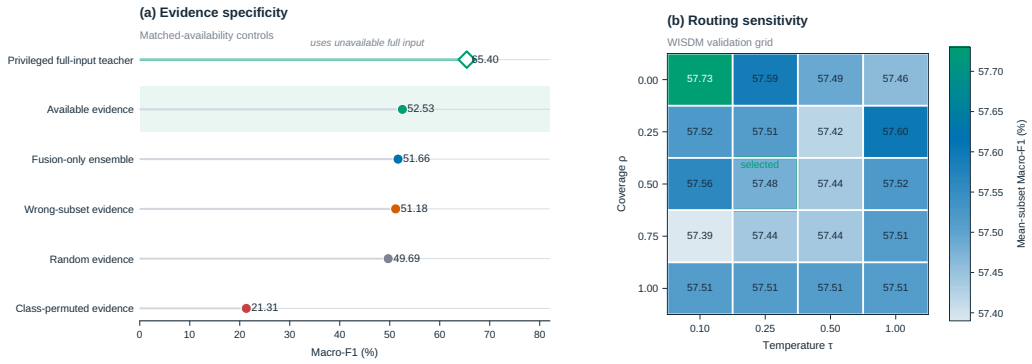

Figure A3: **Validation mechanism diagnostics on WISDM.** (a) Available evidence matched to each subset compared with alternative score sources for the CE path. The open diamond is a teacher with complete input. (b) Subset average Macro-F1 across coverage ρ and temperature τ on fold 0, seed 0; the outline marks the frozen $(0.5, 0.25)$ setting.

Table A11: **Evidence specificity controls on WISDM.** Validation results for alternative score sources in the CE path; ↑ indicates higher is better.

Control	Macro-F1 ↑	Accuracy ↑	ECE ↓	NLL ↓
Available evidence	52.53	57.82	0.270	2.61
Fusion-only ensemble	51.66	56.97	0.347	4.20
Wrong-subset evidence	51.18	56.44	0.362	4.71
Random evidence	49.69	55.11	0.372	4.60
Class-permuted evidence	21.31	24.86	0.638	10.05
Privileged full-input teacher	65.40	69.85	0.270	3.72

The matched reference outperforms alternatives based on fusion alone, the wrong subset, random scores, and permuted classes. The teacher with complete input reaches a higher score by using all modalities. Evidence head F1 has a positive association of $\rho = 0.1757$ with eventual LEAR gain, while the gap between fusion and evidence loss gap has near-zero rank correlation ($\rho = -0.0442$). This supports using evidence quality for distillation confidence and regret for within-training objective weighting.

A validation grid evaluates coverage $\rho \in \{0, .25, .5, .75, 1\}$ and temperature $\tau \in \{.1, .25, .5, 1\}$ on WISDM fold 0, seed 0. Mean-subset F1 ranges from 57.39 to 57.73, and the chosen $(\rho, \tau) = (.5, .25)$ setting gives 57.48.

Table A12: **Routing sensitivity on WISDM.** Validation subset average F1 (%) across coverage ρ and temperature τ.

$\rho \backslash \tau$	.10	.25	.50	1.00
0.00	57.73	57.59	57.49	57.46
0.25	57.52	57.51	57.42	57.60
0.50	57.56	57.48	57.44	57.52
0.75	57.39	57.44	57.44	57.51
1.00	57.51	57.51	57.51	57.51

J Additional Datasets

J.1 PAMAP2

PAMAP2 (Reiss, 2012) uses wrist, chest, and ankle streams as three modalities, giving seven nonempty availability states. Four held-out subjects and three seeds per subject are evaluated. LEAR improves the matched Local CE baseline for all four subjects on every reported endpoint.

Table A13: **PAMAP2 cross-dataset effects.** Difference between LEAR and Local CE, in percentage points.

Endpoint	Mean effect	Positive subjects
Mean-subset F1	+1.70	4/4
Worst-subset F1	+1.24	4/4
Full-input F1	+1.78	4/4
All-subsets correct	+3.48	4/4

K Training and Inference Efficiency

Training and inference measurements use an NVIDIA RTX A6000. Local LEAR uses less measured peak allocation than Global LEAR, while the evidence-aware objective increases update time relative

to Local CE. Detachment reduces stored backward state, whereas evidence processing adds forward and local-loss computation.

Table A14: **WISDM training efficiency.** Median update time and peak allocation for batch size 192 over all 15 states; ↓ indicates lower is better.

Method	Median ms/update ↓	Peak MiB ↓
Local CE	25.07	109.2
LEAR	37.34	109.2
Global CE	27.54	124.2
Global LEAR	43.40	125.0

Table A15: **WISDM inference efficiency.** Median batch latency and peak allocation over all 15 states; ↓ indicates lower is better.

Variant	Median ms/batch ↓	Peak MiB ↓
Local CE	5.743	24.88
Local CE+EGI	5.274	25.07
LEAR	5.327	25.07
Global CE	4.067	24.88
Global CE+EGI	5.308	25.07

Evidence-aware inference adds approximately 0.19 MiB measured peak allocation in this setup. Timing differences at inference are small relative to the training overhead and depend on the device.

L Implementation Details

WISDM uses three blocks of width 128, AdamW with learning rate 10^{-3} and weight decay 10^{-4}, batch size 192, at most 30 epochs, and patience 6. Checkpoint selection maximizes validation mean-subset Macro-F1. Experiments use three prespecified random seeds, and class-balanced sampling uses a deterministic CPU sampler. The fixed routing parameters are $\rho = 0.5$, $\tau = 0.25$, and the EGI weight is $\eta = 2$. CREMA-D uses the separate image and audio training budget in Appendix G.

The WISDM environment uses Python 3.11.7 and PyTorch 2.5.1+cu121 with CUDA 12.1 and cuDNN 9.1.0. NumPy is 1.26.4, SciPy 1.16.3, scikit-learn 1.8.0, and Matplotlib 3.10.8. Experiments use an NVIDIA RTX A6000. Deterministic CuBLAS workspace is set to `:4096:8`.

M Empirical Scope and Research Directions

The experiments reveal a consistent mechanism for available-evidence learning. Informative, modality heads aligned with the labels create useful local reference signals: WISDM provides related inertial streams, CREMA-D combines asymmetric modalities and yields a large training gain, and PAMAP2 extends the positive pattern to a second multi-sensor activity dataset.

MARS and CoMiss provide recent strong task baselines for the two evaluation settings. CoMiss reaches the highest absolute WISDM mean, while LEAR exceeds the adapted MARS baseline on CREMA-D. Together, these comparisons place evidence-aware local learning against current specialized global architectures for incomplete multimodal learning.

N Statistical Evaluation

WISDM. We first average the three seeds within each held out subject fold and then compute paired effects across the five folds. Figure 3(a) reports Student-t intervals over these effects by fold.

CREMA-D. Table 2 averages pooled-test Macro-F1 across states and seeds. The paired analysis first averages state metrics and seeds within each actor, then gives all 19 test actors equal weight. Figure 3(b) and Table A16 report percentile intervals from 10,000 paired actor-bootstrap replicates. The pooled and actor-averaged summaries use different weighting and therefore differ slightly.

Table A16: **CREMA-D statistical contrasts.** Actor averaged effects and paired bootstrap intervals, in percentage points.

Contrast	Mean	95% interval
Local LEAR – Local CE	+9.79	[7.64, 11.73]
Local CE+EGI – Local CE	+6.71	[4.93, 8.40]
Local LEAR – Local CE+EGI	+3.08	[1.47, 4.66]
Global CE+EGI – Global CE	-0.48	[-1.90, 0.97]
Global LEAR – Global CE+EGI	+0.90	[-0.88, 2.70]
Local LEAR – Global LEAR	+0.21	[-1.77, 2.19]

O Complete Training and Inference Diagram

Figure A4 expands the compact overview in Figure 2 into the complete training and inference graph. It makes explicit the sample-wise availability states, fusion and evidence losses, evidence-regret computation, coverage-bounded routing, confidence-weighted distillation, and direct modality-head supervision. The observed-modality reference is detached before it enters the evidence loss, distillation target, and inference score, while normalization and stop-gradient define the inter-block representation boundary. The local objective sums the fusion, distillation, and modality-head terms directly.

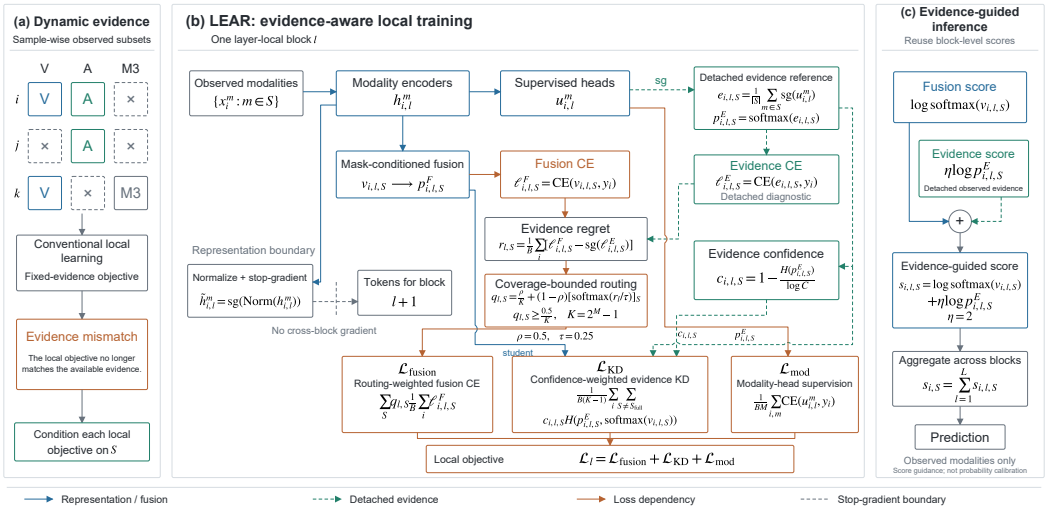

Figure A4: **Complete LEAR training and inference mechanism.** This figure expands the condensed presentation in Figure 2. (a) Each sample exposes an observed modality subset. (b) Within a layer-local block, supervised modality heads form the detached reference $p^E_{i,l,S}$; fusion regret determines coverage-bounded routing, evidence confidence weights distillation, and direct label supervision trains the modality heads. The block objective is $\mathcal{L}_l = \mathcal{L}_{\text{fusion}} + \mathcal{L}_{\text{KD}} + \mathcal{L}_{\text{mod}}$. (c) Evidence-guided inference combines fusion and detached evidence scores and aggregates them across blocks. The index i denotes a sample, l a block, m a modality, and S the observed subset.

P CREMA-D QUALITATIVE AND REPRESENTATION ANALYSES

We examine selected prediction corrections, score concentration on correctly classified clips, and V-only class structure on the CREMA-D split described in Appendix G.

P.1 SELECTED PREDICTION CORRECTIONS

Figure A5 shows six test clips from the fixed seed-0 runs where Local CE is incorrect and LEAR is correct. Two clips are selected within each A/V/AV state by descending LEAR-minus-Local-CE increase in $p(\mathrm{GT})$, with clip identifiers breaking ties and deterministic relaxation of actor and class diversity constraints. The examples include V-only HAP and ANG clips that both baselines misclassify. This selection illustrates individual corrections; aggregate performance is reported in Appendix G.

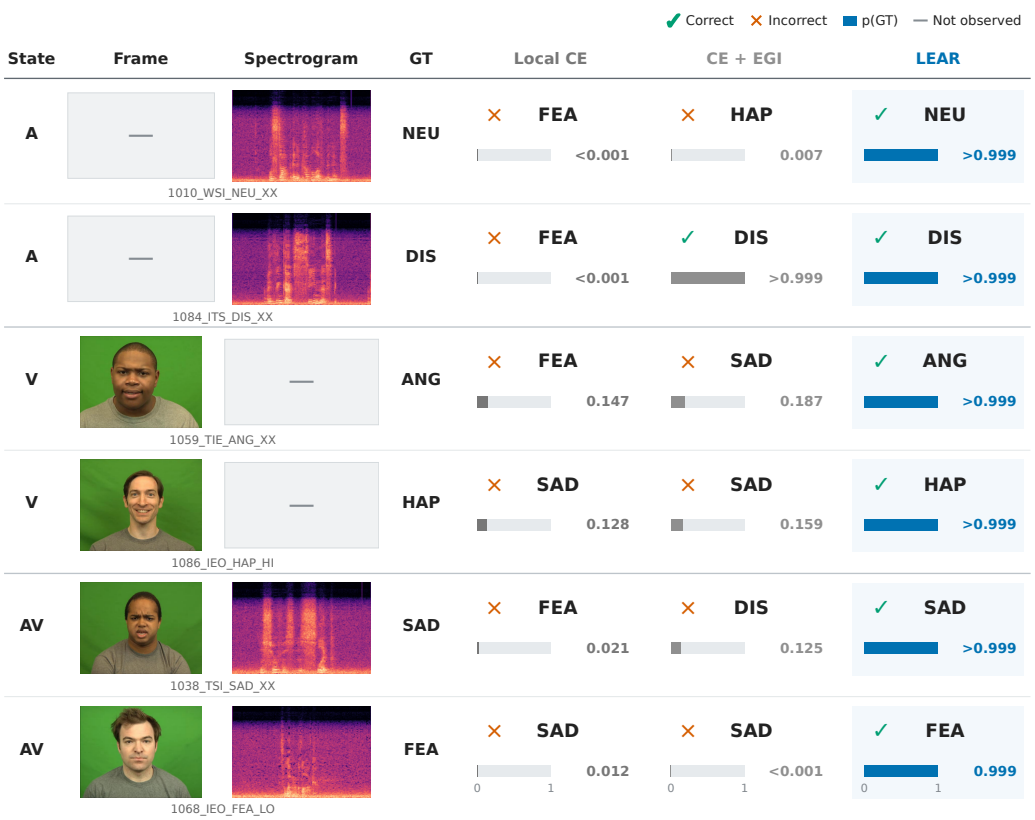

Figure A5: **Selected prediction corrections on CREMA-D.** Six seed-0 test clips for which Local CE is incorrect and LEAR is correct, with two examples each for audio (A), visual (V), and audio-visual (AV) inputs. The media columns show the evaluation frame and a spectrogram of the corresponding audio; a dash denotes an unavailable modality. GT is the annotated emotion. Method columns show the predicted class and $p(\mathrm{GT}) = \mathrm{softmax}(s)_y$, where s is the native final inference score vector. Bars share a scale from 0 to 1. CE + EGI correctly predicts the audio-only DIS example and misclassifies the other five cases.

P.2 Score Concentration on Correct Predictions

Figure A6 applies the same seed-0 selection rule to clips that all three methods classify correctly. For the two V-only SAD clips, Local CE assigns $p(\mathrm{GT})$ values of 0.211 and 0.219, CE + EGI assigns 0.237 in both cases, and LEAR assigns values above 0.999. The predicted labels agree while the normalized score distributions become more concentrated. These native-score distributions describe confidence sharpening and do not establish improved calibration.

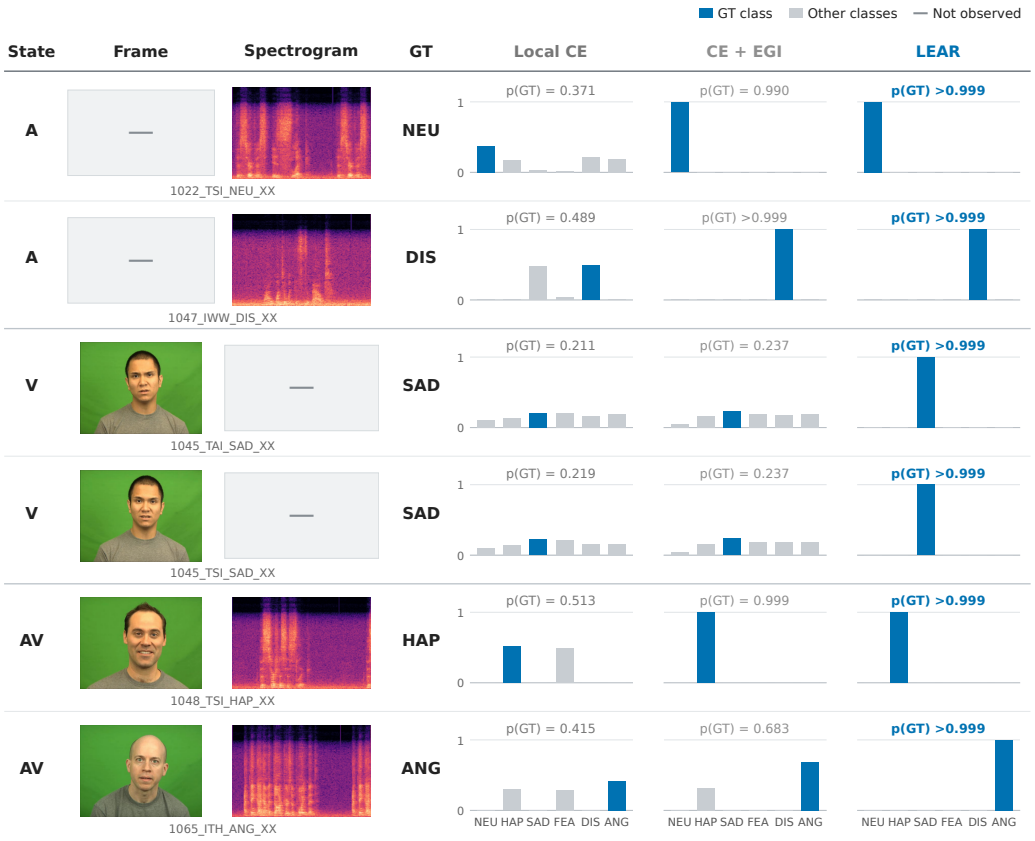

Figure A6: **Score concentration on correctly classified CREMA-D examples.** All three methods correctly classify these six seed-0 test clips, with two clips for each A/V/AV state. Each six-class distribution is the softmax of the method's native final inference scores. Blue bars identify GT and gray bars show the other classes. Panels share the class order NEU, HAP, SAD, FEA, DIS, ANG and a probability scale from 0 to 1. The value above each panel gives $p(\mathrm{GT})$. The V-only SAD examples show diffuse baseline distributions and concentrated LEAR distributions with the same top-1 prediction.

Score display and checkpoint configuration. The display softmax uses temperature 1. Local CE sums fusion log-probabilities across three blocks; CE + EGI and LEAR also include the weight-2 evidence log-probability term from each block. CE + EGI uses the CE training loss in a separate seeded run, with EGI active during validation and test and its own early-stopping checkpoint. Numeric labels use > 0.999 for values strictly above 0.999 and < 0.001 for positive values strictly below 0.001; other values are rounded to three decimal places. Bars retain the full exported values. In particular, LEAR's value of 0.998584 for the AV/FEA correction example is displayed as 0.999.

P.3 V-ONLY CLASS STRUCTURE AND RECALL

Figure A7 combines V-only embeddings with test statistics across seeds 0, 1, and 2. LEAR increases mean recall for HAP, DIS, and NEU by 45.74, 24.69, and 17.98 percentage points, respectively, while FEA recall decreases by 9.77 points. Mean cross-actor purity@10 increases from 25.19% to 35.77%. The paired purity change is positive in all three seeds, with a gain of 0.31 percentage points in seed 2.

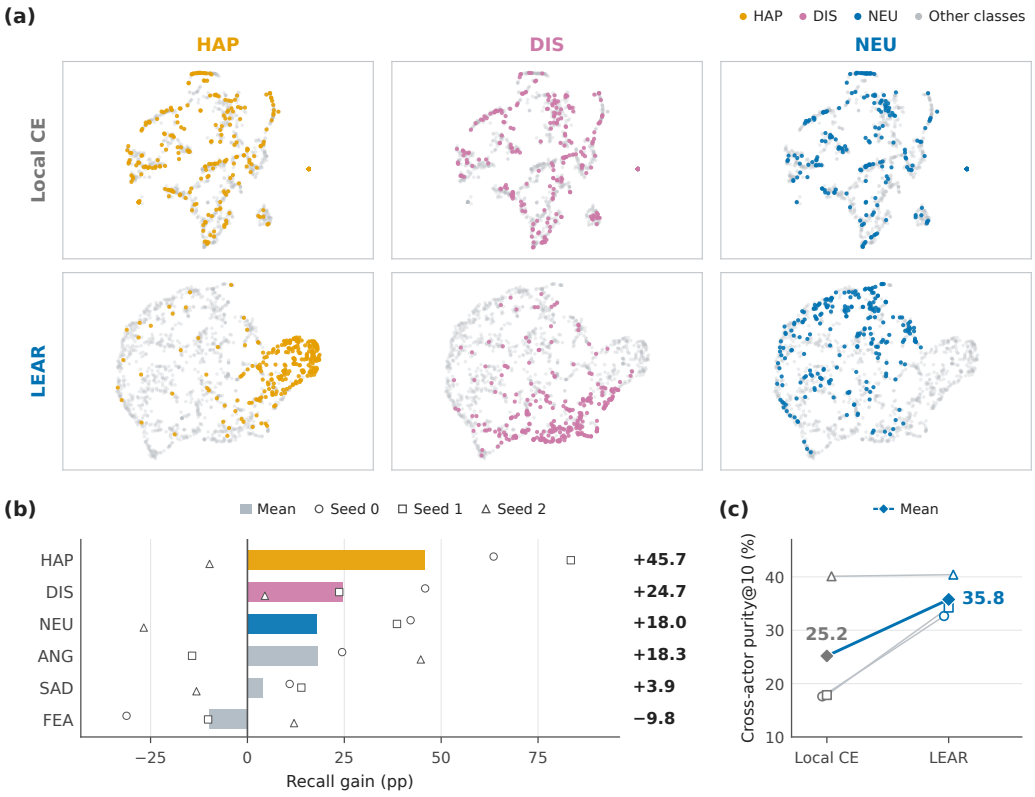

Figure A7: **V-only representations and class-wise performance on CREMA-D.** (a) One-vs-rest views of HAP, DIS, and NEU for Local CE and LEAR, using the same 1,558 test clips and seed-1 checkpoints. Colored points denote the indicated class and gray points denote the rest. The highlighted classes were selected by validation recall improvement. (b) Recall changes for all six classes, measured as LEAR minus Local CE in percentage points. Bars and right-hand values are means over seeds 0, 1, and 2; circles, squares, and triangles show the corresponding seed-level changes. (c) Cross-actor class purity among the 10 nearest neighbors in the standardized 128-dimensional feature space before t-SNE, excluding neighbors from the query clip's actor. Thin lines connect paired seeds, and the bold line connects mean diamonds.

Embedding and neighborhood protocol. The embedding panels use seed-1 test coordinates. Features are standardized separately for each method, seed, and split, with the scaler fitted on that split. The t-SNE configuration uses Euclidean distance, perplexity 40, early exaggeration 8, learning rate 200, PCA initialization, random state 11, and 300 iterations. It was selected from 1,460 candidate mappings on validation data using min-max-normalized trustworthiness, kNN preservation@10, and two-dimensional class purity with weights 0.50, 0.30, and 0.20, respectively, averaged equally across both methods and all three seeds. Ties are resolved by trustworthiness and then neighborhood preservation. The highlighted classes are selected separately by validation recall improvement. Cross-actor purity is computed directly from the standardized 128-dimensional features, excluding all same-actor neighbor candidates.