LLUNA: COLLABORATIVE UNIMODAL MODELS FOR MISSING-MODALITY INFERENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Pretrained unimodal models offer a flexible route to multimodal prediction through collaboration, but this collaboration can become less accurate than a surviving unimodal model when modalities are missing. We study this setting using fixed surviving-model anchors and find that collaborative gains can weaken or reverse as inputs disappear, even when the anchor's own input is unchanged. Our diagnostics further reveal reduced recovery of modality-specific information after aggregation and state-dependent collaborative representations. We propose LLUNA, which combines a direct private prediction path from adapted unimodal features with a common graph-based collaboration branch that provides a residual correction. Progressive Collaboration Consistency organizes richer-to-poorer guidance along single-modality removals, aligning normalized common representations while task supervision adapts predictions to the available evidence. Across eight datasets spanning affective computing, scene understanding, and medical prediction, LLUNA achieves the best evaluated average performance over non-full input states, improving Mean-Missing by 4.26% on average relative to the strongest evaluated baseline. The fixed-anchor diagnostic also shows positive gains in mean collaboration under minimal inputs on all eight tasks. Code is available at https://anonymous.4open.science/r/Lluna.

1 INTRODUCTION

Multimodal systems are essential for perceiving and reasoning about the real world (Baltrušaitis et al., 2018; Alayrac et al., 2022; Girdhar et al., 2023). In practice, however, modalities may be unavailable: audio streams may be missing, sensors may fail, and clinical measurements may be uncollected (Wu et al., 2024; Wang et al., 2023; Yun et al., 2024; Li et al., 2025). Prediction with missing modalities is therefore an important problem in multimodal learning.

Most missing-modality methods learn a single multimodal predictor, using strategies such as modality dropout, aggregation of available representations, or shared-specific modeling (Neverova et al., 2015; Havaei et al., 2016; Wang et al., 2023; Yun et al., 2024; Li et al., 2025; Xie et al., 2026). In contrast, with the abundance of publicly-available pretrained foundation models (Osborne et al., 2024), we study a different regime where independently pretrained unimodal models are composed through *model collaboration*. Specifically, model collaboration can align heterogeneous representations or combine model outputs, as in Freeze-Align (Maniparambil et al., 2025) and FoE (Wang et al., 2024a). This raises a basic question: does collaboration remain effective when some modalities become unavailable? As shown in Figure 1(a), across affective computing, scene understanding, and medical prediction, our experiments show that collaboration gains got smaller as modalities become unavailable and can even turn negative. A surviving unimodal model can then be more accurate than the collaborative predictor, even though its own input remains unchanged.

We next examine how collaboration represents the available evidence. For the illustrative runner in Figure 1(b), "I finally did it" conveys achievement, a relieved voice adds emotional context, and the video reveals physical effort. These complementary cues illustrate the *private signal*, i.e., the modality-specific information collaboration should retain. Yet feature-level probes find that this signal is less recoverable after collaboration across various mechanisms and training policies (Figure 1(b), Appendix E.2). The second challenge is to coordinate collaboration across available missing-modality states. As video and audio disappear subsequently, the same event is represented through different

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

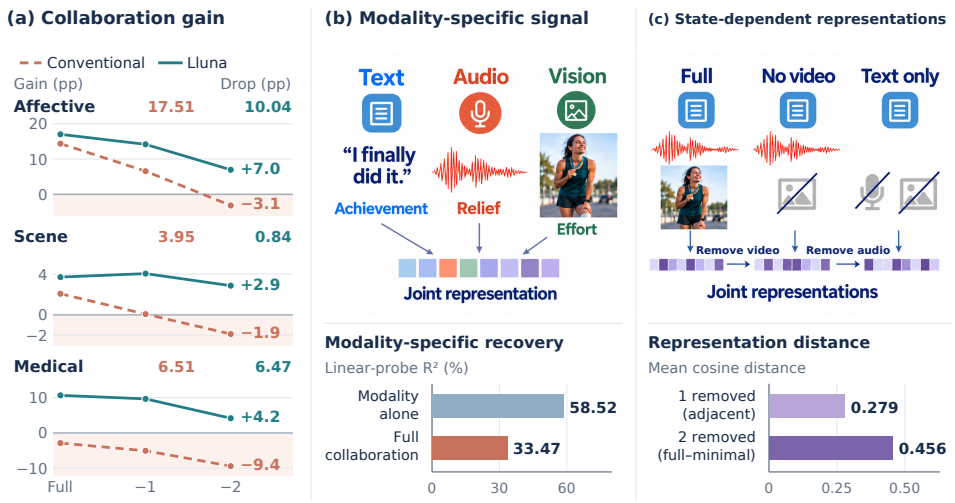

Figure 1: Collaboration under missing modalities. **(a)** Domain-mean accuracy gain in percentage points (pp) over fixed surviving unimodal models; annotations give the absolute decline from Full to two removals in pp. **(b)** Attn. pool's private-signal recovery, averaged over eight tasks, using the linear-probe coefficient of determination (R^2). **(c)** Cosine distances for adjacent and Full-to-Minimal state pairs, averaged over six three-modality tasks. Details: Appendix E.

cues and interactions (Figure 1(c)). This calls for direct access to modality-specific features together with cross-state guidance for the common collaborative branch.

In this paper, we propose **LLUNA**, a collaboration strategy for unimodal models for missing-modality inference. As shown in Figure 2, LLUNA pairs *private prediction* with *common collaboration*. Model-specific adapters provide a direct path from available features to prediction, while a common branch exchanges messages on a compatibility graph weighted by prediction agreement. Its output supplies a learned residual correction to the private prediction. Single-modality removals connect the full input to every nonempty availability state, decomposing larger evidence changes into local steps. Progressive Collaboration Consistency (PCC) uses these steps to train a shared model: richer states guide the common representations of their poorer neighbors, and intermediate states in turn provide targets for subsequent steps. We further establish a local-to-global stability bound, relating the drift of normalized common representations between distant availability states to accumulated discrepancies along single-modality transitions.

Experiments across eight datasets in three application domains show that LLUNA improves Mean-Missing performance over the strongest evaluated baseline (among collaboration methods, missing-modality approaches, and Omni models) on every task, averaging 4.26% relative improvement. The collaboration diagnostic also shows positive mean gains under minimal inputs on all eight tasks. Figure 1a further shows a 37.9% reduction in the domain-averaged absolute decline in collaboration gain after two modality removals, compared with conventional collaboration.

Our contributions are:

- We identify loss of collaborative benefit under missing modalities and characterize the accompanying changes in private-signal recovery and collaborative representations.
- We propose LLUNA, which retains a direct prediction path, adds a common residual correction, and applies progressive richer-to-poorer consistency to the common representation.
- Extensive experiments in three domains show improved Mean-Missing performance on every task, complemented by positive mean collaboration gains under minimal inputs in the diagnostic.

2 Related Work

Model collaboration. Model collaboration combines complementary predictions and representations through ensemble weighting and learned feature interactions (Baltrušaitis et al., 2018; Wolpert, 1992; Jacobs et al., 1991; Vaswani et al., 2017; Lee et al., 2019). Pretrained-model ensembling and routing

often combine models within the same modality (Wang et al., 2024a; Lu et al., 2024; Yadav et al., 2024). At the feature level, shared embeddings align signals across modalities (Girdhar et al., 2023), frozen-encoder alignment connects pretrained unimodal models (Maniparambil et al., 2025), and graph-based aggregation supports variable modality sets (Patapati & Srinivasan, 2026).

Multi-agent Systems. LLM-based MAS coordinate instruction-following agents through dialogue, roles, and task planning (Li et al., 2023; Wu et al., 2023; Hong et al., 2024; Talebirad & Nadiri, 2023). LLUNA uses relatively compact unimodal models, including encoders without a conversational interface, and requires neither instruction following nor agentic behavior. Moreover, modality loss changes the available evidence and can make some models unavailable. LLUNA learns collaboration over representations, rather than assuming all participants communicate via language.

Learning with missing modalities. Missing-modality learning adapts representations, fusion, and supervision to incomplete inputs (Wu et al., 2024). Earlier work uses modality dropout and aggregation over available-modality representations (Neverova et al., 2015; Havaei et al., 2016). Other approaches include shared-specific feature learning, modality-aware routing, and cross-state distillation (Wang et al., 2023; Yun et al., 2024; Li et al., 2025; Baek et al., 2026; Xie et al., 2026), as well as quality-aware fusion and reasoning from incomplete observations (Meng & Wei, 2026; Xu et al., 2026). Most of these methods address missing inputs within a single task-specific multimodal predictor. In contrast, LLUNA studies missing-modality inference as a flexible model-collaboration problem among independently pretrained unimodal models, where modality loss changes both the available evidence and the set of models that can participate in collaboration.

Omni and unified models. Integrating modalities into one predictor offers another route, from pretrained vision–language coupling for few-shot learning (Alayrac et al., 2022) to Omni models combining vision, audio, and language for understanding and interaction (Xu et al., 2025; Tong et al., 2025; Ye et al., 2026; Abd et al., 2026; Zhang et al., 2025). Unified multimodal models bring understanding and generation into a common modeling framework (Luo et al., 2026; Wang et al., 2024b; Xie et al., 2025; Deng et al., 2025).

3 LLUNA

We are given labeled training examples $\{(x, y)\}$, where y is the target label and x has M different modalities: $\mathcal{M} = \{M_i\}_{i=1}^{M}$, with M_i denoting the ith modality. Missing-modality learning aims to learn a function f to infer the labels for a testing set $\{\tilde{x}\}$ that has at most $M - 1$ modalities.

3.1 OVERVIEW OF LLUNA

Different from existing missing-modality approaches that learn a single model (Meng & Wei, 2026; Li et al., 2025; Wang et al., 2023), LLUNA instead learns a collaboration strategy from a set of N *heterogeneous, unimodal* models, $\mathcal{F} = \{F_i\}_{i=1}^{N}$ for robust prediction. Different F_is may specialize in different modalities, e.g., F_1 can be Qwen3-1.7B (Yang et al., 2025) as a textual model, and F_2 can be Whisper (Radford et al., 2023) as an audio model. All F_is remain *frozen* in training phase.

To sustain beneficial collaboration under missing modalities, LLUNA builds on a private prediction path that retains direct access to the available modality-specific features (Figure 2). Using the compatibility graph, the common collaboration path learns interactions among available models to refine this base prediction. Progressive Collaboration Consistency (PCC) further uses richer availability states to guide the common representations learned under progressively reduced inputs.

3.2 COMPATIBILITY GRAPH BUILDING

An availability state $S \subseteq \mathcal{M}$ specifies the nonempty subset of observed modalities. Since models in $\mathcal{F}$ may differ in modality and representation space, LLUNA first adapts each frozen feature $\mathbf{h}_i$ with a model-specific adapter ϕ_i, giving $\mathbf{p}_i = \phi_i(\mathbf{h}_i)$, and then maps it into a common interaction space using a shared aligner C, with $\mathbf{c}_i^{(0)} = C(\mathbf{p}_i)$. Below, i, j index models; unless stated otherwise, message passing, pooling, and their normalization involve only available models. We omit state indices when describing a single state.

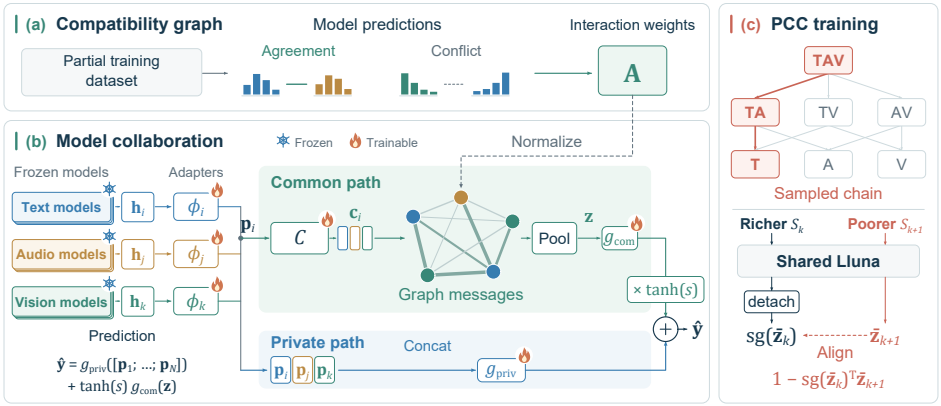

Figure 2: **Overview of LLUNA. (a)** Compatibility graph building using training-partition prediction agreement. **(b)** Residual collaboration of private and common features. **(c)** PCC training. Consecutive availability states S_k, S_{k+1} differ by one removed modality; here k indexes chain positions. PCC aligns their unit-normalized consensuses $\bar{\mathbf{z}}_{k+1}$ and $\bar{\mathbf{z}}_k$, with stop-gradient sg on the richer state. T/A/V denote text/audio/vision. Inference uses only the observed state.

Heterogeneous models can provide complementary information, but effective collaboration also requires their evidence to be sufficiently compatible rather than strongly conflicting. We therefore use agreement in task predictions as a soft measure of compatibility to organize cross-model interactions (Figure 2(a)). For each dataset, we fit a task head to each model's frozen training features and define $A_{ij} = \mathrm{sim}(\hat{\mathbf{y}}_i^{(n)}, \hat{\mathbf{y}}_j^{(n)})$, where $\hat{\mathbf{y}}_i^{(n)}$ is F_i's task-head prediction for n partial training examples. The symmetric similarity sim uses class probabilities for classification and standardized scalar predictions for regression. The graph is estimated once from the training partition and kept fixed throughout collaboration training and inference. For each availability state, we restrict $\mathbf{A}$ to available models and normalize incoming weights for each target model j as $\widehat{A}_{ij} = A_{ij} / \sum_{i'} A_{i'j}$. Positive self-loops ensure a valid incoming connection even for an isolated available model (More details in Appendix F).

3.3 Residual Coordination of Common and Private Information

Our analysis shows that modality-specific signals can become less accessible after collaboration (Appendix E.2), suggesting that common collaboration should refine, rather than replace, the direct evidence provided by individual models. Accordingly, LLUNA decomposes prediction into a private base prediction from adapted model features and a common residual correction from cross-model interaction (Figure 2(b)). The private head g_{priv} is an MLP that reads the adapted features $\mathbf{p}_i$ in a fixed model order, with zeros in unavailable slots, to form the base prediction. The common branch instead operates on the aligned features $\mathbf{c}_i^{(0)}$ through the compatibility graph. Using standard graph message passing (Gilmer et al., 2017), each aligned feature is updated as $\mathbf{c}_j^{(\ell+1)} = \mathrm{LayerNorm}\Big(\mathbf{c}_j^{(\ell)} + \gamma(\mathbf{c}_j^{(\ell)}) \sum_i \widehat{A}_{ij} U(\mathbf{c}_i^{(\ell)})\Big)$, where U is a learned message map and γ is a sigmoid gate. After L layers, attention pooling produces the common consensus $\mathbf{z} = \mathrm{AttentionPool}(\{\mathbf{c}_i^{(L)}\})$ (details in Appendix F).

Crucially, common collaboration refines rather than replaces prediction from model-specific evidence. A common head g_{com}, implemented as an MLP, maps $\mathbf{z}$ to a residual correction:

$$\hat{\mathbf{y}} = g_{\mathrm{priv}}([\mathbf{p}_1; \ldots; \mathbf{p}_N]) + \tanh(s)\, g_{\mathrm{com}}(\mathbf{z}), \tag{1}$$

where s is a learned residual-scaling scalar (Bachlechner et al., 2021). Both g_{priv} and g_{com} additionally receive binary availability indicators, omitted from the notation for clarity.

3.4 Training with Progressive Collaboration Consistency

Missing modalities change both the available evidence and the resulting collaborative representation. Single-modality removals form a connected graph over nonempty availability states, providing a local structure for richer-to-poorer guidance. Along sampled transitions, PCC aligns the poorer

state's common representation with its richer neighbor, while task supervision allows the overall prediction to adapt to the remaining evidence. Intermediate states can in turn guide subsequent removals, yielding *progressive richer-to-poorer* transfer across availability states.

As shown in Figure 2(c), we organize nonempty modality subsets into removal chains $S_0 \supset S_1 \supset \cdots$, where each transition $S_k \to S_{k+1}$ removes exactly one modality; for example, TAV $\to$ TA $\to$ T. Let $\mathbf{z}_k$ denote the common consensus at state S_k and $\bar{\mathbf{z}}_k = \mathbf{z}_k / \|\mathbf{z}_k\|_2$ its unit-normalized form. For each adjacent transition, PCC aligns the poorer representation to its richer neighbor while stopping the gradient through the richer state:

$$\mathcal{L}_{\mathrm{PCC}} = \sum_{(S_k, S_{k+1}) \in \mathcal{E}} \omega_{k,k+1} \left[1 - \mathrm{sg}(\bar{\mathbf{z}}_k)^\top \bar{\mathbf{z}}_{k+1}\right], \tag{2}$$

where sg denotes stop-gradient, $\mathcal{E}$ contains the sampled adjacent transitions, and $\omega_{k,k+1}$ balances contributions across different availability states. Because an intermediate state can receive guidance from a richer neighbor and subsequently guide a poorer one, these local consistency constraints propagate progressively along the removal chain. Normalizing $\mathbf{z}_k$ focuses the guidance on the direction of the common consensus while allowing its magnitude to vary with the available evidence (details in Appendix F.1). Appendix G further relates accumulated local discrepancies to representation changes between distant states. The overall objective combines state-wise task supervision, auxiliary supervision of the common prediction, and progressive consistency:

$$\mathcal{L} = \mathbb{E}[\mathcal{L}_{\mathrm{task}}(y, \widehat{\mathbf{y}}) + \lambda_{\mathrm{com}} \mathcal{L}_{\mathrm{task}}(y, g_{\mathrm{com}}(\mathbf{z}))] + \lambda_{\mathrm{PCC}} \mathcal{L}_{\mathrm{PCC}}. \tag{3}$$

Here, λ_{com} and λ_{PCC} control the auxiliary and consistency terms. Task supervision adapts the complete prediction to each availability state, while PCC regularizes only the common collaboration.

Efficient chain sampling. Training with randomly dropped modalities is an established strategy for incomplete-input robustness (Neverova et al., 2015; Havaei et al., 2016; Wang et al., 2023). However, exhaustively evaluating all $2^M - 1$ nonempty availability states at every training step becomes increasingly expensive as the number of modalities grows. To improve training efficiency while retaining progressive guidance, LLUNA samples K maximal removal chains on the same training examples, reusing states and adjacent transitions shared across chains. Validation and test evaluate all legal availability states, while inference evaluates only the observed state.

3.5 Local-to-Global Representation Stability

Although PCC regularizes only adjacent availability states, local control extends along the availability graph. Any two nonempty states can be connected through a sequence of single-modality changes.

Theorem 1 (Local-to-global representation stability). *Fix an input and a path $\{S_i\}_{i=0}^{D}$ of $D \geq 1$ single-modality transitions. Let $d_k = 1 - \bar{\mathbf{z}}_{k-1}^\top \bar{\mathbf{z}}_k$ denote the cosine discrepancy between consecutive normalized common representations. If the common consensus is nonzero at every state, then*

$$\|\bar{\mathbf{z}}_0 - \bar{\mathbf{z}}_D\|_2 \leq \sum_{k=1}^{D} \sqrt{2d_k} \leq \sqrt{2D \sum_{k=1}^{D} d_k}. \tag{4}$$

Theorem 1 shows that controlling local PCC discrepancies also limits representation drift between more distant availability states. Thus, progressive consistency provides pathwise global control without directly forcing distant evidence configurations to match. The proof is given in Appendix G.

4 Experiments

4.1 Setup

Benchmarks and protocol. Our experiments cover eight datasets in three domains: *affective computing*, *scene understanding*, and *medical prediction*. Affective computing includes CH-SIMS v2 (**SIMS**) (Liu et al., 2022), CMU-MOSI (**MOSI**) (Zadeh et al., 2016), MELD (Poria et al., 2019), and UR-FUNNY v2 (**URF**) (Hasan et al., 2019). Scene understanding includes derived protocols from Daily-Omni (**Daily**) (Zhou et al., 2025) and WorldSense (**WS**) (Hong et al., 2026). Medical

Table 1: **Mean-Missing performance across eight datasets.** Metrics: mean absolute error (MAE), support-weighted F1 (W-F1), accuracy (Acc.), and balanced accuracy (BAcc.). MAE means and Standard Deviations (SDs) are multiplied by 100; F1 and accuracy means are percentages, with SDs in percentage points. Protocol: Appendix A.1.

Method	SIMS MAE ↓	MOSI MAE ↓	MELD W-F1 ↑	URF Acc. ↑	Daily Acc. ↑	WS Acc. ↑	CirCor BAcc. ↑	ADNI BAcc. ↑	Avg. rank ↓
GENERAL COLLABORATION									
Mean ens.	53.24±0.02	145.71±0.25	31.31±0.00	61.26±0.54	33.16±0.81	31.10±0.12	50.00±0.00	66.88±6.12	12.06
Logit stack (1992)	49.62±0.93	144.12±2.54	22.70±0.57	60.92±1.07	30.11±3.73	31.20±0.15	51.95±2.20	71.92±3.20	13.25
Concat MLP	42.27±0.73	130.21±4.21	44.13±2.98	64.54±1.37	31.19±1.85	30.52±0.65	74.73±1.89	79.71±3.03	8.00
Attn. pool	38.87±0.75	115.21±6.28	49.93±0.88	66.60±0.98	30.00±1.10	30.40±0.62	72.61±4.76	77.75±3.90	7.62
Soft MoE (1991)	58.21±4.88	133.86±13.86	47.11±1.33	66.34±0.49	29.83±2.02	30.22±0.99	70.28±3.54	73.54±2.07	11.38
FoE (2024a)	53.88±2.27	148.02±7.96	31.31±0.00	52.39±2.28	31.81±1.92	29.75±1.81	49.92±0.19	64.96±5.84	15.44
Set Trans. (2019)	42.95±0.73	138.54±12.45	48.32±1.38	63.76±0.89	29.66±2.39	29.03±2.09	71.32±3.00	73.25±5.85	12.12
CLARGA (2026)	39.99±0.34	122.90±2.63	52.48±0.81	66.33±0.67	29.49±0.95	30.73±0.27	73.63±4.83	77.00±3.38	8.25
MISSING-MODALITY METHODS									
GIML (2026)	68.09±20.73	181.62±96.15	33.72±6.49	53.96±3.06	32.94±1.29	30.92±0.18	74.83±1.26	72.88±2.67	11.38
MARS (2026)	40.83±3.79	132.58±16.83	51.94±0.30	67.47±0.50	31.47±1.61	30.35±0.58	63.48±10.72	76.62±1.58	8.38
LIMSSR (2026)	35.90±0.21	102.34±0.56	53.98±0.33	68.86±0.82	28.59±2.34	29.91±0.84	65.98±13.55	73.58±3.09	7.75
SimMLM (2025)	39.64±0.43	116.10±3.07	53.09±0.25	68.05±0.43	30.17±1.47	30.93±0.43	76.60±2.30	79.00±2.93	5.00
ShaSpec (2023)	37.39±0.27	116.82±3.38	52.92±0.48	67.25±0.70	30.79±1.15	30.95±0.23	74.59±1.16	78.08±3.12	5.50
Flex-MoE (2024)	43.25±7.98	125.34±7.87	50.35±1.47	67.02±1.27	30.34±2.33	29.82±0.80	68.34±4.85	73.88±2.36	9.62
OMNI MODELS									
Q2.5-O-3B (2025)	171.86	161.69	22.53	55.00	35.88	28.49	61.50	52.92	15.62
Q2.5-O-7B (2025)	94.56	99.67	39.58	53.91	36.72	29.71	59.80	53.75	12.00
InterO-8B (2025)	182.38	135.56	26.52	51.36	24.86	22.94	58.83	53.54	18.38
OmniV-9B (2026)	106.25	104.34	34.39	56.17	31.92	25.63	75.17	50.00	12.38
Gemma4-12B (2026)	366.55	193.88	30.53	52.93	32.49	30.06	52.19	50.83	16.50
StreamO-8B (2025)	169.47	387.08	17.23	51.96	25.71	28.24	55.44	51.04	19.38
LLUNA (ours)	**34.24**±0.30	**96.71**±0.63	**56.10**±0.21	**70.62**±0.23	**37.29**±0.96	**33.90**±0.30	**81.76**±0.55	**82.21**±1.26	**1.00**
Δ_{rel} (%)	+4.61	+2.98	+3.91	+2.56	+1.54	+8.65	+6.73	+3.14	**+4.26%**

prediction includes the four-site CirCor DigiScope task (**CirCor**) (Oliveira et al., 2021) and MCI-to-dementia conversion prediction on ADNI (Mueller et al., 2005). The main comparison evaluates every nonempty modality subset on a fixed test split. We define **Mean-Missing (MM)** as the task metric averaged equally over all nonempty, incomplete modality subsets in a run. For trainable same-pool methods, we select one checkpoint per run by validation MM for all test states and report the mean and standard deviation (SD) across five runs. Each dataset uses one LLUNA configuration shared by five runs. Task definitions, data splits, and training details are in Appendix B; aggregation follows Appendix A.1.

Baselines. We compare against three groups. (1) *General collaboration methods* include mean ensembling (Mean ens.) and concatenation MLP (Baltrušaitis et al., 2018), learned logit stacking (Logit stack) (Wolpert, 1992), attention pooling (Attn. pool), soft MoE (Jacobs et al., 1991), FoE (Wang et al., 2024a), Set Transformer (Set Trans.) (Lee et al., 2019), and CLARGA (Patapati & Srinivasan, 2026). (2) *Missing-modality methods* include GIML (Meng & Wei, 2026), MARS (Baek et al., 2026), LIMSSR (Xu et al., 2026), SimMLM (Li et al., 2025), ShaSpec (Wang et al., 2023), and Flex-MoE (Yun et al., 2024). (3) *Omni comparators* include Qwen2.5-Omni 3B/7B (Xu et al., 2025), InteractiveOmni 8B (Tong et al., 2025), OmniVinci 9B (Ye et al., 2026), Gemma 4 12B IT (Abd et al., 2026), and Stream-Omni-8B (Zhang et al., 2025).

Shared model pool. For each dataset, LLUNA and all general collaboration and missing-modality baselines use the same frozen pool and identical pre-extracted representations, data splits, and modality-availability masks. They use these representations or derived task outputs, without updating or replacing upstream encoders or accessing additional raw inputs. Removing a modality jointly masks every representation derived from it. This controls information available for collaboration;

trainable parameter counts are reported separately in Appendix D.1. The main comparison uses a nine-model pool. The shared text/audio/vision pool uses three publicly available models per modality from different families. Omni baselines instead process available raw inputs with their frozen backbones. Appendix B.2 specifies the task-specific pools and interfaces.

4.2 MAIN RESULTS

We evaluate LLUNA and its competitors using Mean-Missing performance. In Table 1, **Avg. rank** summarizes cross-dataset performance: we rank all 21 methods by their Mean-Missing mean separately on each dataset, then average each method's eight ranks with equal weight. Full-input results, with all modalities available, and results for each individual missing-input state are reported in Appendix A.2. LLUNA improves MM over the strongest evaluated baseline on all eight datasets, averaging 4.26% relative improvement. For example, WS accuracy rises from 31.20% to 33.90%, an 8.65% relative gain, while SIMS MAE decreases from 0.3590 to 0.3424. These results demonstrate LLUNA's effectiveness in combining frozen unimodal models under missing modalities across classification and regression tasks.

Training with missing-input views. To test whether LLUNA's advantage persists under missing-input training, we evaluate ten representative strong baselines on one dataset per domain. The collaboration baselines receive missing-input supervision and select checkpoints by validation Mean-Missing. These controls include I^2MoE (Xin et al., 2025) and use the shared frozen model pool. The three Omni models retain frozen backbones and use a shared task head fitted on complete and missing-input views. In Table 2, LLUNA exceeds all ten controls on each task: the strongest control reaches 0.3546 versus 0.3421 MAE on SIMS, 32.64% versus 33.82% accuracy on WS, and 82.08% versus 84.17% balanced accuracy on ADNI. LLUNA remains ahead of all evaluated controls, supporting its effectiveness when the baselines are trained and selected for missing-input performance. Appendix B.4 gives the training protocol and all test-view results.

Table 2: **Missing-input training.**

Method	SIMS↓	WS↑	ADNI↑
Concat MLP	37.90	32.64	70.21
Attn. pool	35.46	32.15	78.13
Soft MoE	38.08	31.17	80.83
FoE	47.59	31.03	63.33
Set Trans.	37.42	31.17	82.08
CLARGA	40.18	31.17	80.63
I^2MoE	37.31	30.27	77.71
Qwen2.5-Omni 3B	51.55	30.51	66.88
Qwen2.5-Omni 7B	48.61	30.54	69.79
Stream-Omni	65.71	29.92	74.17
LLUNA	**34.21**	**33.82**	**84.17**

4.3 ABLATION STUDY

Private vs. common paths. To assess the contribution of each prediction path, we compare LLUNA with w/o Private and w/o Common. The former removes the direct private prediction; the latter removes the common branch. Removing either branch worsens MM on all eight tasks (Table 3a). On ADNI, balanced accuracy drops by 18.54 pp without private prediction and by 4.79 points without common collaboration, supporting the complementary roles of the two paths. To examine how the paths should be combined, Equal addition fixes the common residual scale to one. The learned combination performs better on most tasks, supporting more consistent cross-task performance by adapting the contribution of the common prediction. More analyses are in Appendix C.

Graph interaction controls. To test whether prediction compatibility helps organize model interaction, we compare the fitted compatibility graph with Uniform Graph, which assigns equal weights over all available nodes (Table 3b). LLUNA outperforms Uniform Graph on seven tasks and ties on ADNI. These results favor compatibility-guided interaction over uniform aggregation.

Consistency-pair selection. To examine which state relationships should be regularized, we compare LLUNA with w/o PCC, Full-to-Minimal, and All-Pairs, retaining the architecture and remaining training settings (Table 3c). W/o PCC removes the consistency objective. Full-to-Minimal aligns the full state directly with each single-modality state, whereas All-Pairs aligns every pair of distinct nonempty states. LLUNA's adjacent-state PCC outperforms removing PCC and both alternative pair selections on most tasks. These results favor concentrating consistency on single-modality transitions. Further details are given in Appendix C.

Direction of consistency. To test the value of directional guidance, w/o Stop-gradient retains the PCC loss but allows its gradients to update both the richer and poorer states. This worsens MM on most tasks (Table 3c), favoring richer-to-poorer guidance over mutual adjustment.

Table 3: **Ablation study.** MM follows the metrics in Table 1. The final Δ_{rel} row averages relative changes across eight tasks compared with LLUNA; negative values indicate degradation.

		(a) Paths			(b) Graph	(c) PCC				(d) Target	
Dataset	LLUNA	w/o Private	w/o Common	Equal addition	Uniform Graph	w/o PCC	Full-to-Minimal	All-Pairs	w/o Stop-gradient	PCC-Private	PCC-Both
SIMS↓	<u>34.21</u>	35.35	35.52	**33.88**	34.77	34.87	34.74	34.92	35.21	34.59	35.18
MOSI↓	**98.34**	101.41	101.40	100.21	100.74	99.87	<u>99.10</u>	99.40	103.17	99.95	99.62
MELD↑	<u>56.08</u>	55.49	55.15	**56.16**	55.20	55.96	54.58	55.08	55.93	55.96	55.08
URF↑	**70.61**	68.01	69.62	70.12	69.62	70.00	<u>70.22</u>	70.17	70.04	70.00	69.99
Daily↑	**38.42**	37.57	36.16	28.81	<u>38.14</u>	**38.42**	**38.42**	<u>38.14</u>	28.53	<u>38.14</u>	37.57
WS↑	**33.82**	32.50	32.25	33.30	<u>33.51</u>	32.15	33.30	32.46	32.54	32.36	32.29
CirCor↑	**82.56**	71.88	78.12	74.89	78.71	78.71	76.92	81.93	77.42	<u>82.26</u>	82.16
ADNI↑	**84.17**	65.63	79.38	<u>83.96</u>	**84.17**	80.83	81.04	<u>83.96</u>	**84.17**	82.50	83.33
Δ_{rel} (%)	**0.00**	−6.53	−3.95	−4.70	−1.67	−2.27	−2.20	−1.42	−5.58	−1.40	−1.88

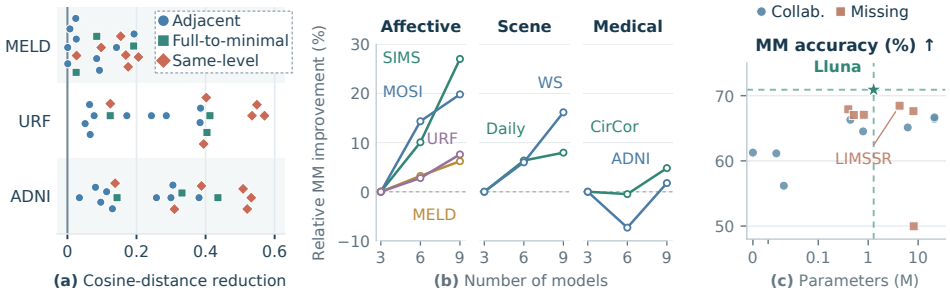

Figure 3: **Analyses of LLUNA.** (a) Reduction in common-state drift with PCC. (b) Relative MM gains over three-model pools. (c) MM versus trainable parameters on URF. Additional pool-size and parameter results are in Appendix D.1.

Consistency target. To identify where cross-state consistency is most effective, we apply PCC to the common representation (LLUNA, or PCC-Common), the private representation (PCC-Private), or both (PCC-Both), while retaining the two prediction paths (Table 3d). PCC-Common outperforms both alternatives on all eight tasks. These results support focusing consistency on the common representation while preserving a separate path for modality-specific information.

Effect of PCC on common-representation drift. To examine how PCC shapes relationships among availability states, we compare common-state cosine distances in matched models with and without PCC. For each state pair, absolute reduction is $d_{\text{off}} - d_{\text{on}}$: the pair's mean distance without PCC minus that with PCC, each averaged over the same test examples. Figure 3(a) shows individual pair responses on MELD, URF, and ADNI. Both adjacent and more distant state pairs exhibit contraction in these examples, illustrating that local consistency can reduce drift beyond the directly regularized transitions. This complements the MM evidence in Table 3c.

4.4 UNDERSTANDING MODEL COLLABORATIONS

Varying model-pool size. We vary LLUNA's pool with $3, 6, 9$ models to examine the benefit of additional models (Figure 3(b)). Nine models achieve the best observed MM, with relative improvements of 1.77%–27.02% over the three-model pools. Gains vary by task, and intermediate expansions are not uniformly beneficial. Appendix D.1 reports the scores and pool composition.

Collaboration gains. We test whether collaboration remains beneficial as inputs disappear. For each modality, we fix a unimodal reference model and measure collaboration accuracy minus its accuracy, in percentage points, across states retaining that modality. Figure 4 compares LLUNA with the equally weighted mean of seven baselines. From full to minimal inputs, the baseline mean gain declines on all eight tasks and is negative at Minimal on six. LLUNA retains positive gains on every task and exceeds the baseline mean at both endpoints, supporting beneficial collaboration under severe missingness. Appendix E.1 gives the baseline identities, aggregation protocol, and complete results.

Visualization analysis. We compare the training-estimated compatibility graphs of MELD, WS, and ADNI (Figure 5). Sparse cross-modality bridges on MELD contrast with stronger audio–visual

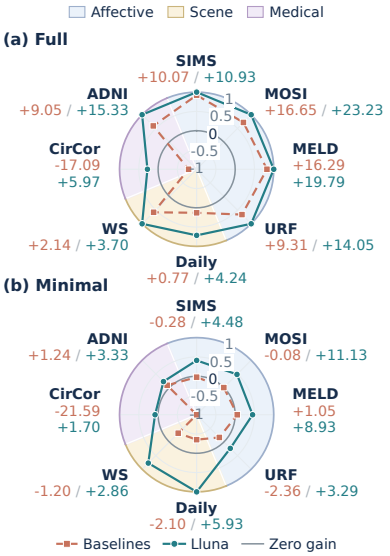

Figure 4: Collaboration gains.

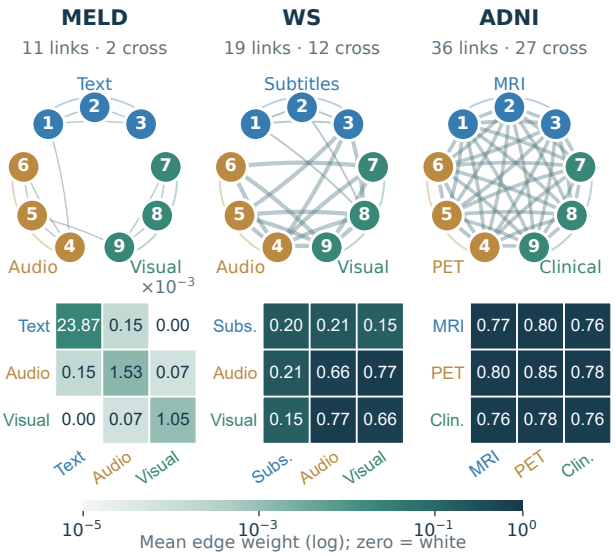

Figure 5: Model compatibility across tasks.

coupling on WS and dense connections on ADNI, illustrating task-dependent interaction patterns. Appendix B.2 provides model identities and visualization details.

4.5 Training Efficiency and Hyperparameter Sensitivity

PCC budgets and hyperparameters. To assess sensitivity to consistency coverage and model configuration, we vary the number of adjacent PCC pairs per step and use a nine-configuration design covering PCC strength, private/common width ratio, and graph depth. Larger pair budgets are not uniformly better: SIMS favors four pairs, whereas URF favors nine. Averaged by factor level, the base PCC strength performs best on all four affective tasks, while preferred widths and depths vary. These results support task-specific choices rather than uniformly increasing coverage or capacity. Appendix C gives the complete settings and results.

Trainable parameters. To examine the trainable capacity needed for strong missing-input performance, we compare MM with the number of optimized collaboration parameters, excluding frozen backbones and separately fitted frozen heads. Lluna achieves the best MM among the compared methods on all eight tasks. On URF (Figure 3(c)), it reaches 70.91% accuracy with 1.29M parameters, versus 68.46% with 4.25M for the strongest baseline, LIMSSR. This 2.45-point gain with 69.8% fewer parameters illustrates effective collaboration with moderate trainable capacity. Appendix D.1 reports the complete comparison and counting protocol.

Chain sampling. To test whether fewer PCC comparisons can preserve missing-input performance, we compare task-specific sparse configurations with full-lattice training. Their equal-task mean relative MM change is -0.28% across eight tasks. On MELD, one sampled chain reduces comparisons per step from nine to two while improving MM by 0.49% relative to full training. Thus, sparse training retains comparable average MM with fewer consistency comparisons. Appendix D.3 provides all operating points and a controlled sweep varying only the number of sampled chains.

5 Conclusion

We study collaboration among pretrained unimodal models under missing modalities and propose Lluna, which combines direct access to modality-specific features with residual common collaboration and progressive cross-state consistency. Across eight datasets, Lluna improves Mean-Missing performance over the strongest evaluated baseline on every task, averaging 4.26% relative improvement. It also maintains positive mean collaboration gains under minimal inputs, supporting robust collaboration as available evidence decreases. We discuss limitations in Appendix H.

AI USE STATEMENT

We used generative AI tools to assist with manuscript revision, literature search, and code development. The authors take responsibility for the accuracy and integrity of the final manuscript.

REPRODUCIBILITY STATEMENT

Dataset protocols, preprocessing, model pools, and training hyperparameters are provided in Appendix B. Appendix F details graph construction and PCC implementation, while Appendix A.1 specifies the evaluation and aggregation procedures. The assumptions and proofs of the theoretical results are given in Appendix G. Code available at https://anonymous.4open.science/r/Lluna/README.md.

REFERENCES

Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Carbune, Michelle Casbon, Mayank Chaturvedi, Aditya Chawla, Victor Cotruta, Alice Coucke, Phil Culliton, Robert Dadashi, Lucas Dixon, Mohamed Elhawaty, Utku Evci, Clément Farabet, Johan Ferret, Filippo Galgani, Sertan Girgin, Jean-Bastien Grill, Maarten Grootendorst, Jiaxian Guo, Cassidy Hardin, Yanzhang He, Steven M. Hernandez, Omri Homburger, Léonard Hussenot, Ju-yeong Ji, Armand Joulin, Aishwarya Kamath, Parnian Kassraie, Olivier Lacombe, Preethi Lahoti, Gaël Liu, Gus Martins, Luciano Martins, Tatiana Matejovicova, Ramona Merhej, Nikola Momchev, Sneha Mondal, Ryan Mullins, Sindhu Raghuram Panyam, Shreya Pathak, Sarah Perrin, André Susano Pinto, Etienne Pot, Angéline Pouget, Alexandre Ramé, Sabela Ramos, Douglas Reid, David Rim, Morgane Rivière, Karsten Roth, Louis Rouillard, Omar Sanseviero, Pier Giuseppe Sessa, Shane Settle, Danila Sinopalnikov, Sara Smoot, Piotr Stanczyk, Andreas Steiner, Lawrence Stewart, Ilya O. Tolstikhin, Michael Tschannen, Anton Tsitsulin, Nino Vieillard, Renjie Wu, Pingmei Xu, Haichuan Yang, Edouard Yvinec, Biao Zhang, Li Zhang, Joe Zou, Nicolas Aagnes, Abdelrahman Abdelhamed, Jakub Adámek, Shivani Agrawal, Shubham Agrawal, Ibrahim Alabdulmohsin, Jean-Baptiste Alayrac, Uri Alon, Chandramouli Amarnath, Ankesh Anand, Chrysovalantis Anastasiou, Setareh Ariafar, François-Xavier Aubet, Kyriakos Axiotis, Federico Barbero, Joelle K. Barral, Alexei Bendebury, Urs Bergmann, Stanley Bileschi, Kat Black, Mathieu Blondel, Sebastian Borgeaud, and Arthur Brazinskas. Gemma 4 technical report. *CoRR*, abs/2607.02770, 2026. doi: 10.48550/ARXIV.2607.02770. URL `https://doi.org/10.48550/arXiv.2607.02770`.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Thomas Bachlechner, Bodhisattwa Prasad Majumder, Henry Mao, Gary Cottrell, and Julian McAuley. Rezero is all you need: Fast convergence at large depth. In *Uncertainty in artificial intelligence*, pp. 1352–1361. PMLR, 2021.

Seunghun Baek, Jihwan Park, Jaeyoon Sim, Minjae Jeong, Hoseok Lee, and Won Hwa Kim. Residual-guided expert specialization for incomplete multimodal learning. In *European Conference on Computer Vision*, pp. 76–93. Springer, 2026.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2): 423–443, 2018.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *International conference on machine learning*, pp. 1263–1272. Pmlr, 2017.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind one embedding space to bind them all. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15180–15190. IEEE, 2023.

Md Kamrul Hasan, Wasifur Rahman, AmirAli Bagher Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, and Mohammed Ehsan Hoque. Ur-funny: A multimodal language dataset for understanding humor. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 2046–2056, 2019.

Mohammad Havaei, Nicolas Guizard, Nicolas Chapados, and Yoshua Bengio. Hemis: Hetero-modal image segmentation. In *International conference on medical image computing and computer-assisted intervention*, pp. 469–477. Springer, 2016.

Jack Hong, Shilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Weidi Xie. Worldsense: Evaluating real-world omnimodal understanding for multimodal llms. In *International Conference on Learning Representations*, volume 2026, pp. 52423–52443, 2026.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. Metagpt: Meta programming for A multi-agent collaborative framework. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=VtmBAGCN7o.

Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International conference on machine learning*, pp. 3744–3753. PMLR, 2019.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for" mind" exploration of large language model society. *Advances in neural information processing systems*, 36:51991–52008, 2023.

Sijie Li, Chen Chen, and Jungong Han. Simmlm: A simple framework for multi-modal learning with missing modality. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 24068–24077. IEEE, 2025.

Yihe Liu, Ziqi Yuan, Huisheng Mao, Zhiyun Liang, Wanqiuyue Yang, Yuanzhe Qiu, Tie Cheng, Xiaoteng Li, Hua Xu, and Kai Gao. Make acoustic and visual cues matter: Ch-sims v2. 0 dataset and av-mixup consistent module. In *Proceedings of the 2022 international conference on multimodal interaction*, pp. 247–258, 2022.

Jinliang Lu, Ziliang Pang, Min Xiao, Yaochen Zhu, Rui Xia, and Jiajun Zhang. Merge, ensemble, and cooperate! a survey on collaborative strategies in the era of large language models. *arXiv preprint arXiv:2407.06089*, 2024.

Yinyi Luo, Wenwen Wang, Hayes Bai, Hongyu Zhu, Hao Chen, Pan He, Marios Savvides, Sharon Li, and Jindong Wang. Torchumm: A unified multimodal model codebase for evaluation, analysis, and post-training. *arXiv preprint arXiv:2604.10784*, 2026.

Mayug Maniparambil, Raiymbek Akshulakov, Yasser Abdelaziz Dahou Djilali, Sanath Narayan, Ankit Singh, and Noel E O'Connor. Harnessing frozen unimodal encoders for flexible multimodal alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29847–29857, 2025.

Xiangyu Meng and Shicai Wei. General incomplete multimodal learning via dynamic quality perception. *arXiv preprint arXiv:2607.06943*, 2026.

Susanne G Mueller, Michael W Weiner, Leon J Thal, Ronald C Petersen, Clifford Jack, William Jagust, John Q Trojanowski, Arthur W Toga, and Laurel Beckett. The alzheimer's disease neuroimaging initiative. *Neuroimaging Clinics of North America*, 15(4):869, 2005.

Natalia Neverova, Christian Wolf, Graham Taylor, and Florian Nebout. Moddrop: adaptive multi-modal gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38 (8):1692–1706, 2015.

Jorge Oliveira, Francesco Renna, Paulo Dias Costa, Marcelo Nogueira, Cristina Oliveira, Carlos Ferreira, Alípio Jorge, Sandra Mattos, Thamine Hatem, Thiago Tavares, et al. The circor digiscope dataset: from murmur detection to murmur classification. *IEEE journal of biomedical and health informatics*, 26(6):2524–2535, 2021.

Cailean Osborne, Jennifer Ding, and Hannah Rose Kirk. The ai community building the future? a quantitative analysis of development activity on hugging face hub. *Journal of Computational Social Science*, 7(2):2067–2105, 2024.

Santosh Patapati and Trisanth Srinivasan. Multimodal graph representation learning over arbitrary sets of modalities. In *2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 7104–7115. IEEE, 2026.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 527–536, 2019.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pp. 28492–28518. PMLR, 2023.

Yashar Talebirad and Amirhossein Nadiri. Multi-agent collaboration: Harnessing the power of intelligent llm agents. *arXiv preprint arXiv:2306.03314*, 2023.

Wenwen Tong, Hewei Guo, Dongchuan Ran, Jiangnan Chen, Jiefan Lu, Kaibin Wang, Keqiang Li, Xiaoxu Zhu, Jiakui Li, Kehan Li, et al. Interactiveomni: A unified omni-modal model for audio-visual multi-turn dialogue. *arXiv preprint arXiv:2510.13747*, 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Hongyi Wang, Felipe Maia Polo, Yuekai Sun, Souvik Kundu, Eric Xing, and Mikhail Yurochkin. Fusing models with complementary expertise. In *International Conference on Learning Representations*, volume 2024, pp. 45284–45306, 2024a.

Hu Wang, Yuanhong Chen, Congbo Ma, Jodie Avery, Louise Hull, and Gustavo Carneiro. Multi-modal learning with missing modality via shared-specific feature modelling. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15878–15887. IEEE, 2023.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024b.

David H Wolpert. Stacked generalization. *Neural networks*, 5(2):241–259, 1992.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, et al. Autogen: Enabling next-gen llm applications via multi-agent conversation. *arXiv preprint arXiv:2308.08155*, 2023.

Renjie Wu, Hu Wang, Hsiang-Ting Chen, and Gustavo Carneiro. Deep multimodal learning with missing modality: A survey. *arXiv preprint arXiv:2409.07825*, 2024.

Dongqing Xie, Yonghuang Wu, Zisheng Ai, Jun Min, Zhencun Jiang, Shaojin Geng, and Lei Wang. Ccsd: Cross-modal compositional self-distillation for robust brain tumor segmentation with missing modalities. *Knowledge-Based Systems*, pp. 116136, 2026.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. In *International Conference on Learning Representations*, volume 2025, pp. 28240–28264, 2025.

Jiayi Xin, Sukwon Yun, Jie Peng, Inyoung Choi, Jenna L Ballard, Tianlong Chen, and Qi Long. I2moe: Interpretable multimodal interaction-aware mixture-of-experts. *Proceedings of machine learning research*, 267:68870, 2025.

Huangbiao Xu, Huanqi Wu, Xiao Ke, and Yuxin Peng. Limssr: Llm-driven sequence-to-score reasoning under training-time incomplete multimodal observations. *arXiv preprint arXiv:2605.00434*, 2026.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report. *CoRR*, abs/2503.20215, 2025. doi: 10.48550/ARXIV.2503.20215. URL https://doi.org/10.48550/arXiv.2503.20215.

Prateek Yadav, Colin Raffel, Mohammed Muqeeth, Lucas Caccia, Haokun Liu, Tianlong Chen, Mohit Bansal, Leshem Choshen, and Alessandro Sordoni. A survey on model moerging: Recycling and routing among specialized experts for collaborative learning. *arXiv preprint arXiv:2408.07057*, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Hanrong Ye, Chao-Han Huck Yang, Arushi Goel, Wei Huang, Zhen Wan, Jinchuan Tian, An-Chieh Cheng, Ligeng Zhu, Yuanhang Su, Yuming Lou, et al. Omnivinci: Enhancing architecture and data for omni-modal understanding llm. In *International Conference on Learning Representations*, volume 2026, pp. 56101–56138, 2026.

Sukwon Yun, Inyoung Choi, Jie Peng, Yangfan Wu, Jingxuan Bao, Qiyiwen Zhang, Jiayi Xin, Qi Long, and Tianlong Chen. Flex-moe: Modeling arbitrary modality combination via the flexible mixture-of-experts. *Advances in Neural Information Processing Systems*, 37:98782–98805, 2024.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*, 2016.

Shaolei Zhang, Shoutao Guo, Qingkai Fang, Yan Zhou, and Yang Feng. Stream-omni: Simultaneous multimodal interactions with large language-vision-speech model. *arXiv preprint arXiv:2506.13642*, 2025.

Ziwei Zhou, Rui Wang, Zuxuan Wu, and Yu-Gang Jiang. Daily-omni: Towards audio-visual reasoning with temporal alignment across modalities. *arXiv preprint arXiv:2505.17862*, 2025.

APPENDIX CONTENTS

A Complete Five-Run Evaluation **15**
A.1 Selection and aggregation protocol . . . 15
A.2 All methods under every availability state . . . 16

B Main Experimental Setup **24**
B.1 Tasks, splits, and availability states . . . 24
B.2 Frozen inputs and baseline interfaces . . . 25
B.3 Main-run optimization settings . . . 28
B.4 Missing-input training controls . . . 28

C Additional Ablation Results **31**

D Additional Analyses **33**
D.1 Model-Pool and Parameter Analyses . . . 33
D.2 Modality-Utility Results . . . 39
D.3 Chain-Sampling Results . . . 41

E Motivation Diagnostics **42**
E.1 Fixed-anchor collaboration gains . . . 42
E.2 Private-signal recoverability . . . 48
E.3 Representation changes under modality removal . . . 49
E.4 Prediction changes across availability states . . . 49

F Implementation and Reproducibility **51**
F.1 PCC weighting and discrepancy configuration . . . 51

G Additional Derivations **52**
G.1 Paths between arbitrary modality subsets . . . 52
G.2 Local-to-global representation stability . . . 52
G.3 Supplementary output-drift bound . . . 53

H Limitations **54**

I Broader Impact **54**

A Complete Five-Run Evaluation

A.1 Selection and aggregation protocol

Table 1 reports five training runs for each trainable method. Every retrained baseline uses seeds $11, 47, 103, 211, 389$. Each dataset uses one adopted LLUNA configuration across its five runs; all five runs contribute to the reported mean and standard deviation. Omni baselines use fixed backbone/head checkpoints with deterministic inference; their repeated evaluations reproduce the same score, reported without a training SD.

Mean-Missing and relative improvement. Let $\mathcal{V}_d^-$ be the legal, nonempty, non-full availability states of dataset d, and let $m_{d,c}(S)$ be the task score for run c in state S. Mean-Missing is computed within each run before averaging runs:

$$\mathrm{MM}_{d,c} = \frac{1}{|\mathcal{V}_d^-|} \sum_{S \in \mathcal{V}_d^-} m_{d,c}(S). \tag{5}$$

Each main-table cell reports the corresponding method's MM result. Relative improvement uses the best baseline mean for that dataset. For an individual view, the reference is the strongest baseline mean in that view. Thus, baseline MM is not an average of the different per-view winners. For the displayed five-run mean $\bar{m}$ and baseline score B, relative improvement Δ_{rel} (in percent) is

$$\Delta_{\mathrm{rel}} = \begin{cases} 100(B - \bar{m})/B, & \text{MAE (lower is better)}, \\ 100(\bar{m} - B)/B, & \text{F1 or accuracy (higher is better)}. \end{cases} \tag{6}$$

Positive Δ_{rel} indicates improvement; a negative value indicates a shortfall relative to the strongest baseline. In all result tables, MAE means and SDs are multiplied by 100. F1 and accuracy means are percentages, with SDs in percentage points. Every mean and SD uses two decimal places; for MAE this retains the precision of four decimals in native units. The relative improvement reported in the main text averages Δ_{rel} equally across the eight datasets, using unrounded values. The last cell of the Δ_{rel} row in Table 1 reports this same mean in percent. The Avg. rank column averages each method's rank across the eight datasets. On each dataset, all 21 methods are ranked by their unrounded Mean-Missing means. Rank 1 is best: lower MAE and higher F1 or accuracy receive better ranks. Exact ties receive the average of their occupied positions. We rank the five-run means, not individual runs; every dataset receives equal weight. This summarizes cross-dataset standing without averaging incompatible metrics.

View notation. For CH-SIMS v2, CMU-MOSI, MELD, and UR-FUNNY v2, T/A/V denote text/audio/vision. Daily-Omni uses A/V for audio/video. WorldSense uses V/A/S for visual/audio/subtitles. ADNI uses M/P/C for MRI/PET/clinical inputs, including states without clinical inputs. CirCor-4Site uses AV/PV/TV/MV for four auscultation sites; these are site identifiers, not audio/visual modality abbreviations. Required question-and-choice context is retained for the scene-understanding tasks while the listed modalities vary.

A.2 ALL METHODS UNDER EVERY AVAILABILITY STATE

Tables 4–11 give the complete method-by-state results for all eight datasets. Bold marks the best mean and underlining the second-best mean in each column, using unrounded values and preserving ties. These marks indicate rank rather than statistical significance.

Table 4: **SIMS: complete results.** (Liu et al., 2022) MAE $\times 100 \downarrow$. Trained models: five-run mean $\pm$ sample SD; fixed models: single score. F denotes full input; MM averages non-full states.

Method	TAV^F	TA	TV	AV	T	A	V	MM
GENERAL COLLABORATION								
Mean ens.	53.24±0.02	53.52±0.02	53.20±0.02	53.00±0.02	53.72±0.02	53.31±0.03	52.71±0.03	53.24±0.02
Logit stack	43.39±0.12	46.37±1.32	49.59±1.82	47.31±0.55	51.82±1.40	48.92±0.14	53.69±2.43	49.62±0.93
Concat MLP	33.37±0.37	37.18±1.60	40.82±2.16	38.80±0.63	42.39±1.13	44.71±0.33	49.71±2.95	42.27±0.73
Attn. pool	29.39±0.42	32.05±0.61	37.93±1.58	34.53±1.04	41.55±0.92	39.59±1.17	47.55±0.59	38.87±0.75
Soft MoE	32.50±0.76	51.42±7.17	44.68±3.68	37.40±3.71	91.03±22.57	58.75±13.70	65.97±10.85	58.21±4.88
FoE	43.02±0.25	52.63±3.89	53.94±5.75	50.68±3.39	57.39±6.81	53.80±4.81	54.86±2.25	53.88±2.27
Set Trans.	31.21±0.52	34.87±0.86	41.47±1.97	36.63±1.02	48.80±5.38	42.99±1.72	52.95±1.75	42.95±0.73
CLARGA	34.12±0.41	36.77±0.25	38.35±0.61	36.64±0.44	41.99±0.36	39.74±0.32	46.47±0.66	39.99±0.34
MISSING-MODALITY METHODS								
GIML	37.18±0.97	41.76±3.45	50.96±11.33	58.26±33.74	88.41±48.71	63.18±21.87	105.96±61.01	68.09±20.73
MARS	33.38±4.65	38.07±4.98	38.26±3.35	37.30±4.58	43.06±2.08	42.83±6.95	45.47±1.82	40.83±3.79
LIMSSR	29.13±0.14	31.92±0.49	33.91±0.47	32.05±0.27	38.48±0.53	35.36±0.32	43.66±0.85	35.90±0.21
SimMLM	33.84±0.72	35.84±0.31	37.42±0.98	36.05±0.90	41.32±1.33	39.99±0.47	47.25±1.56	39.64±0.43
ShaSpec	30.89±0.29	33.69±0.32	35.87±0.38	32.88±0.70	40.79±0.45	37.35±0.52	43.76±0.51	37.39±0.27
Flex-MoE	35.26±1.08	37.46±2.52	41.73±8.54	38.42±4.38	41.63±1.69	40.96±3.11	59.33±29.04	43.25±7.98
OMNI MODELS								
Q2.5-O-3B	56.67	110.33	71.76	188.61	134.87	223.27	302.29	171.86
Q2.5-O-7B	68.39	79.44	78.59	111.14	106.07	85.78	106.33	94.56
InterO-8B	100.85	115.39	122.96	99.45	506.63	123.59	126.27	182.38
OmniV-9B	89.49	128.16	100.19	95.36	130.77	90.84	92.17	106.25
Gemma4-12B	69.27	326.82	118.87	75.83	1100.45	452.45	124.85	366.55
StreamO-8B	81.51	253.89	151.98	101.33	270.40	83.02	156.18	169.47
LLUNA	**28.08**±0.46	**31.52**±0.62	**31.60**±0.46	**29.98**±0.21	**38.15**±0.40	**34.45**±0.65	**39.74**±0.57	**34.24**±0.30

Table 5: **MOSI: complete results.** (Zadeh et al., 2016) MAE $\times 100 \downarrow$. Trained models: five-run mean $\pm$ sample SD; fixed models: single score. F denotes full input; MM averages non-full states.

Method	TAV^{F}	TA	TV	AV	T	A	V	MM
GENERAL COLLABORATION								
Mean ens.	145.70±0.25	145.35±0.23	146.12±0.25	145.62±0.26	145.85±0.23	144.88±0.24	146.41±0.27	145.71±0.25
Logit stack	138.67±0.70	139.49±5.69	147.24±5.65	140.99±4.40	146.92±7.75	140.23±2.02	149.86±7.89	144.12±2.54
Concat MLP	111.99±10.96	116.53±18.79	110.09±11.13	145.14±9.84	98.70±14.06	158.94±14.80	151.85±4.98	130.21±4.21
Attn. pool	86.91±2.00	85.33±3.31	92.72±4.78	132.37±20.39	94.51±5.30	123.60±15.17	162.73±12.85	115.21±6.28
Soft MoE	91.53±2.75	99.92±6.34	101.44±10.89	148.52±13.77	129.56±37.78	162.77±20.77	160.94±13.87	133.86±13.86
FoE	137.19±1.75	148.69±21.35	149.40±20.31	142.62±5.96	147.86±10.77	149.39±18.20	150.13±8.07	148.02±7.96
Set Trans.	109.41±7.49	101.92±5.69	114.51±2.81	164.81±32.12	110.01±9.45	154.98±35.11	185.01±18.98	138.54±12.45
CLARGA	108.77±2.68	102.97±4.00	118.58±5.26	128.12±4.31	108.70±6.24	116.69±1.91	162.33±5.52	122.90±2.63
MISSING-MODALITY METHODS								
GIML	101.97±4.43	114.13±15.74	122.10±33.72	205.84±175.43	169.07±86.79	216.65±147.07	261.91±211.86	181.62±96.15
MARS	113.68±29.33	110.70±28.88	119.06±23.97	153.22±19.35	115.90±28.19	138.12±13.62	158.47±12.44	132.58±16.83
LIMSSR	83.12±1.81	81.22±1.96	88.23±2.52	102.55±1.81	86.28±2.88	99.80±2.37	155.98±6.18	102.34±0.56
SimMLM	103.65±3.14	98.56±2.76	110.93±5.61	116.15±2.78	102.40±7.27	111.46±2.25	157.12±3.61	116.10±3.07
ShaSpec	100.84±3.99	96.37±4.13	111.89±4.20	121.22±3.32	103.46±4.68	111.07±4.41	156.93±2.01	116.82±3.38
Flex-MoE	112.98±3.23	105.63±6.73	131.93±20.38	126.68±5.29	109.83±7.37	119.18±7.59	158.77±9.21	125.34±7.87
OMNI MODELS								
Q2.5-O-3B	61.60	84.63	82.39	123.39	345.11	213.42	121.18	161.69
Q2.5-O-7B	**59.33**	80.88	**80.69**	**73.50**	173.18	**78.81**	110.98	99.67
InterO-8B	113.02	165.92	136.56	109.28	104.15	162.37	135.09	135.56
OmniV-9B	77.36	121.04	83.91	91.18	96.18	135.97	**97.73**	104.34
Gemma4-12B	67.96	150.36	193.24	102.94	265.61	197.35	253.80	193.88
StreamO-8B	65.54	432.38	105.13	108.56	590.18	831.16	255.07	387.08
LLUNA	78.36±1.17	**77.34**±1.09	83.03±0.98	95.38±2.36	**82.01**±0.88	94.75±2.41	147.72±1.63	**96.71**±0.63

Table 6: **MELD: complete results.** (Poria et al., 2019) Weighted F1 (%) ↑. Trained models: five-run mean ± sample SD; fixed models: single score. F denotes full input; MM averages non-full states.

Method	TAV^{F}	TA	TV	AV	T	A	V	MM
GENERAL COLLABORATION								
Mean ens.	31.31±0.00	31.31±0.00	31.31±0.00	31.31±0.00	31.31±0.00	31.31±0.00	31.31±0.00	31.31±0.00
Logit stack	52.23±0.10	43.76±1.63	12.71±1.67	38.24±1.45	7.83±3.22	31.59±0.28	2.10±0.00	22.70±0.57
Concat MLP	60.99±0.51	62.58±0.62	50.59±5.16	38.64±2.06	51.51±4.98	34.65±1.57	26.82±4.27	44.13±2.98
Attn. pool	62.16±0.48	61.06±2.42	59.11±1.53	44.31±1.26	56.57±3.93	46.72±1.20	31.78±2.92	49.93±0.88
Soft MoE	62.45±0.24	62.77±0.71	60.75±1.14	38.25±4.48	61.35±0.71	37.49±8.08	22.05±7.09	47.11±1.33
FoE	31.31±0.00	31.31±0.00	31.31±0.00	31.31±0.00	31.31±0.00	31.31±0.00	31.31±0.00	31.31±0.00
Set Trans.	61.13±0.36	60.23±1.03	57.07±2.25	43.33±1.12	56.77±2.44	44.04±1.32	28.49±5.08	48.32±1.38
CLARGA	60.90±0.86	61.20±0.78	59.51±0.88	48.95±1.48	59.58±1.18	49.04±1.19	36.60±0.65	52.48±0.81
MISSING-MODALITY METHODS								
GIML	60.16±1.24	33.78±17.01	43.37±11.26	31.31±0.00	31.25±16.28	31.31±0.00	31.31±0.00	33.72±6.49
MARS	59.52±0.47	61.47±0.73	57.87±0.60	47.08±0.88	60.19±1.08	49.02±0.94	35.98±0.45	51.94±0.30
LIMSSR	62.77±1.43	63.11±1.21	61.53±0.94	51.39±0.45	61.72±0.57	51.33±0.36	34.84±0.86	53.98±0.33
SimMLM	61.38±0.51	61.54±0.53	60.16±0.93	49.66±0.78	60.13±0.86	50.24±0.66	36.81±0.26	53.09±0.25
ShaSpec	61.27±0.68	61.23±0.48	60.09±1.02	49.58±0.68	60.45±0.56	49.87±0.44	36.30±0.53	52.92±0.48
Flex-MoE	60.91±0.57	60.84±0.70	58.57±0.44	44.47±4.16	59.01±1.28	45.70±2.80	33.50±1.04	50.35±1.47
OMNI MODELS								
Q2.5-O-3B	52.42	2.95	46.96	33.43	2.48	23.24	26.10	22.53
Q2.5-O-7B	54.58	47.41	49.73	35.88	40.16	35.77	28.52	39.58
InterO-8B	40.40	26.74	32.80	38.80	4.38	26.91	29.51	26.52
OmniV-9B	49.63	25.62	47.12	36.49	35.13	29.63	32.33	34.39
Gemma4-12B	51.90	31.05	38.94	27.63	31.50	26.97	27.10	30.53
StreamO-8B	52.10	23.34	20.09	32.62	17.28	0.97	9.06	17.23
LLUNA	**64.55**±0.18	**64.75**±0.19	**63.19**±0.36	**53.54**±0.59	**63.60**±0.27	**53.63**±0.63	**37.86**±0.57	**56.10**±0.21

Table 7: **URF: complete results.** (Hasan et al., 2019) Accuracy (%) ↑. Trained models: five-run mean ± sample SD; fixed models: single score. F denotes full input; MM averages non-full states.

Method	TAV^F	TA	TV	AV	T	A	V	MM
GENERAL COLLABORATION								
Mean ens.	62.58±2.13	63.82±1.69	63.32±0.87	59.34±0.74	67.04±0.45	59.40±0.60	54.63±0.61	61.26±0.54
Logit stack	67.16±0.35	67.75±0.83	66.80±0.44	56.00±4.00	66.80±0.83	55.17±3.39	53.00±1.44	60.92±1.07
Concat MLP	75.67±1.26	72.56±2.13	71.01±1.91	61.49±1.09	73.42±1.07	56.16±2.92	52.64±0.89	64.54±1.37
Attn. pool	77.26±0.97	76.94±0.79	74.51±0.61	60.60±2.74	73.80±0.58	60.95±2.39	52.78±1.16	66.60±0.98
Soft MoE	76.52±1.04	76.22±0.91	74.10±1.32	60.58±1.57	74.00±1.39	59.64±1.57	53.52±0.75	66.34±0.49
FoE	65.92±1.21	54.57±6.68	55.81±7.35	49.58±0.63	52.96±5.86	51.83±2.90	49.58±0.63	52.39±2.28
Set Trans.	75.31±0.48	71.13±3.24	70.00±1.53	61.13±1.42	71.47±1.64	57.10±3.19	51.73±1.25	63.76±0.89
CLARGA	73.10±0.75	73.50±0.86	71.33±1.58	63.42±0.52	72.21±0.87	64.02±1.11	53.48±0.42	66.33±0.67
MISSING-MODALITY METHODS								
GIML	73.68±0.97	55.47±5.56	61.95±9.20	51.35±0.95	53.52±6.76	50.14±0.77	51.33±2.27	53.96±3.06
MARS	73.40±0.31	75.69±1.42	72.45±0.59	62.54±0.90	74.57±1.37	63.78±0.97	55.81±0.55	67.47±0.50
LIMSSR	76.72±0.85	76.72±0.56	74.63±1.01	66.16±1.92	74.69±0.51	66.72±1.36	54.25±1.82	68.86±0.82
SimMLM	75.13±1.14	75.03±0.65	72.94±1.20	65.92±0.70	72.60±0.61	65.90±0.47	55.92±0.91	68.05±0.43
ShaSpec	74.55±1.24	74.27±1.03	72.45±1.29	64.49±1.18	72.86±1.13	64.83±1.83	54.63±1.02	67.25±0.70
Flex-MoE	74.08±1.47	74.10±0.79	72.09±1.75	64.61±1.99	71.85±1.16	64.95±1.99	54.51±0.96	67.02±1.27
OMNI MODELS								
Q2.5-O-3B	73.84	58.75	58.75	55.13	51.11	51.61	54.63	55.00
Q2.5-O-7B	69.72	57.95	58.45	55.13	50.70	50.00	51.21	53.91
InterO-8B	56.14	50.00	50.80	56.14	49.30	51.61	50.30	51.36
OmniV-9B	66.90	56.24	60.06	58.15	53.02	56.14	53.42	56.17
Gemma4-12B	64.99	51.01	53.82	58.85	48.49	54.23	51.21	52.93
StreamO-8B	59.05	51.11	58.15	50.20	53.02	50.70	48.59	51.96
LLUNA	**77.61**±0.51	**77.38**±0.22	**75.88**±0.39	**67.67**±0.97	**75.59**±0.54	**69.17**±0.46	**58.03**±0.42	**70.62**±0.23

Table 8: **Daily: complete results.** (Zhou et al., 2025) Accuracy (%) ↑. Trained models: five-run mean ± sample SD; fixed models: single score. F denotes full input; MM averages non-full states.

Method	AV^F	A	V	MM
GENERAL COLLABORATION				
Mean ens.	34.12±0.31	34.46±0.56	31.86±1.36	33.16±0.81
Logit stack	33.90±0.69	34.12±1.17	26.10±6.89	30.11±3.73
Concat MLP	33.22±1.09	33.22±1.01	29.15±3.16	31.19±1.85
Attn. pool	33.11±1.90	32.88±2.27	27.12±2.00	30.00±1.10
Soft MoE	32.77±1.44	32.77±1.44	26.89±3.14	29.83±2.02
FoE	33.11±0.95	33.56±0.95	30.06±4.42	31.81±1.92
Set Trans.	32.43±1.17	32.43±0.86	26.89±5.03	29.66±2.39
CLARGA	31.41±1.53	31.53±1.76	27.46±1.72	29.49±0.95
MISSING-MODALITY METHODS				
GIML	35.25±0.86	34.01±1.35	31.86±1.30	32.94±1.29
MARS	34.12±2.38	33.33±3.29	29.60±2.58	31.47±1.61
LIMSSR	31.86±1.30	31.98±1.42	25.20±5.66	28.59±2.34
SimMLM	32.43±1.63	32.32±1.67	28.02±1.68	30.17±1.47
ShaSpec	31.75±2.01	31.86±1.94	29.72±1.53	30.79±1.15
Flex-MoE	33.33±1.32	33.33±1.60	27.34±5.37	30.34±2.33
OMNI MODELS				
Q2.5-O-3B	47.46	32.20	<u>39.55</u>	35.88
Q2.5-O-7B	**55.93**	31.07	**42.37**	<u>36.72</u>
InterO-8B	25.99	25.42	24.29	24.86
OmniV-9B	35.59	33.33	30.51	31.92
Gemma4-12B	<u>50.28</u>	**37.29**	27.68	32.49
StreamO-8B	27.12	24.86	26.55	25.71
LLUNA	36.27±0.62	<u>36.61</u>±1.09	37.97±2.24	**37.29**±0.96

Table 9: **WS: complete results.** (Hong et al., 2026) Accuracy (%) ↑. Trained models: five-run mean ± sample SD; fixed models: single score. F denotes full input; MM averages non-full states.

Method	VAS^{F}	VA	VS	AS	V	A	S	MM
GENERAL COLLABORATION								
Mean ens.	31.97±0.27	29.08±0.44	32.55±0.19	32.34±0.24	29.00±0.41	29.04±0.31	**34.60**±0.32	31.10±0.12
Logit stack	33.85±0.62	28.95±0.48	33.35±0.64	**33.81**±0.50	29.16±0.94	28.74±0.60	33.18±0.70	31.20±0.15
Concat MLP	32.47±0.40	29.46±0.83	32.34±0.50	32.47±0.27	28.12±2.13	28.45±2.47	32.30±0.35	30.52±0.65
Attn. pool	32.51±0.57	29.21±1.47	32.51±0.67	32.30±0.35	28.33±0.83	27.32±2.96	32.72±0.60	30.40±0.62
Soft MoE	33.05±0.80	28.12±1.79	33.18±0.76	33.10±0.70	28.20±1.37	25.69±2.36	33.05±0.90	30.22±0.99
FoE	33.89±0.61	27.62±2.83	33.47±0.33	33.31±0.62	25.40±4.49	25.65±4.02	33.05±0.51	29.75±1.81
Set Trans.	32.55±0.67	24.56±4.23	32.85±1.36	32.47±0.63	25.40±4.88	27.62±3.99	31.30±0.67	29.03±2.09
CLARGA	32.68±0.65	29.62±0.76	32.18±0.40	31.76±0.68	30.17±1.22	28.54±0.70	32.09±0.43	30.73±0.27
MISSING-MODALITY METHODS								
GIML	33.89±0.71	28.87±0.30	**34.27**±0.50	33.10±0.18	29.04±0.34	28.41±0.18	31.84±0.63	30.92±0.18
MARS	32.93±0.88	27.74±1.31	32.64±0.71	32.76±1.02	27.45±0.63	29.41±2.24	32.09±0.92	30.35±0.58
LIMSSR	32.18±0.54	28.03±1.00	32.47±0.40	32.01±0.82	28.08±1.18	26.57±2.63	32.30±0.72	29.91±0.84
SimMLM	33.26±0.90	28.83±0.62	33.26±0.80	32.93±0.76	28.79±0.89	28.95±0.87	32.80±0.77	30.93±0.43
ShaSpec	32.68±0.23	28.91±0.85	32.80±0.27	32.85±0.36	28.45±0.91	29.87±0.72	32.80±0.18	30.95±0.23
Flex-MoE	32.43±0.33	28.33±0.87	32.55±0.32	32.43±0.49	27.03±2.72	26.23±2.53	32.34±0.38	29.82±0.80
OMNI MODELS								
Q2.5-O-3B	33.05	29.08	31.38	26.99	31.59	24.69	27.20	28.49
Q2.5-O-7B	**37.66**	**36.40**	32.85	24.27	32.85	26.57	25.31	29.71
InterO-8B	29.08	24.06	21.55	20.92	24.27	24.06	22.80	22.94
OmniV-9B	28.45	20.92	25.52	25.73	23.85	32.85	24.90	25.63
Gemma4-12B	33.26	28.03	31.80	26.78	32.43	29.08	32.22	30.06
StreamO-8B	31.38	30.75	23.01	27.41	31.38	30.75	26.15	28.24
LLUNA	33.01±0.60	36.15±0.90	33.31±0.94	32.97±0.69	**35.36**±0.26	**33.26**±0.44	32.34±0.35	**33.90**±0.30

Table 10: **CirCor: complete results.** (Oliveira et al., 2021) Balanced accuracy (%) ↑. Panel 1/2. Trained models: five-run mean ± sample SD; fixed models: single score. F denotes full input; MM averages non-full states. AV/PV/TV/MV denote auscultation sites.

Method	AV+PV+ TV+MV^{F}	AV+PV+ TV	AV+PV+ MV	AV+TV+ MV	PV+TV+ MV	AV+PV	AV+TV	AV+MV
GENERAL COLLABORATION								
Mean ens.	50.00±0.00	50.00±0.00	50.00±0.00	50.00±0.00	50.00±0.00	50.00±0.00	50.00±0.00	50.00±0.00
Logit stack	67.38±6.57	51.98±2.75	53.14±7.03	54.83±9.32	54.03±8.62	51.11±2.48	50.57±1.28	52.08±4.65
Concat MLP	82.60±3.37	83.19±1.77	78.89±2.39	74.57±3.13	83.43±2.01	78.43±3.54	73.67±1.33	66.81±2.69
Attn. pool	80.17±4.95	80.06±4.47	73.46±11.63	76.24±1.67	80.19±3.10	72.27±10.61	76.83±1.36	60.81±7.37
Soft MoE	76.27±2.90	76.56±2.92	67.17±10.02	75.02±2.16	76.41±2.98	69.30±10.90	74.87±2.08	59.27±6.63
FoE	50.00±0.00	48.84±2.59	50.00±0.00	50.00±0.00	50.00±0.00	50.00±0.00	50.00±0.00	50.00±0.00
Set Trans.	79.22±3.86	77.32±3.84	71.70±11.19	75.40±1.46	78.94±4.82	70.16±8.77	74.08±2.16	65.70±10.41
CLARGA	79.76±3.47	77.95±4.37	75.95±9.96	76.44±1.76	76.03±7.67	74.10±9.50	72.68±2.11	74.41±3.24
MISSING-MODALITY METHODS								
GIML	**87.25**±1.72	**87.30**±1.95	**85.92**±1.52	83.29±1.57	**87.75**±2.62	78.32±3.44	77.90±2.27	73.19±1.68
MARS	66.63±13.51	64.43±11.58	66.65±14.60	64.16±11.23	67.06±14.81	64.95±14.97	60.75±9.31	64.00±9.92
LIMSSR	74.95±8.24	67.89±15.02	70.02±14.35	66.87±15.73	70.94±13.87	69.02±14.54	64.76±13.56	63.76±12.61
SimMLM	75.29±2.90	75.84±3.21	82.14±3.61	74.73±1.50	75.43±2.80	82.13±3.53	74.59±1.34	75.71±2.91
ShaSpec	76.35±1.38	75.08±1.92	74.68±1.26	75.79±1.44	75.38±1.16	75.37±3.85	73.00±1.17	73.84±0.37
Flex-MoE	73.57±5.45	71.17±4.33	71.08±8.03	70.65±8.66	70.79±6.70	67.00±9.11	71.02±2.62	67.73±6.73
OMNI MODELS								
Q2.5-O-3B	85.32	80.16	73.25	76.11	57.38	63.17	60.48	66.11
Q2.5-O-7B	77.70	68.97	71.19	69.76	74.68	62.30	58.02	60.24
InterO-8B	64.37	55.95	57.94	64.92	63.49	57.54	55.48	59.60
OmniV-9B	76.35	77.06	81.90	72.14	83.25	79.13	76.27	77.06
Gemma4-12B	71.35	44.37	45.16	50.16	65.63	50.56	52.78	49.84
StreamO-8B	54.60	57.94	58.10	59.76	49.29	56.43	56.43	51.43
LLUNA	84.38±1.08	84.08±0.61	82.56±2.59	**83.40**±2.11	84.65±1.39	**83.19**±1.00	**81.16**±2.23	**80.92**±1.39

Table 10 (continued): **CirCor: complete results.** (Oliveira et al., 2021) Balanced accuracy (%) ↑. Panel 2/2. Trained models: five-run mean ± sample SD; fixed models: single score. AV/PV/TV/MV denote auscultation sites.

Method	PV+TV	PV+MV	TV+MV	AV	PV	TV	MV	MM
GENERAL COLLABORATION								
Mean ens.	50.00±0.00	50.00±0.00	50.00±0.00	50.00±0.00	50.00±0.00	50.00±0.00	50.00±0.00	50.00±0.00
Logit stack	51.11±1.52	50.56±1.24	51.67±3.73	50.00±0.00	50.00±0.00	56.16±9.72	50.00±0.00	51.95±2.20
Concat MLP	83.60±1.48	76.57±0.97	70.98±3.82	64.24±6.22	75.62±7.29	75.29±2.67	61.00±6.38	74.73±1.89
Attn. pool	80.90±3.85	70.94±11.55	78.21±2.79	60.49±5.80	72.11±11.47	77.27±2.08	56.79±5.29	72.61±4.76
Soft MoE	76.56±2.94	69.25±11.19	74.60±1.94	59.56±5.93	71.52±11.89	74.17±2.03	59.65±6.25	70.28±3.54
FoE	50.00±0.00	50.00±0.00	50.00±0.00	50.00±0.00	50.00±0.00	50.00±0.00	50.00±0.00	49.92±0.19
Set Trans.	77.32±5.84	68.40±11.63	75.92±4.34	60.87±6.84	66.49±2.74	74.67±0.90	61.56±10.77	71.32±3.00
CLARGA	76.21±5.78	72.57±13.23	74.87±1.96	64.10±8.26	76.97±15.19	73.78±0.80	64.75±8.69	73.63±4.83
MISSING-MODALITY METHODS								
GIML	**84.29**±3.17	80.97±4.04	79.86±2.08	53.89±3.17	60.56±5.69	58.89±2.32	55.56±3.93	74.83±1.26
MARS	62.81±13.66	66.41±13.77	67.02±10.52	55.30±6.95	63.57±17.68	60.68±12.55	60.86±9.25	63.48±10.72
LIMSSR	65.71±14.68	68.76±17.50	65.60±14.59	61.19±10.45	68.17±16.60	63.48±12.32	57.56±8.12	65.98±13.55
SimMLM	76.54±4.04	82.02±3.21	74.59±1.43	68.44±2.23	80.71±4.36	74.59±1.43	74.92±2.88	76.60±2.30
ShaSpec	75.51±1.35	76.51±3.04	75.94±1.17	69.92±1.79	77.02±2.01	74.81±1.71	71.37±3.81	74.59±1.16
Flex-MoE	73.73±2.88	72.75±7.49	71.48±7.77	58.75±8.56	62.62±12.76	64.60±9.98	63.38±9.81	68.34±4.85
OMNI MODELS								
Q2.5-O-3B	58.57	53.02	66.75	51.59	50.71	50.87	52.86	61.50
Q2.5-O-7B	50.95	58.10	63.02	50.00	50.00	50.00	50.00	59.80
InterO-8B	61.03	67.94	64.37	51.19	54.05	54.76	55.40	58.83
OmniV-9B	76.11	75.56	71.83	70.00	81.59	62.54	67.94	75.17
Gemma4-12B	57.06	54.21	62.78	45.87	45.87	54.29	52.14	52.19
StreamO-8B	63.41	58.97	55.71	44.68	50.95	53.65	59.37	55.44
LLUNA	83.22±0.98	**82.25**±0.64	**83.54**±1.86	**72.10**±2.76	**84.24**±3.58	**81.56**±2.60	**77.73**±2.03	**81.76**±0.55

Table 11: **ADNI: complete results.** (Mueller et al., 2005) Balanced accuracy (%) ↑. Trained models: five-run mean ± sample SD; fixed models: single score. F denotes full input; MM averages non-full states.

Method	MPC^{F}	MP	MC	PC	M	P	C	MM
GENERAL COLLABORATION								
Mean ens.	73.25±8.18	64.00±11.71	64.50±9.42	76.75±6.99	48.75±12.02	70.25±6.40	77.00±4.56	66.88±6.12
Logit stack	78.00±3.60	74.25±6.71	76.25±4.59	79.75±3.79	53.75±7.40	74.25±9.54	73.25±2.09	71.92±3.20
Concat MLP	86.25±3.85	84.75±4.28	81.25±6.67	83.00±2.44	68.50±4.45	81.25±3.95	79.50±4.56	79.71±3.03
Attn. pool	89.75±1.05	83.25±2.09	77.25±6.70	88.00±4.29	59.25±7.84	82.75±5.76	76.00±3.35	77.75±3.90
Soft MoE	86.25±5.80	80.25±3.47	75.50±4.89	82.00±6.03	53.75±6.90	77.25±8.54	72.50±6.90	73.54±2.07
FoE	76.25±12.34	72.00±12.70	64.25±7.84	75.75±9.30	50.75±7.48	65.00±12.12	62.00±11.91	64.96±5.84
Set Trans.	81.00±4.45	74.75±12.23	75.75±5.20	80.50±5.77	53.75±13.55	77.25±6.34	77.50±5.30	73.25±5.85
CLARGA	83.75±5.08	78.75±4.15	77.25±4.63	87.50±2.93	61.00±7.42	82.00±6.16	75.50±3.71	77.00±3.38
MISSING-MODALITY METHODS								
GIML	**90.50**±3.81	73.25±5.49	78.75±7.02	**88.25**±3.81	52.25±3.24	68.00±10.88	76.75±3.38	72.88±2.67
MARS	88.00±1.90	79.50±3.91	**82.75**±2.71	87.25±1.37	51.75±6.99	78.50±4.71	**80.00**±5.80	76.62±1.58
LIMSSR	82.75±6.70	77.25±7.83	77.25±3.69	85.25±4.95	53.50±6.52	77.25±6.52	71.00±2.85	73.58±3.09
SimMLM	87.50±5.38	84.00±3.24	79.25±6.77	87.25±6.34	64.50±7.94	81.50±3.69	77.50±4.42	79.00±2.93
ShaSpec	87.75±2.40	84.50±3.14	76.25±2.93	87.25±4.18	60.50±10.81	84.50±2.27	75.50±2.74	78.08±3.12
Flex-MoE	80.50±5.97	74.00±3.79	77.50±4.15	82.25±5.18	56.50±7.15	78.50±11.23	74.50±3.38	73.88±2.36
OMNI MODELS								
Q2.5-O-3B	65.00	50.00	63.75	55.00	48.75	50.00	50.00	52.92
Q2.5-O-7B	76.25	50.00	71.25	51.25	50.00	50.00	50.00	53.75
InterO-8B	70.00	55.00	40.00	62.50	48.75	65.00	50.00	53.54
OmniV-9B	82.50	50.00	53.75	50.00	50.00	46.25	50.00	50.00
Gemma4-12B	90.00	50.00	53.75	57.50	42.50	51.25	50.00	50.83
StreamO-8B	81.25	50.00	51.25	57.50	47.50	50.00	50.00	51.04
LLUNA	88.00±4.11	**89.75**±2.40	78.75±3.75	82.25±3.24	**76.00**±4.09	**87.25**±1.63	79.25±2.59	**82.21**±1.26

B MAIN EXPERIMENTAL SETUP

B.1 TASKS, SPLITS, AND AVAILABILITY STATES

SIMS and MOSI are sentiment-regression tasks; MELD and URF are emotion and humor classification tasks. Daily and WS use supervised, video-disjoint QA splits derived from the public benchmarks. CirCor predicts murmur presence from four auscultation sites, and ADNI predicts MCI-to-dementia conversion using MRI, PET, and clinical inputs. The medical tasks use patient- or subject-disjoint splits. The task-specific paragraphs below specify outcome windows, eligibility rules, and preprocessing.

Table 12 specifies the evaluation population, split unit, and task metric. The same test examples are evaluated under every legal nonempty modality subset. Full inputs are reported separately; Mean-Missing equally weights the remaining states within each run before the five-run aggregation in Appendix A.1. Accuracy is abbreviated Acc.; balanced accuracy (BAcc.) is the mean of the two class recalls. MELD uses support-weighted F1 over its seven emotion classes.

Affective tasks. SIMS and MOSI predict continuous utterance-level sentiment in $[-1, 1]$ and $[-3, 3]$, respectively. MELD predicts anger, disgust, fear, joy, neutral, sadness, or surprise; URF predicts funny versus not funny. URF v2 refers to the fixed preprocessing protocol used here.

Table 12: **Dataset protocols.** Entries count utterances (SIMS/MOSI/MELD), humor examples (URF), question–answer items (Daily/WS), visits (CirCor), or labeled subjects (ADNI). Splits are train / validation / test; state counts include Full. Input abbreviations follow Appendix A.1.

Dataset	Inputs	Train / val. / test	States	Split unit	Task / metric
SIMS	T/A/V	2,722 / 647 / 1,034	7	Source video	Sentiment / MAE
MOSI	T/A/V	1,284 / 229 / 686	7	Source video	Sentiment / MAE
MELD	T/A/V	9,988 / 1,108 / 2,608	7	Dialogue	Emotion / W-F1
URF	T/A/V	7,614 / 980 / 994	7	Source video	Humor / Acc.
Daily	A/V	836 / 183 / 177	3	Video	Four-choice QA / Acc.
WS	V/A/S	2,214 / 480 / 478	7	Video	3/4-choice QA / Acc.
CirCor	AV/PV/TV/MV	409 / 88 / 88	15	Patient group	Murmur / BAcc.
ADNI	M/P/C	202 / 32 / 50	7	Subject	MCI conversion / BAcc.

Each task varies text, audio, and vision, giving one Full state, three two-modality states, and three one-modality states.

Derived audio–visual question answering. Daily and WS use supervised, video-disjoint derived splits of their public QA resources. Daily contains 479 / 103 / 102 videos; WS contains 1,164 / 249 / 249. The question and candidate answers remain visible as task context in every state. Daily has four choices and two evidence modalities, audio and video. Its locally generated ASR text depends on audio and disappears with it. WS has three or four choices and varies video, raw audio, and official subtitles independently; padded option positions are masked before prediction. Candidate-order prefixes are removed before encoding. Generated captions and answer-bearing annotations are not evidence inputs.

CirCor. The four-site derived task retains 585 visits from 556 patient groups with all four auscultation sites and a Present/Absent murmur label; Unknown labels are excluded. Patient groups connect records sharing Patient ID or Additional ID. The train / validation / test splits contain 387 / 84 / 85 groups with zero overlap, and 84 / 18 / 18 positive visits. AV, PV, TV, and MV denote the aortic, pulmonary, tricuspid, and mitral sites. Removing a site removes its waveform and every representation derived from that waveform. Age, sex, height, weight, and pregnancy status remain available as context; diagnostic annotations, identifiers, murmur descriptors, and segmentation labels are excluded from model inputs. Mean-Missing averages all 14 incomplete site combinations.

ADNI. The task predicts MCI-to-dementia conversion with subsequent diagnostic confirmation. The frozen subject partition contains 384 / 72 / 101 subjects before outcome eligibility filtering; the labeled subset in Table 12 contains 41 / 10 / 10 converters. MRI, amyloid PET, ADAS, FAQ, and CDR measurements are matched within a 90-day window. The prediction index is the latest date among these selected measurements, so the outcome clock starts after all input measurements. A converter has a first dementia diagnosis after the index and no later than 36 months plus 90 days (approximately 39 months), followed by a confirming dementia diagnosis on a distinct later date. This cutoff applies to the first diagnosis, not to the confirmation visit, which may occur beyond that window. Stable MCI requires no dementia through the first-diagnosis window and an explicit MCI assessment at or after 36 months minus 90 days. Feature filtering, imputation, standardization, and PCA are fitted only on the 384 training subjects; task heads and compatibility estimates use the training subset eligible under this outcome definition. MRI, PET, and the clinical instruments form three independently removable modalities, including clinical-absent states.

B.2 Frozen inputs and baseline interfaces

Main model pools. Each task uses a pool of nine frozen model representations or task-specific feature sets. SIMS, MOSI, MELD, URF, WS, Daily, and CirCor share the nine encoders in Table 13(a), with task-specific inputs summarized in Table 14. ADNI uses the MRI and PET embeddings, imaging measurements, and clinical assessment features listed in Table 13(b). Each pool entry has its own adapter.

Table 13: **Frozen model pools and parameter counts.** (a) Shared encoders for SIMS, MOSI, MELD, URF, WS, Daily, and CirCor. Counts are approximate billions of frozen parameters in the components used for feature extraction under complete inputs, excluding trainable collaboration modules. (b) ADNI inputs. Entry indices are preserved for both pools; Table 14 specifies task-specific input mappings.

(a) Shared text/audio/vision model pool

Entry	Modality	Model / component used	Params (B)
1	Text	Qwen3-1.7B	1.721
2		SmolLM2-1.7B-Instruct	1.711
3		XLM-RoBERTa-large	0.560
4	Audio	Whisper-large-v3 (encoder)	0.637
5		WavLM-base-plus	0.094
6		HuBERT-large-ls960-ft	0.315
7	Vision	DINOv2-giant	1.136
8		SigLIP2-giant-opt-patch16-256 (vision tower)	1.163
9		VideoMAE-base-finetuned-kinetics	0.087
Total (9 models)			≈ 7.43

(b) ADNI model representations and measurements

Entry	MRI	Entry	PET	Entry	Clinical
1	DINOv2-giant	4	DINOv2-giant	7	ADAS-Cog features
2	SigLIP2-giant	5	SigLIP2-giant	8	FAQ features
3	FS morphometry	6	SUVR summaries	9	CDR features

Table 14: **Input mapping across all eight datasets.** Entry numbers refer to Table 13. Availability follows the originating input, with task-specific masking described below.

Datasets	Inputs to 1–3	Inputs to 4–6	Inputs to 7–9
SIMS, MOSI, MELD, URF	Utterance text	Speech audio	Video frames
WS	Subtitles	Audio	Video frames
Daily	ASR transcripts	Audio	Video frames
CirCor	Demographic text	PCG waveforms	PCG log-spectrograms
ADNI	MRI images and morphometry	PET images and SUVR measurements	Clinical assessments

Input availability. SIMS, MOSI, MELD, URF, and WS independently mask their text/subtitle, audio, and visual inputs. Daily's ASR transcripts depend on audio, so its three text and three acoustic streams disappear together when audio is removed; the other three streams depend on video. ADNI groups its streams into MRI, PET, and clinical inputs.

CirCor varies the four auscultation sites while retaining the permitted demographic template. The audio encoders process PCG waveforms and the visual encoders process deterministic PCG log-spectrograms. Each signal stream preserves four site slots; all six signal encoders lose the same slot when that site is removed. Each signal adapter projects each site separately, applies the site mask after projection and site conditioning, and concatenates the four ordered slots with the four-bit mask. This yields one node per encoder. A signal node stays available while any site survives, and all three demographic nodes remain available; every legal CirCor state therefore retains nine graph nodes while their signal contents change.

ADNI feature definitions. FS denotes FreeSurfer morphometry, and SUVR denotes standardized uptake value ratio. ADAS, FAQ, and CDR denote ADAS-Cog item/total scores, Functional Activities Questionnaire measures, and Clinical Dementia Rating features.

Graph node indexing. For the three case studies in Figure 5, entries 1–9 of Table 13 give the node order: MELD and WS use panel (a), and ADNI uses panel (b). The graph case studies visualize model relations under each task's graph-weighting and pruning settings. Graph weights are fitted on training-partition predictions.

Feature preparation and adapted methods. For the affective tasks, text is encoded in nonoverlapping chunks of up to 256 tokens and pooled with token-count weights; 16-kHz audio is encoded in 20-second chunks and pooled with duration weights. Visual encoders use 16 uniformly sampled frames across the clip. The QA pipeline retains eight ordered temporal bins and uses question-and-option-conditioned scoring of the frozen evidence. Missing inputs zero the corresponding feature slots and are accompanied by an explicit availability mask. Same-pool baselines share these cached inputs and boundaries of independently maskable inputs. Their task outputs are adapted to regression, classification, or masked candidate scoring. Each baseline retains its defining collaboration or missing-modality mechanism within this common frozen-feature interface. LIMSSR uses Qwen2.5-0.5B with rank-4 LoRA and task adapters. Frozen unimodal encoders are excluded from trainable-parameter counts.

Baseline training and checkpoint selection. Each baseline retains its training procedure after adaptation to the shared frozen-feature interface. Table 15 summarizes the availability states used by the general collaboration and staged missing-modality baselines. For each trainable baseline, we select one checkpoint by validation Mean-Missing and evaluate it across all test states. Mean ens. averages the outputs of fixed unimodal task heads and has no trainable collaboration layer.

Table 15: **Baseline training views.**

Methods	Training availability states
Logit stack, Concat MLP, Attn. pool, Soft MoE, FoE, Set Trans.	Complete inputs.
CLARGA, LIMSSR, ShaSpec	One uniformly sampled legal nonempty state per batch.
SimMLM	Five epochs of complete-input expert training, followed by sampled pairs in which one state's modalities strictly contain the other's.
Flex-MoE	Five complete-input generalization epochs, followed by specialization on uniformly sampled incomplete states.

GIML retains feature-degradation training, whereas MARS uses adaptive partial-view sampling.

Omni models. Each Omni comparator receives the currently available raw evidence through its own input processor, independently of the nine-model feature cache. Unavailable modalities are physically removed before processing. The backbone remains frozen; a final prompt-token representation is standardized and fitted with one task head using complete-input *training examples only*. Classification uses balanced logistic regression with inverse regularization strength 1; regression uses ridge regression with regularization strength 10. The same head evaluates every test state, with no missing-input head training, state-specific heads, or test-time adaptation. Questions and choices remain in QA prompts, with invalid candidate positions masked. CirCor retains permitted demographic context; ADNI inputs use the available image montages and clinical fields. These are frozen-backbone task-head evaluations, not zero-shot answer generation. Repeated evaluation is deterministic and therefore is reported without a training SD.

Graph case-study details. To examine how model interactions are organized across modalities, we visualize the fixed model graphs of MELD, WS, and ADNI (Figure 5). All three graphs use identical node coordinates. MELD and WS also share model identities at each position, with text interpreted as subtitles in WS; ADNI uses the same three-group layout for MRI, PET, and clinical models. Edge weights encode dataset-level prediction compatibility estimated on the training

partition. Positive undirected edges share one width scale; self-loops are excluded and weights precede incoming normalization, without additional thresholding. The matrices average between-group weights including zeros, using shared logarithmic colors with zero shown in white. MELD annotations are scaled by 10^3.

MELD has nine within-modality links and two weak cross-modality bridges: Qwen3–Whisper and WavLM–VideoMAE. These bridges connect the three modality groups into one component. WS has nineteen links, twelve of which cross modalities, with stronger mean audio–visual coupling (0.77) than subtitle–audio (0.21) or subtitle–visual (0.15) coupling. ADNI retains all thirty-six possible links, with cross-modality means of 0.76–0.80. These patterns show dataset-dependent interaction structures, ranging from sparse modality bridges to dense cross-modality connections. The private path remains available to every active model irrespective of these connections. Appendix D.2 analyzes modality contributions and utility interactions of the resulting predictor, complementing the prediction compatibility used here to construct graph weights.

B.3 Main-run optimization settings

Table 16 summarizes the main training settings. Within each task, the five runs share all settings except the random seeds for initialization, training, and chain sampling. AdamW updates the adapters and collaboration layers while keeping the unimodal encoders fixed. Training uses task-specific availability-state weights, whereas evaluation averages all non-full states equally. Task-specific configurations specify branch-wise learning-rate and regularization multipliers, state weights, and loss schedules. For each run, one checkpoint is selected by validation Mean-Missing and used across all availability states.

Table 16: **Lluna main-run settings.** Private/common widths are $d_{\text{priv}}/d_{\text{com}}$; L is graph depth and K the chain budget per step. Learning rates and loss coefficients are rounded to three significant digits. Patience is in epochs.

(a) Architecture and training objectives

Dataset	d_{priv}	d_{com}	L	λ_{com}	λ_{PCC}	K
SIMS	64	192	2	0.300	1.58×10^{-2}	2
MOSI	144	576	2	0.278	9.60×10^{-2}	2
MELD	128	384	2	0.0860	6.25×10^{-4}	1
URF	80	160	1	0.0125	1.00×10^{-2}	1
Daily	128	256	1	0.0500	1.81×10^{-3}	1
WS	128	768	1	0.0500	1.00×10^{-3}	2
CirCor	320	640	1	0.200	1.00×10^{-1}	2
ADNI	96	384	1	0.0625	9.23×10^{-3}	2

(b) Optimization

Dataset	Batch	Learning rate	Weight decay	Max epochs	Patience
SIMS	128	3.26×10^{-4}	10^{-4}	105	25
MOSI	192	3.22×10^{-4}	10^{-3}	60	15
MELD	200	8.88×10^{-5}	2.5×10^{-5}	45	10
URF	384	7.50×10^{-5}	10^{-3}	180	25
Daily	64	4.00×10^{-4}	10^{-3}	150	30
WS	64	2.03×10^{-4}	0.05	240	64
CirCor	6	7.50×10^{-5}	10^{-3}	24	8
ADNI	32	1.05×10^{-3}	5×10^{-4}	420	80

B.4 Missing-input training controls

This experiment evaluates whether the comparison changes when the baselines receive missing-input views during training. We use SIMS, WS, and ADNI to cover the three domains, with the same splits

and test sets as the main evaluation. Complete training examples are augmented with complete and nonempty missing-input views.

Collaboration controls. Concat MLP, Attn. pool, Soft MoE, FoE, Set Transformer, CLARGA, and I^{2}MoE use the same frozen model pool and modality-level masking. The first six controls use the corresponding LLUNA configuration's task-view weights and batch construction; FoE, Set Transformer, and CLARGA also use its sampled-chain view schedule. FoE freezes its unimodal task heads before fitting the fuser, while CLARGA retains its contrastive objective. The I^{2}MoE adaptation receives missing-input views through its own training procedure. Each control retains its defining fusion mechanism and uses one validation-MM-selected checkpoint across all test views.

Omni controls. Qwen2.5-Omni 3B/7B and Stream-Omni use their frozen backbones. For each of the seven nonempty views, we remove the unavailable raw modalities before feature extraction. One shared standardizer and one shared linear task head are fitted on the training representations from all seven views and used unchanged across test states. Regression uses ridge regression with regularization strength 10; classification uses balanced logistic regression with inverse regularization strength 1. View weights follow the task-view objective and sum to one per original training example. In the column order of Tables 17–19, the normalized weights are $(0.042553, 0.191489, 0.191489, 0.191489, 0.127660, 0.127660, 0.127660)$ for SIMS, $(0.058824, 0.078431, 0.078431, 0.078431, 0.235294, 0.235294, 0.235294)$ for WS, and $(0.102564, 0.076923, 0.205128, 0.205128, 0.153846, 0.102564, 0.153846)$ for ADNI. This matches the support and relative weighting of missing-view supervision.

Tables 17–19 report every test state for the ten controls. MM averages the six non-full states equally, irrespective of their training weights.

Table 17: **Missing-input training on CH-SIMS v2.** Results for every test view; column labels specify the retained modalities. MAE is multiplied by 100. MM averages the six non-full views. T/A/V denote text, audio, and vision.

Method	TAV	TA	TV	AV	T	A	V	MM
Concat MLP	32.68	34.46	35.68	36.12	39.78	37.63	43.74	37.90
Attn. pool	29.24	31.47	33.30	31.72	37.13	35.03	44.10	35.46
Soft MoE	35.53	33.87	38.16	37.70	38.72	36.33	43.72	38.08
FoE	44.93	44.84	47.62	45.90	49.06	47.65	50.48	47.59
Set Trans.	30.89	33.00	36.45	33.16	40.17	37.04	44.66	37.42
CLARGA	34.20	36.91	38.61	36.90	41.74	40.88	46.06	40.18
I^{2}MoE	30.71	33.87	36.06	32.31	40.52	37.56	43.51	37.31
Qwen2.5-Omni 3B	48.69	53.76	50.51	46.14	57.34	53.04	48.51	51.55
Qwen2.5-Omni 7B	51.18	48.11	49.39	45.97	58.04	44.12	46.03	48.61
Stream-Omni	59.27	61.41	61.34	65.72	63.95	80.02	61.83	65.71

Table 18: **Missing-input training on WorldSense.** Results for every test view; column labels specify the retained modalities. Accuracy is reported as a percentage. MM averages the six non-full views. V/A/S denote vision, audio, and subtitles.

Method	VAS	VA	VS	AS	V	A	S	MM
Concat MLP	32.64	34.31	31.59	31.59	35.15	31.59	31.59	32.64
Attn. pool	34.10	31.80	30.13	30.13	32.01	32.22	36.61	32.15
Soft MoE	29.92	33.47	29.92	29.50	33.26	31.38	29.50	31.17
FoE	32.64	28.87	33.68	32.85	28.87	28.87	33.05	31.03
Set Trans.	32.01	29.71	33.05	32.01	30.96	29.29	32.01	31.17
CLARGA	33.68	28.87	34.73	32.85	29.29	28.03	33.26	31.17
I^2MoE	32.22	28.66	32.43	32.22	26.36	29.92	32.01	30.27
Qwen2.5-Omni 3B	29.29	38.08	28.24	28.45	32.01	31.59	24.69	30.51
Qwen2.5-Omni 7B	34.94	32.85	31.80	32.22	29.71	30.13	26.57	30.54
Stream-Omni	30.96	30.54	29.71	28.66	31.17	29.08	30.34	29.92

Table 19: **Missing-input training on ADNI.** Results for every test view; column labels specify the retained modalities. Balanced accuracy is reported as a percentage. MM averages the six non-full views. M/P/C denote MRI, PET, and clinical information.

Method	MPC	MP	MC	PC	M	P	C	MM
Concat MLP	87.50	76.25	67.50	81.25	46.25	76.25	73.75	70.21
Attn. pool	87.50	91.25	68.75	76.25	80.00	82.50	70.00	78.13
Soft MoE	75.00	83.75	80.00	80.00	81.25	80.00	80.00	80.83
FoE	76.25	77.50	62.50	57.50	70.00	52.50	60.00	63.33
Set Trans.	95.00	87.50	82.50	88.75	73.75	81.25	78.75	82.08
CLARGA	88.75	78.75	80.00	86.25	77.50	82.50	78.75	80.63
I^2MoE	73.75	86.25	75.00	83.75	58.75	86.25	76.25	77.71
Qwen2.5-Omni 3B	67.50	60.00	71.25	85.00	38.75	61.25	85.00	66.88
Qwen2.5-Omni 7B	81.25	56.25	85.00	83.75	50.00	56.25	87.50	69.79
Stream-Omni	81.25	75.00	62.50	76.25	67.50	86.25	77.50	74.17

C Additional Ablation Results

Table 3 compares four ablation groups. The tables below detail these variants and give every sampled-pair and L9 setting in Section 4.5. The final Δ_{rel} row in Table 3 averages each variant's percentage change from LLUNA equally across the eight tasks, reversing the sign for MAE so that negative values always indicate degradation. Bold and underlining identify the best and second-best values in each dataset row.

Prediction-path controls. The w/o Common variant predicts using the private head alone and disables the common auxiliary loss and PCC. The w/o Private variant predicts using the unscaled common head alone. The remaining training settings are retained within each task.

Common/private combination. We retain both branches and compare the learned residual scale in Eq. (1) with fixed coefficients $\alpha \in \{0.25, 0.5, 1, 2\}$ multiplying the common prediction. Equal addition is the case $\alpha = 1$. Concat instead feeds the concatenated private features, post-interaction common consensus, and availability indicators to an MLP; its hidden width and dropout follow the private head. All variants retain PCC and common auxiliary supervision. Table 20 reports the complete sweep. Concatenation changes the prediction head as well as its parameter count. The learned residual rule improves over equal addition on six of eight tasks, while the preferred fixed coefficient varies across tasks.

Table 20: **Common/private combination variants.** MM for feature concatenation (Concat) and fixed residual scales. Equal addition corresponds to $\alpha = 1$. LLUNA uses the learned residual scale. Metrics follow Table 1; bold identifies the best displayed value in each row.

Dataset	LLUNA	Concat	$\alpha = 0.25$	$\alpha = 0.5$	$\alpha = 1$	$\alpha = 2$
SIMS↓	34.21	34.79	34.52	34.37	**33.88**	34.69
MOSI↓	**98.34**	103.19	99.55	105.66	100.21	99.36
MELD↑	56.08	55.59	55.65	55.91	**56.16**	56.05
URF↑	**70.61**	69.06	70.47	70.09	70.12	69.90
Daily↑	**38.42**	31.92	29.10	29.66	28.81	32.20
WS↑	**33.82**	31.10	33.33	33.54	33.30	33.17
CirCor↑	**82.56**	81.58	78.91	78.27	74.89	77.69
ADNI↑	84.17	83.00	83.79	**84.21**	83.96	80.63

Graph interaction controls. The graph panel in Table 3 changes only the interaction mode. LLUNA uses the fixed compatibility graph. Uniform Graph assigns equal weights to the complete active-node graph, including self-loops, followed by incoming normalization; it therefore changes connectivity as well as weights when the compatibility graph is sparse.

Consistency-pair selection and budget. These experiments test which state relationships should be regularized and how many are needed per step. With two, three, and four independently maskable inputs, adjacent PCC has 2, 9, and 28 eligible pairs; Full-to-Minimal has 2, 3, and 4; All-Pairs has 3, 21, and 105. These are candidate-set sizes, not per-step comparison counts. All-Pairs includes nonnested pairs and uses the endpoint with more modalities as the stop-gradient target, breaking ties by a fixed view order. For two modalities, Full-to-Minimal and adjacent PCC select the same pair set. The separate PCC pair-budget experiment selects up to Q unique single-modality-removal pairs per step while retaining all task-view forwards. Daily's $Q = 4, 9$ cells reuse $Q = 2$, giving thirty distinct budget runs. CirCor has twenty-eight cover edges, so $Q = 9$ covers 32.1% per step.

PCC direction. The w/o Stop-gradient variant removes detachment from the richer state's normalized common representation. It does not add a second loss or an explicit midpoint target: the forward cosine-distance value is unchanged for fixed representations, but both endpoints receive its gradient. The task losses continue to supervise both states. This variant retains the graph, adjacent-pair construction, common-state target, and PCC coefficient and schedule. Table 3c reports its MM on all eight tasks.

Consistency target. Table 3d retains both prediction paths and varies the representation targeted by PCC. PCC-Common corresponds to LLUNA, applying consistency to the common representation. PCC-Private instead targets the hidden representation immediately before the private head's final linear layer. PCC-Both averages the cosine losses on the separately normalized common and private representations, retaining the same overall PCC coefficient. The comparison examines whether cross-state consistency is most effective when focused on the common representation.

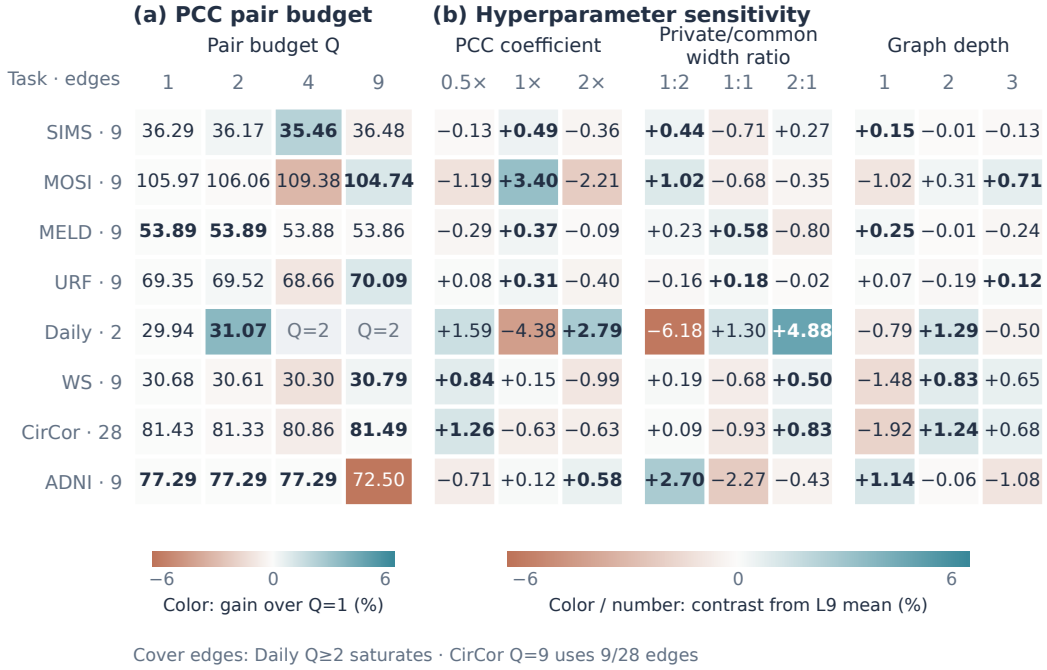

Figure 6: **PCC pair budget and hyperparameter sensitivity.** (a) Numbers: test MM (MAE $\times 100$ for SIMS/MOSI; percentages otherwise). Colors: oriented relative change from $Q = 1$. Task labels give the numbers of single-modality-removal pairs; gray Daily cells reuse $Q = 2$. (b) Numbers/colors: oriented relative contrasts from each task's nine-configuration L9 mean, averaging three distinct configurations per factor level; bold marks each task/factor's best level. Positive means better in both panels.

Table 21: **PCC pair budgets.** MM; metrics and units follow Table 1. Q counts cover pairs selected per step. Daily's $Q = 4, 9$ cells reuse $Q = 2$; CirCor samples at most 9 of 28 edges. Bold marks each task's best budget.

Setting	SIMS↓	MOSI↓	MELD↑	URF↑	Daily↑	WS↑	CirCor↑	ADNI↑
$Q = 1$	36.29	105.97	**53.89**	69.35	29.94	30.68	81.43	**77.29**
$Q = 2$	36.17	106.06	**53.89**	69.52	**31.07**	30.61	81.33	**77.29**
$Q = 4$	**35.46**	109.38	53.88	68.66	**31.07**	30.30	80.86	**77.29**
$Q = 9$	36.48	**104.74**	53.86	**70.09**	**31.07**	**30.79**	**81.49**	72.50

Hyperparameter sensitivity: L9 design. The L9 design examines sensitivity to consistency strength, private/common width ratio, and interaction depth. The PCC coefficient is $0.5\times$, $1\times$, or $2\times$ its task-specific base value, and graph depth is one, two, or three layers. For fixed private width d_{priv}, the common width d_{com} is $2d_{\text{priv}}$, d_{priv}, or $\lfloor d_{\text{priv}}/2 \rfloor$ at ratios 1:2, 1:1, and 2:1; width and depth changes also change model capacity. For each task, we average native MM across the three configurations at each factor level. Figure 6b subtracts the nine-configuration mean from each factor-level mean, divides by the absolute nine-configuration mean, and multiplies by 100. We reverse the sign for MAE so that positive values indicate improvement. Each cell averages three different

configurations, not seed replicates. L9 balances factor levels but does not separately estimate their interactions.

Table 22: **Nine L9 configurations.** MM; metrics and units follow Table 1. Weights multiply each task's base PCC coefficient. Bold marks the best configuration per task.

Setting	SIMS↓	MOSI↓	MELD↑	URF↑	Daily↑	WS↑	CirCor↑	ADNI↑
L9-1	35.70	109.18	54.03	69.42	28.81	29.92	81.18	76.46
L9-2	36.20	111.13	54.51	69.47	**33.90**	**31.21**	81.47	74.17
L9-3	36.13	112.45	53.53	69.50	33.33	**31.21**	82.49	73.75
L9-4	35.80	106.35	54.57	69.35	28.53	30.96	80.81	77.50
L9-5	35.99	**104.11**	54.31	**69.85**	29.38	30.09	80.93	71.67
L9-6	**35.57**	107.21	54.26	69.67	32.49	30.65	78.82	77.08
L9-7	35.91	109.95	54.30	69.11	31.36	30.86	80.32	**78.12**
L9-8	36.46	115.83	**54.65**	69.28	32.49	29.64	77.45	75.00
L9-9	35.90	110.34	53.44	69.00	33.33	30.16	**82.80**	74.17

L9 row: (weight multiplier, private/common width ratio, graph depth)

L9-1: (0.5×, 1:2, 1)	L9-2: (0.5×, 1:1, 2)	L9-3: (0.5×, 2:1, 3)
L9-4: (1×, 1:2, 2)	L9-5: (1×, 1:1, 3)	L9-6: (1×, 2:1, 1)
L9-7: (2×, 1:2, 3)	L9-8: (2×, 1:1, 1)	L9-9: (2×, 2:1, 2)

D Additional Analyses

D.1 Model-Pool and Parameter Analyses

Individual pool-entry performance. Table 23 reports each entry's test performance using the inputs specified in Table 14. Outside CirCor, an entry's prediction is unchanged whenever its required input remains available. Table 24 gives the CirCor results as auscultation sites are removed, with demographic inputs retained.

Table 23: **Individual pool-entry performance.** Full-input test scores; each entry uses its own features and task head. Metrics and units follow Table 1.

(a) Shared encoder pool

Model	SIMS MAE ↓	MOSI MAE ↓	MELD W-F1 ↑	URF Acc. ↑	Daily Acc. ↑	WS Acc. ↑	CirCor BAcc. ↑
Qwen3-1.7B	64.43	144.78	50.99	66.60	28.25	33.47	46.90
SmolLM2-1.7B	65.84	125.47	51.93	68.01	35.03	32.22	52.62
XLM-RoBERTa-large	51.14	128.34	54.93	60.16	32.20	28.66	60.16
Whisper-large-v3	46.91	142.29	44.99	69.32	32.20	29.92	79.05
WavLM-base-plus	50.42	133.07	45.58	63.28	26.55	28.87	80.32
HuBERT-large	54.27	142.42	43.50	60.06	34.46	27.82	81.11
DINOv2-giant	59.37	250.63	29.98	52.82	30.51	28.45	87.38
SigLIP2-giant	47.10	358.27	35.95	56.84	32.77	27.82	81.83
VideoMAE-base	54.61	227.03	33.34	53.82	31.64	29.08	82.54

(b) ADNI pool: balanced accuracy (%)

MRI	Score	PET	Score	Clinical	Score
DINOv2-giant	53.75	DINOv2-giant	70.00	ADAS	67.50
SigLIP2-giant	67.50	SigLIP2-giant	62.50	FAQ	70.00
FreeSurfer	66.25	SUVR	78.75	CDR	73.75

Table 24: **Individual CirCor entries across available-site subsets.** Balanced accuracy (%); MM averages the 14 non-full states. Site abbreviations follow Table 12.

(a) Four or three available sites

Model	Full	AV+PV +TV	AV+PV +MV	AV+TV +MV	PV+TV +MV
Qwen3-1.7B	46.90	46.90	46.90	46.90	46.90
SmolLM2-1.7B	52.62	52.62	52.62	52.62	52.62
XLM-RoBERTa-large	60.16	60.16	60.16	60.16	60.16
Whisper-large-v3	79.05	76.11	50.00	62.06	74.21
WavLM-base-plus	80.32	55.79	50.00	50.00	58.73
HuBERT-large	81.11	50.00	55.71	61.11	50.00
DINOv2-giant	87.38	70.00	60.71	57.14	55.00
SigLIP2-giant	81.83	50.71	50.00	50.00	77.78
VideoMAE-base	82.54	77.30	50.00	73.41	50.71

(b) Two available sites

Model	AV+PV	AV+TV	AV+MV	PV+TV	PV+MV	TV+MV
Qwen3-1.7B	46.90	46.90	46.90	46.90	46.90	46.90
SmolLM2-1.7B	52.62	52.62	52.62	52.62	52.62	52.62
XLM-RoBERTa-large	60.16	60.16	60.16	60.16	60.16	60.16
Whisper-large-v3	50.00	61.11	50.00	79.76	50.00	72.06
WavLM-base-plus	50.00	50.00	50.00	50.71	50.00	50.00
HuBERT-large	50.00	72.54	60.40	50.00	50.00	50.00
DINOv2-giant	50.00	50.00	50.00	50.00	50.00	50.00
SigLIP2-giant	50.00	50.00	50.00	52.14	50.00	50.00
VideoMAE-base	50.00	65.87	50.00	50.00	50.00	50.00

(c) One available site and Mean-Missing

Model	AV	PV	TV	MV	MM
Qwen3-1.7B	46.90	46.90	46.90	46.90	46.90
SmolLM2-1.7B	52.62	52.62	52.62	52.62	52.62
XLM-RoBERTa-large	60.16	60.16	60.16	60.16	60.16
Whisper-large-v3	50.00	50.00	71.98	50.00	60.52
WavLM-base-plus	50.00	50.00	50.00	50.00	51.09
HuBERT-large	50.00	50.00	50.00	50.00	53.55
DINOv2-giant	50.00	50.00	50.00	50.00	53.06
SigLIP2-giant	50.00	50.00	50.00	50.00	52.19
VideoMAE-base	50.00	50.00	50.00	50.00	54.81

Effect of model-pool size. We examine how LLUNA's predictive performance changes with the number of available models. Each dataset uses nested pools of three, six, and nine frozen models. Table 25 lists the retained models. Figure 3(b) summarizes relative MM changes across all eight tasks; Figure 7 gives the task scores at each pool size.

Table 26 reports MM and percentage changes from the three-model pool: the score difference divided by the reference score, with its sign reversed for MAE so that positive means better. Nine models give the best observed MM on all eight tasks. Performance improves at both expansions on six tasks, whereas CirCor and ADNI decline at six models before improving at nine. Daily increases from 35.59% to 37.85% and 38.42%; ADNI changes from 82.71% to 76.67% and 84.17%. Thus, the nine-model pool provides the strongest performance among the evaluated sizes, while the benefit of an intermediate expansion remains task-dependent.

Table 25: **Nested LLUNA pools.** Entries refer to Table 13: panel (a) for the first seven datasets and panel (b) for ADNI. Entries are listed in inventory order; each nine-model pool contains entries 1–9.

Dataset	3-member pool	6-member pool
SIMS	3, 5, 9	2, 3, 5, 6, 8, 9
MOSI	1, 5, 8	1, 2, 4, 5, 8, 9
MELD	2, 5, 7	1, 2, 4, 5, 7, 9
URF	2, 6, 7	2, 3, 5, 6, 7, 9
Daily	1, 6, 7	1, 2, 6, 7, 8, 9
WS	3, 6, 9	1, 3, 5, 6, 7, 9
CirCor	1, 4, 6	1, 3, 4, 5, 6, 8
ADNI	3, 4, 9	2, 3, 4, 5, 7, 9



Figure 7: **LLUNA at different pool sizes.** MM versus the number of models; point labels give scores. Higher positions indicate better performance, with inverted MAE axes. Numerical results: Table 26.

Table 26: **LLUNA at different pool sizes.** MM; units follow Table 1. Bold marks the best count within each dataset. Relative changes use the three-model pool as reference, with positive values indicating improvement.

Dataset	3 models	6 models	9 models	3→6 (%)	3→9 (%)
SIMS↓	46.87	42.16	**34.21**	+10.06	+27.02
MOSI↓	122.63	105.04	**98.34**	+14.34	+19.81
MELD↑	52.81	54.50	**56.08**	+3.21	+6.20
URF↑	65.63	67.45	**70.61**	+2.78	+7.59
Daily↑	35.59	37.85	**38.42**	+6.35	+7.95
WS↑	29.11	30.86	**33.82**	+5.99	+16.16
CirCor↑	78.78	78.40	**82.56**	-0.48	+4.80
ADNI↑	82.71	76.67	**84.17**	-7.30	+1.77

Sensitivity to model-pool composition. The scaling experiment changes pool size; here we keep nine models and vary their composition (Figure 8). Starting from the original pool, we replace one designated text model, then an audio model, then a visual model; ADNI uses clinical, MRI, and PET models in that order. Each replacement pool is retrained, its graph is refitted on training data, and its hyperparameters are retuned. The experiment characterizes performance under changes in model composition after re-optimization.

For original metric m_0 and metric m_r after r cumulative model replacements, relative change is $100(m_r - m_0)/m_0$ for higher-better tasks and its negative for MAE. Replacing one, two, and three models gives mean direction-corrected relative changes of -1.19%, -4.65%, and -4.27% from the original pool, respectively. Sensitivity varies across tasks: larger replacements produce the largest changes on SIMS and MOSI, while the effect does not grow uniformly with the number of replaced models. Pool cardinality alone therefore does not determine performance; the representations supplied by the selected models also matter. Together with the scaling experiment, these results identify model composition as a task-dependent design choice at a fixed pool size.

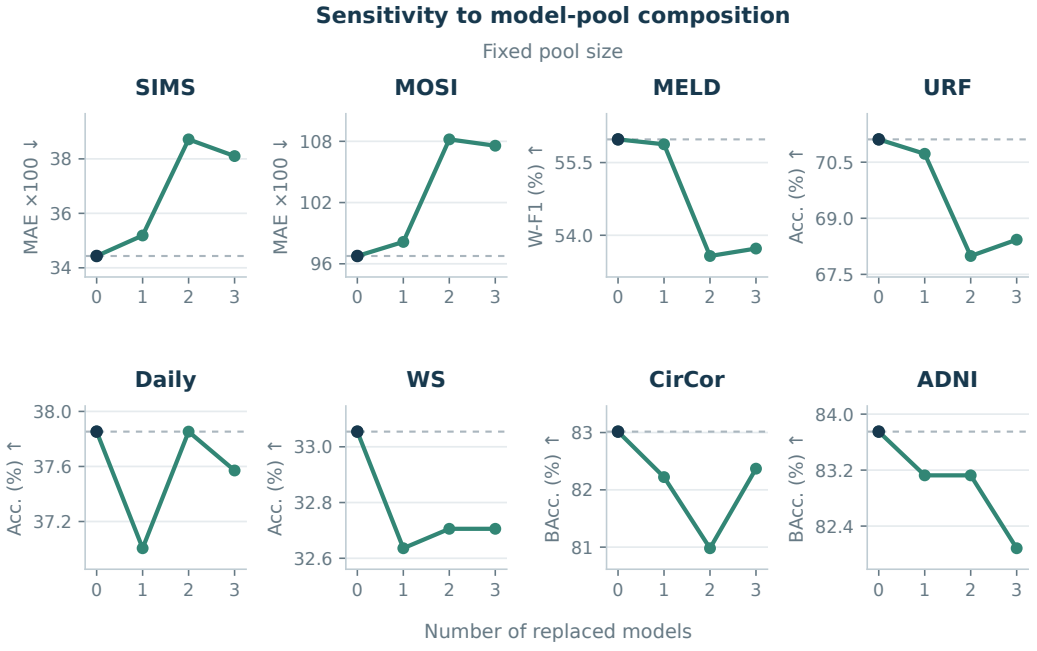

Figure 8: **Sensitivity to model-pool composition.** Zero replacements and dashed lines mark the original nine-model pool. SIMS/MOSI report MAE $\times 100$; other scores are percentages.

Table 27: **Nine-model pool replacements.** Replace 1–3 cumulatively changes model membership, followed by retraining and task-specific retuning. Scores are MM with Table 1 metrics and units. Δ_r is the direction-corrected relative MM change from the original pool after r cumulative replacements.

Dataset	Original	Replace 1	Replace 2	Replace 3	Δ_1 (%)	Δ_2 (%)	Δ_3 (%)
SIMS	34.43	35.19	38.72	38.10	-2.19	-12.44	-10.66
MOSI	96.76	98.14	108.19	107.57	-1.42	-11.81	-11.17
MELD	55.98	55.88	53.57	53.73	-0.18	-4.30	-4.02
URF	71.11	70.72	67.99	68.43	-0.54	-4.39	-3.77
Daily	37.85	37.01	37.85	37.57	-2.24	+0.00	-0.75
WS	33.05	32.64	32.71	32.71	-1.27	-1.05	-1.05
CirCor	83.02	82.22	80.98	82.37	-0.96	-2.45	-0.78
ADNI	83.75	83.12	83.12	82.08	-0.75	-0.75	-1.99

Parameter counts. The count includes parameters optimized during collaboration training and excludes frozen backbones, adjacency buffers, and separately fitted frozen task heads. Thus Mean ens. has zero optimized collaboration parameters, and LIMSSR's frozen language backbone is excluded. Figure 9 extends the UR-FUNNY view in Figure 3(c) to all eight tasks.

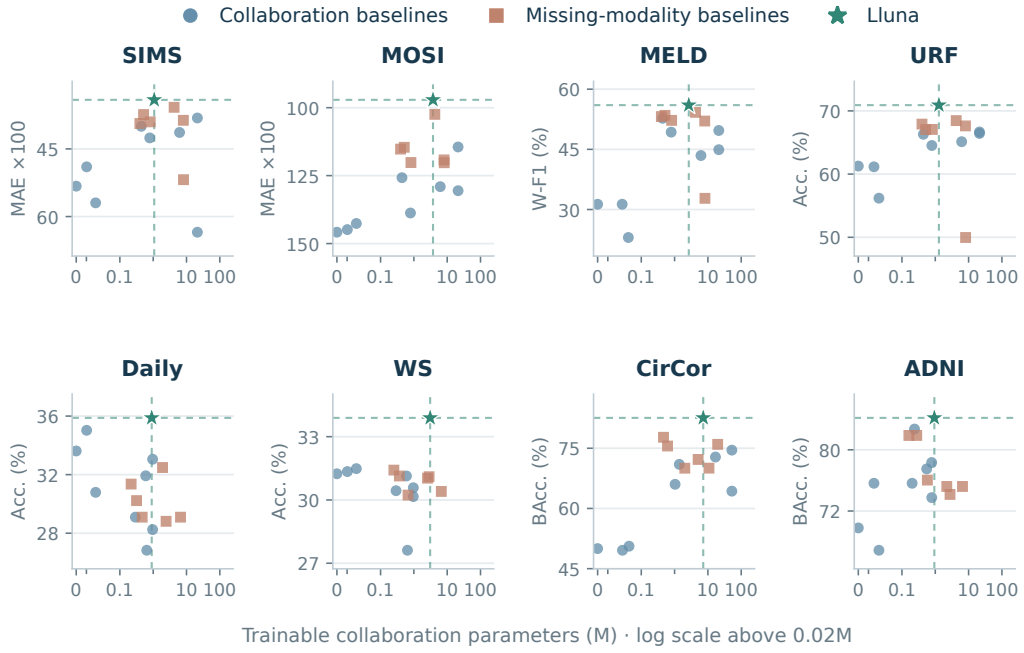

Figure 9: **Trainable parameters versus MM.** Stars: LLUNA; circles/squares: collaboration/missing-modality baselines; Omni omitted. SIMS/MOSI use MAE $\times 100$; other scores are percentages. Inverted MAE axes make higher positions better throughout. Dashed lines mark LLUNA. The parameter axis is logarithmic above 0.02M and linear near zero. Counts: Table 28.

Table 28: **Trainable collaboration parameters (millions).** Counts match Figure 9 and exclude frozen backbones, feature encoders, and separately fitted frozen task heads. Mean ens. optimizes no collaboration parameters; Omni models are omitted.

Method	SIMS	MOSI	MELD	URF	Daily	WS	CirCor	ADNI
Mean ens.	0	0	0	0	0	0	0	0
Logit stack	0.0102	0.0102	0.0410	0.0154	0.0102	0.0102	0.0430	0.0154
Concat MLP	6.1901	6.1901	6.1932	6.1906	0.5978	0.5978	16.8582	0.2380
Attn. pool	21.2199	21.2199	21.2214	21.2201	0.9608	0.9608	52.3718	0.7802
Soft MoE	21.2219	21.2219	21.2358	21.2242	0.9628	0.9628	52.3851	0.7843
FoE	0.0191	0.0191	0.0268	0.0204	0.0191	0.0191	0.0273	0.0204
Set Trans.	0.7892	0.7892	0.7900	0.7894	0.6418	0.6418	1.3861	0.5522
CLARGA	0.4391	0.4391	0.4399	0.4392	0.2916	0.2916	1.0360	0.2021
GIML	8.1434	8.1434	8.1526	8.1449	1.9118	2.5728	19.6885	2.2154
MARS	8.0437	8.0437	8.0467	8.0442	6.6179	6.6510	10.7647	6.5618
LIMSSR	4.2492	4.2492	4.2508	4.2495	2.4397	2.8584	5.0976	2.7725
SimMLM	0.3983	0.3983	0.4007	0.3987	0.2176	0.2509	0.4656	0.1615
ShaSpec	0.5146	0.5146	0.5153	0.5147	0.3172	0.3671	0.6145	0.2775
Flex-MoE	0.8157	0.8157	0.8165	0.8158	0.4673	0.6682	2.0268	0.5786
Lluna	1.0743	3.7523	2.6690	1.2852	0.9035	3.0790	7.2514	0.9374

D.2 Modality-Utility Results

We examine modality contributions and pair interactions using exact Shapley values over all modality subsets of a fixed Lluna model. Utility is negative absolute error on SIMS/MOSI and the log probability of the true class or valid answer on the other tasks. The empty-input reference predicts zero for regression and uses a uniform class/candidate distribution otherwise.

For the modality set $\mathcal{M}$ with M modalities, let $v(S)$ denote one example's scalar predictive utility under available subset S, including the empty-input reference. The Shapley contribution of modality M_i is

$$\text{Shapley}(M_i) = \sum_{S \subseteq \mathcal{M} \setminus \{M_i\}} \frac{|S|!(M - |S| - 1)!}{M!} \big[v(S \cup \{M_i\}) - v(S)\big]. \tag{7}$$

Pair interactions measure departures from additive predictive utility. For distinct modalities M_i, M_j, the interaction averages the utility second difference $v(S \cup \{M_i, M_j\}) - v(S \cup \{M_i\}) - v(S \cup \{M_j\}) + v(S)$ over $S \subseteq \mathcal{M} \setminus \{M_i, M_j\}$ with weight $|S|!(M - |S| - 2)!/(M - 1)!$, without an additional factor of $1/2$. For three modalities, this equally weights the third modality being absent and present. The reported median and positive/largest-contribution frequencies describe variation across test examples, not uncertainty across training seeds. The attribution uses a fixed model across all modality subsets; it measures that model's modality use rather than the effect of retraining without a modality.

The leading modality differs across tasks (Figure 10): audio on SIMS, text on MOSI/MELD/URF, vision on Daily, the TV auscultation site on CirCor, and clinical inputs on ADNI. On WS, audio and vision contribute similarly; subtitles have a mean contribution of -0.007 nats but positive interactions with both audio and vision (about $+0.029$ each). ADNI clinical and PET inputs both contribute positively, while their -0.174-nat interaction indicates a joint gain below the additive reference. Individual contributions and pair interactions thus reveal different aspects of evidence use. Section 4.4 examines the prediction compatibility used to construct the model graph; Tables 29 and 30 give complete modality-utility statistics.

Table 29: **Modality contributions.** Single-checkpoint Shapley utilities across test examples. Mean/median use negative-absolute-error units for SIMS/MOSI and nats otherwise; Positive/Largest are percentages of test examples.

Task	Modality	Mean	Median	Positive (%)	Largest (%)
SIMS	A	+0.152	+0.187	72.24	38.88
	T	+0.128	+0.161	67.50	32.40
	V	+0.076	+0.085	63.15	28.72
MOSI	A	+0.171	+0.168	68.37	33.97
	T	+0.276	+0.313	73.91	53.35
	V	−0.054	−0.037	40.52	12.68
MELD	A	+0.298	+0.479	75.00	25.23
	T	+0.495	+0.714	79.33	60.47
	V	+0.150	+0.264	70.97	14.30
URF	A	+0.070	+0.113	67.10	22.03
	T	+0.145	+0.335	74.55	62.37
	V	+0.008	+0.001	50.50	15.59
Daily	A	−0.335	−0.102	46.89	45.76
	V	+0.080	+0.033	59.89	54.24
WS	A	+0.013	+0.002	53.35	25.73
	S	−0.007	+0.035	53.56	53.14
	V	+0.013	+0.002	52.93	21.13
CirCor	AV	−0.222	+0.075	60.23	7.95
	MV	−0.061	+0.168	70.45	6.82
	PV	+0.149	+0.217	84.09	23.86
	TV	+0.256	+0.267	87.50	61.36
ADNI	C	+0.231	+0.291	90.00	70.00
	M	+0.094	+0.157	72.00	12.00
	P	+0.178	+0.256	84.00	18.00

Table 30: **Complete pair interactions.** Positive interactions increase joint utility above the additive reference; negative values are subadditive. Units follow Table 29; synergy is the share of examples with positive interaction.

Task	Pair	Mean	Median	Synergy (%)
SIMS	A+T	−0.141	−0.149	33.46
	A+V	−0.069	−0.078	39.26
	T+V	−0.051	−0.065	41.30
MOSI	A+T	−0.289	−0.324	26.97
	A+V	+0.022	+0.006	51.31
	T+V	+0.027	+0.021	56.27
MELD	A+T	−0.449	−0.576	22.24
	A+V	−0.219	−0.312	34.28
	T+V	−0.205	−0.339	34.47
URF	A+T	−0.064	−0.081	39.34
	A+V	+0.004	−0.030	40.14
	T+V	+0.036	−0.013	45.98
Daily	A+V	+0.160	+0.061	59.89
WS	A+S	+0.029	+0.003	52.30
	A+V	−0.015	−0.002	46.03
	S+V	+0.029	+0.003	52.30
CirCor	AV+MV	+0.156	−0.044	47.73
	AV+PV	+0.235	−0.113	44.32
	AV+TV	+0.211	−0.083	42.05
	MV+PV	−0.060	−0.210	32.95
	MV+TV	−0.115	−0.231	25.00
	PV+TV	−0.170	−0.249	20.45
ADNI	C+M	−0.002	−0.206	38.00
	C+P	−0.174	−0.369	22.00
	M+P	−0.049	−0.195	38.00

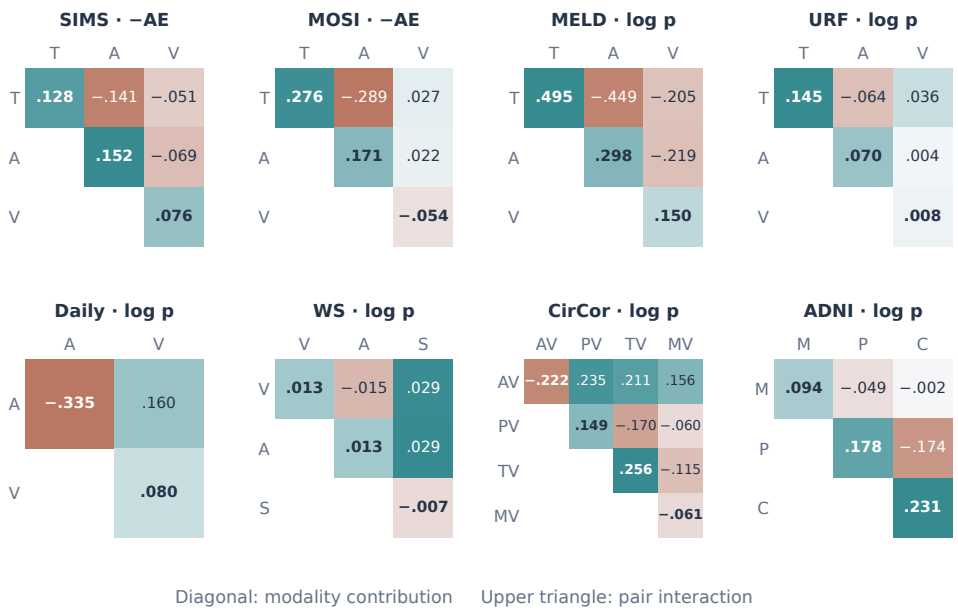

Figure 10: **All-task modality contributions and interactions.** Diagonal cells show mean Shapley values; upper-triangle cells show mean pair interactions. Teal/brown denote positive/negative values, with color ranges scaled within each task. In panel labels, AE is absolute error and p is the probability of the true class or answer. Units are native negative absolute error on SIMS/MOSI and nats elsewhere. Modality abbreviations follow Appendix A.1.

D.3 Chain-Sampling Results

To test how chain coverage affects MM, we vary the number of sampled maximal chains while keeping the remaining settings fixed within each task. Table 31 gives every budget in Figure 11. Repeated pairs within a training step are counted once, so their mean count can be noninteger. The counts measure PCC comparisons rather than wall-clock time.

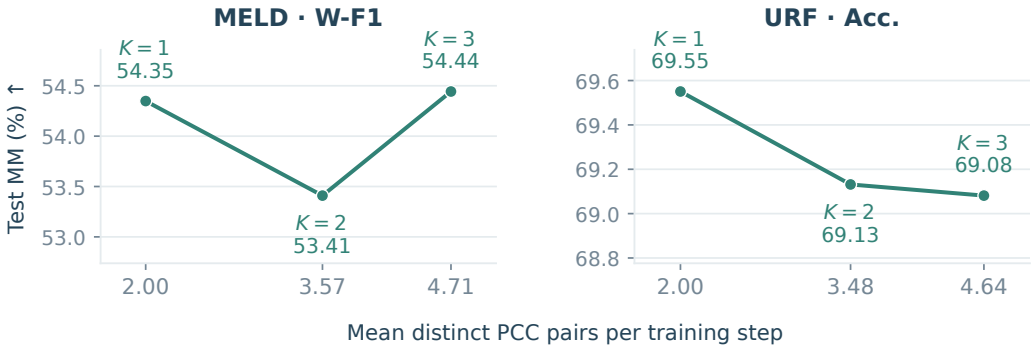

Figure 11: **Chain-sampling budget and predictive performance.** MM versus the mean number of distinct PCC pairs per step. Labels give the number of sampled chains K and the MM score. Only K changes within each task. Numeric results appear in Appendix D.3.

As K increases from one to three, MELD MM changes from 54.35% to 53.41% and 54.44%, while URF changes from 69.55% to 69.13% and 69.08%. More sampled chains therefore do not uniformly improve MM.

Table 31: **Complete chain-budget results.** MM is weighted F1 for MELD and accuracy for URF, in percent. K counts sampled maximal chains; pairs are the mean distinct PCC comparisons per training step.

Dataset	K	PCC pairs	MM (%)
MELD	1	2.00	54.35
	2	3.57	53.41
	3	4.71	54.44
URF	1	2.00	69.55
	2	3.48	69.13
	3	4.64	69.08

Task-specific sparse configurations. To assess whether sparse training retains full-lattice predictive performance across tasks, Table 32 compares task-specific sparse operating points with their full-lattice references. SIMS, MELD, URF, Daily, and CirCor use one independently sampled maximal chain per step; MOSI uses two. WS uses four cyclically scheduled chains, and ADNI uses three. The equal-task mean relative MM change is -0.28%, ranging from $+1.40\%$ on SIMS to -2.85% on CirCor. At these task-specific configurations, sparse training achieves comparable average missing-modality performance with fewer PCC comparisons per training step. The complementary fixed-setting comparison in Figure 11 varies only K and selects checkpoints by validation Mean-Missing.

Table 32: **Sparse versus full training.** Full denotes full-lattice training. Scores: MM in Table 1 units; bold marks better. Δ: direction-corrected relative MM change (positive is better). # PCC: mean distinct comparisons per step (Full $\rightarrow$ Sparse). All nonempty states are tested.

Dataset	Full	Sparse	Δ (%)	# PCC
SIMS	35.87	**35.37**	$+1.40$	$9 \rightarrow 2$
MOSI	100.99	**100.08**	$+0.91$	$9 \rightarrow 3.48$
MELD	55.28	**55.55**	$+0.49$	$9 \rightarrow 2$
URF	**70.82**	70.09	-1.04	$9 \rightarrow 2$
Daily	**36.44**	36.16	-0.78	$2 \rightarrow 1$
WS	**33.05**	33.02	-0.11	$9 \rightarrow 7$
CirCor	**82.32**	79.97	-2.85	$28 \rightarrow 3$
ADNI	**83.75**	83.54	-0.25	$9 \rightarrow 5$

PCC comparison counts. The last column of Table 32 gives the mean number of distinct PCC comparisons per training step as full-lattice $\rightarrow$ sparse. Pairs shared by sampled chains are counted once; averaging over training steps can yield noninteger counts. Full-lattice training evaluates $M(2^{M-1}-1)$ cover pairs per step for M maskable modalities. For example, sparse training reduces comparisons from nine to two on MELD and from twenty-eight to three on CirCor. These counts describe training-time PCC computation; test evaluation still includes every nonempty availability state.

E Motivation Diagnostics

This appendix gives the protocols and complete results underlying Figure 1 and the fixed-anchor analysis in Section 4.4: collaboration gains under modality removal, recoverability of modality-specific features, and changes in collaborative representations.

E.1 Fixed-anchor collaboration gains

Let $\mathcal{M}$ denote the independently maskable inputs and $S \subseteq \mathcal{M}$ the available subset. For each modality $M_i \in S$, we select a unimodal anchor f_i^{uni} by validation performance and keep its predictions fixed

across all views retaining that modality. Each collaborative predictor f_S also uses one checkpoint across views. Collaboration gain is

$$\Delta_{\text{collab}}(S; M_i) = \text{Acc}(f_S) - \text{Acc}(f_i^{\text{uni}}). \tag{8}$$

A negative gain means that collaboration is less accurate than the fixed surviving anchor. *Harmed* is the share of all test examples on which the anchor is correct and collaboration is wrong; *Rescued* is the reverse. Their difference, Rescued minus Harmed, equals Δ_{collab} in percentage points.

We average legal anchor/view comparisons within each removal level and dataset. The conventional result additionally averages Logit stack, Concat MLP, Attn. pool, and Soft MoE equally. LLUNA uses the same evaluation inputs and fixed anchors. SIMS and MOSI use auxiliary binary sentiment accuracy, Acc-2(has0); the other tasks use ordinary classification accuracy for this diagnostic. Required task context is retained, and CirCor's maskable inputs are its four auscultation sites. For CirCor, we fit a `StandardScaler` and balanced logistic-regression head ($C = 1$) on training examples only for each of the six signal experts at each site; validation balanced accuracy selects one anchor per site, whose head remains fixed across states retaining that site.

Seven-baseline summary. Figure 4 combines the four conventional methods above with I^{2}MoE, CLARGA, and Set Transformer trained with missing-input views. All use the same fixed anchors, test examples, and legal availability states for this diagnostic. We first average legal anchor/view comparisons within each method and removal level, then give each of the seven methods equal weight. Thus, the two subgroups contribute four-sevenths and three-sevenths, respectively. Full uses all modalities; Minimal retains one independently maskable input. The combined mean gain declines from Full to Minimal on all eight tasks and is negative at Minimal on six; LLUNA's gain remains positive and exceeds this mean at both endpoints on every task. Table 33 gives the corresponding unnormalized gains and test counts.

Figure 4 uses task-normalized radii, while outer labels give raw gains in percentage points, in Baselines / LLUNA order. For each dataset, the shared scale is the largest absolute gain among the baseline mean and LLUNA at Full and Minimal. Dividing all four gains by this scale preserves their signs and relative magnitudes within that dataset; a common radial offset accommodates negative gains. Both panels use the same normalization, and the zero-gain circle separates beneficial from harmful collaboration.

Table 33: **Fixed-anchor gains (pp).** Min. retains one maskable input; Baselines averages seven methods equally. N_{test}: test examples.

		Baselines		Lluna	
Dataset	N_{test}	Full	Min.	Full	Min.
SIMS	1034	+10.07	-0.28	+10.93	+4.48
MOSI	686	+16.65	-0.08	+23.23	+11.13
MELD	2608	+16.29	+1.05	+19.79	+8.93
URF	994	+9.31	-2.36	+14.05	+3.29
Daily	177	+0.77	-2.10	+4.24	+5.93
WS	478	+2.14	-1.20	+3.70	+2.86
CirCor	88	-17.09	-21.59	+5.97	+1.70
ADNI	50	+9.05	+1.24	+15.33	+3.33

Four-method subgroup and LLUNA transitions. Table 34 reports the complete harmed/rescued rates and net gains for the original four-method conventional subgroup and LLUNA. At Minimal, LLUNA retains positive mean gains on all eight datasets, whereas this subgroup's mean is negative on all eight. Minimal retains one maskable input, which may still supply features from several models. CirCor uses site-specific anchors and retains demographic context across availability states. These are dataset-level averages: individual anchor/view comparisons can remain negative.

Table 34: **Fixed-anchor prediction transitions.** H/R are harmed/rescued percentages of all test examples; Net = R − H (pp). Independent rounding can yield 0.01-pp discrepancies. All tasks use ordinary accuracy, including auxiliary Acc-2(has0) for SIMS/MOSI. Conventional averages four methods equally.

		Conventional mean			Lluna		
Dataset	Removed	H (%)	R (%)	Net (pp)	H (%)	R (%)	Net (pp)
SIMS	0	8.09	18.05	+9.96	7.03	17.96	+10.93
	1	11.67	16.55	+4.88	7.41	16.51	+9.09
	2	16.53	14.35	-2.18	8.74	13.22	+4.48
MOSI	0	10.22	30.49	+20.27	8.79	32.02	+23.23
	1	16.33	25.24	+8.90	9.50	30.98	+21.48
	2	22.18	19.38	-2.81	12.88	24.00	+11.13
MELD	0	9.88	25.99	+16.11	7.17	26.96	+19.79
	1	13.53	21.57	+8.04	7.29	23.60	+16.31
	2	18.84	14.65	-4.18	7.89	16.82	+8.93
URF	0	11.33	22.46	+11.13	8.69	22.74	+14.05
	1	15.43	19.87	+4.43	10.63	20.39	+9.76
	2	20.23	16.83	-3.40	13.48	16.77	+3.29
Daily	0	17.16	17.44	+0.28	16.95	21.19	+4.24
	1	17.87	14.69	-3.18	14.12	20.06	+5.93
WS	0	14.07	16.14	+2.07	14.23	17.92	+3.70
	1	11.99	12.06	+0.07	13.08	17.12	+4.04
	2	10.04	8.16	-1.88	10.53	13.39	+2.86
CirCor	0	23.72	8.66	-15.06	3.69	9.66	+5.97
	1	24.81	8.40	-16.41	4.36	10.32	+5.97
	2	26.70	8.29	-18.42	5.02	10.04	+5.02
	3	30.75	8.45	-22.30	7.95	9.66	+1.70
ADNI	0	7.33	16.67	+9.33	3.33	18.67	+15.33
	1	8.92	15.25	+6.33	4.33	17.67	+13.33
	2	10.83	10.50	-0.33	5.33	8.67	+3.33

Additional controls trained with missing-input views. We repeat the fixed-anchor evaluation for I^{2}MoE, CLARGA, and Set Transformer after missing-input training, using the same anchors, test examples, legal views, and gain definition. We first average over legal anchor/view comparisons at each removal level within each method, then average the three methods equally. This subgroup contributes to the seven-baseline mean in Figure 4; its results are also summarized separately below. Table 35 gives Full and Minimal gains; Tables 36–39 give harmed/rescued rates and net gains for every method and removal level. The eight-task evaluation comprises 324 method/anchor/view comparisons, summarized into 72 method/level rows. The subgroup's mean gain declines from Full to Minimal on all tasks and is negative at Minimal on URF, Daily, WS, and CirCor; this does not imply that every intermediate transition or individual method behaves monotonically.

Table 35: **Cooperation gains after missing-input training.** Net accuracy gain over a fixed surviving model, averaged equally over I^{2}MoE, CLARGA, and Set Transformer. Full retains all modalities; Minimal retains one variable modality. Drop is Full minus Minimal, in percentage points.

Dataset	Full	Minimal	Drop
SIMS	+10.22	+2.27	7.95
MOSI	+11.81	+3.56	8.25
MELD	+16.53	+8.02	8.50
URF	+6.88	-0.97	7.85
Daily	+1.41	-0.66	2.07
WS	+2.23	-0.28	2.51
CirCor	-19.79	-20.64	0.85
ADNI	+8.67	+3.33	5.33

Table 36: **Fixed-anchor transitions on SIMS and MOSI.** Harmed and Rescued are percentages of test examples; Net is Rescued minus Harmed, in percentage points. Missing counts removed modalities.

SIMS

Method	Missing	Harmed	Rescued	Net
I^{2}MoE	0	7.16	18.57	+11.41
	1	8.85	16.84	+8.00
	2	10.83	13.44	+2.61
CLARGA	0	8.00	18.15	+10.16
	1	8.98	17.20	+8.22
	2	11.54	14.18	+2.64
Set Trans.	0	8.45	17.54	+9.09
	1	9.48	15.93	+6.45
	2	11.73	13.28	+1.55

MOSI

Method	Missing	Harmed	Rescued	Net
I^{2}MoE	0	13.36	26.68	+13.31
	1	13.51	25.17	+11.66
	2	16.18	20.46	+4.28
CLARGA	0	13.02	26.63	+13.61
	1	14.09	25.12	+11.03
	2	15.94	19.19	+3.26
Set Trans.	0	16.04	24.54	+8.50
	1	17.10	22.69	+5.59
	2	16.28	19.44	+3.16

Table 37: **Fixed-anchor transitions on MELD and URF.** Harmed and Rescued are percentages of test examples; Net is Rescued minus Harmed, in percentage points. Missing counts removed modalities.

MELD

Method	Missing	Harmed	Rescued	Net
I^{2}MoE	0	7.41	25.09	+17.68
	1	7.91	22.25	+14.34
	2	9.00	16.72	+7.72
CLARGA	0	7.62	23.72	+16.10
	1	8.05	21.47	+13.42
	2	8.63	16.72	+8.09
Set Trans.	0	7.99	23.79	+15.80
	1	8.44	21.26	+12.82
	2	8.30	16.55	+8.26

URF

Method	Missing	Harmed	Rescued	Net
I^{2}MoE	0	12.68	20.19	+7.51
	1	13.92	19.32	+5.40
	2	17.40	16.53	-0.87
CLARGA	0	13.58	19.99	+6.41
	1	14.64	18.63	+3.99
	2	17.00	15.63	-1.38
Set Trans.	0	13.62	20.32	+6.71
	1	15.11	19.06	+3.96
	2	17.20	16.53	-0.67

Table 38: **Fixed-anchor transitions on Daily and WS.** Harmed and Rescued are percentages of test examples; Net is Rescued minus Harmed, in percentage points. Missing counts removed modalities.

Daily

Method	Missing	Harmed	Rescued	Net
I^2MoE	0	16.67	15.82	-0.85
	1	14.97	12.71	-2.26
CLARGA	0	16.38	18.36	+1.98
	1	13.84	13.56	-0.28
Set Trans.	0	16.10	19.21	+3.11
	1	13.28	13.84	+0.57

WS

Method	Missing	Harmed	Rescued	Net
I^2MoE	0	13.81	15.62	+1.81
	1	9.83	10.53	+0.70
	2	6.49	5.51	-0.98
CLARGA	0	11.58	14.85	+3.28
	1	8.65	10.39	+1.74
	2	4.95	4.74	-0.21
Set Trans.	0	13.81	15.41	+1.60
	1	9.31	10.50	+1.19
	2	5.58	5.93	+0.35

Table 39: **Fixed-anchor transitions on CirCor and ADNI.** Harmed and Rescued are percentages of test examples; Net is Rescued minus Harmed, in percentage points. Missing counts removed modalities.

CirCor

Method	Missing	Harmed	Rescued	Net
I^2MoE	0	6.53	10.23	+3.69
	1	7.29	9.85	+2.56
	2	7.67	9.28	+1.61
	3	7.10	8.24	+1.14
CLARGA	0	6.82	8.24	+1.42
	1	5.11	8.52	+3.41
	2	5.30	8.05	+2.75
	3	6.25	7.67	+1.42
Set Trans.	0	72.16	7.67	-64.49
	1	72.16	7.67	-64.49
	2	72.16	7.67	-64.49
	3	72.16	7.67	-64.49

ADNI

Method	Missing	Harmed	Rescued	Net
I^2MoE	0	10.00	15.33	+5.33
	1	8.33	14.33	+6.00
	2	6.67	8.00	+1.33
CLARGA	0	10.67	16.00	+5.33
	1	12.00	14.00	+2.00
	2	6.67	11.33	+4.67
Set Trans.	0	5.33	20.67	+15.33
	1	6.67	16.00	+9.33
	2	6.67	10.67	+4.00

Sensitivity to baseline composition. Excluding Set Transformer on all eight tasks, the equally weighted mean of the remaining six baselines still declines from Full to Minimal on every task and is negative at Minimal on seven; LLUNA's gains remain positive and exceed this mean at both endpoints on all eight. On CirCor, I^{2}MoE and CLARGA alone average $+2.56$ pp at Full and $+1.28$ pp at Minimal, compared with $+5.97$ and $+1.70$ pp for LLUNA. Omitting each of the seven baselines in turn also preserves the decline in mean gain and LLUNA's advantage over the remaining mean at both endpoints on every task.

Exact-removal domain view. Figure 1a compares Full input and removal of exactly one or two modalities on the same seven eligible datasets. Affective computing contains SIMS, MOSI, MELD, and URF; scene understanding contains WS; medical prediction contains CirCor and ADNI. Daily is excluded at all three states because removing both of its modalities would leave no input for a surviving anchor. CirCor's Remove 2 endpoint averages two-site states; the other eligible tasks have three modalities, so Remove 2 coincides with Minimal. Datasets receive equal weight within each domain.

Figure annotations report the absolute decline in domain-mean collaboration gain from Full to Remove 2, in percentage points. Table 40 reports every plotted endpoint. Conventional gains are negative after two removals in all three domains, whereas LLUNA retains gains of $+6.96$, $+2.86$, and $+4.17$ pp, respectively.

Absolute declines decrease from 17.51 to 10.04 pp in affective computing and from 3.95 to 0.84 pp in scene understanding. They are nearly unchanged in medical prediction (6.51 versus 6.47 pp). Averaging the three absolute declines equally gives 9.32 pp for conventional collaboration and 5.79 pp for LLUNA, a 37.9% reduction. This is a domain-averaged comparison, not a percentage decrease in accuracy.

Table 40: **Exact-removal collaboration gains (pp).** Conventional averages four methods equally. Drop is Full minus Remove 2; datasets are equally weighted within each domain.

Domain	Predictor	Full	Remove 1	Remove 2	Drop (pp)
Affective	Conventional	14.37	6.56	−3.14	17.51
	LLUNA	17.00	14.16	6.96	10.04
Scene	Conventional	2.07	0.07	−1.88	3.95
	LLUNA	3.70	4.04	2.86	0.84
Medical	Conventional	−2.86	−5.04	−9.37	6.51
	LLUNA	10.65	9.65	4.17	6.47

Figure 12 gives individual conventional-method curves and prediction transitions for MELD, WS, and ADNI.

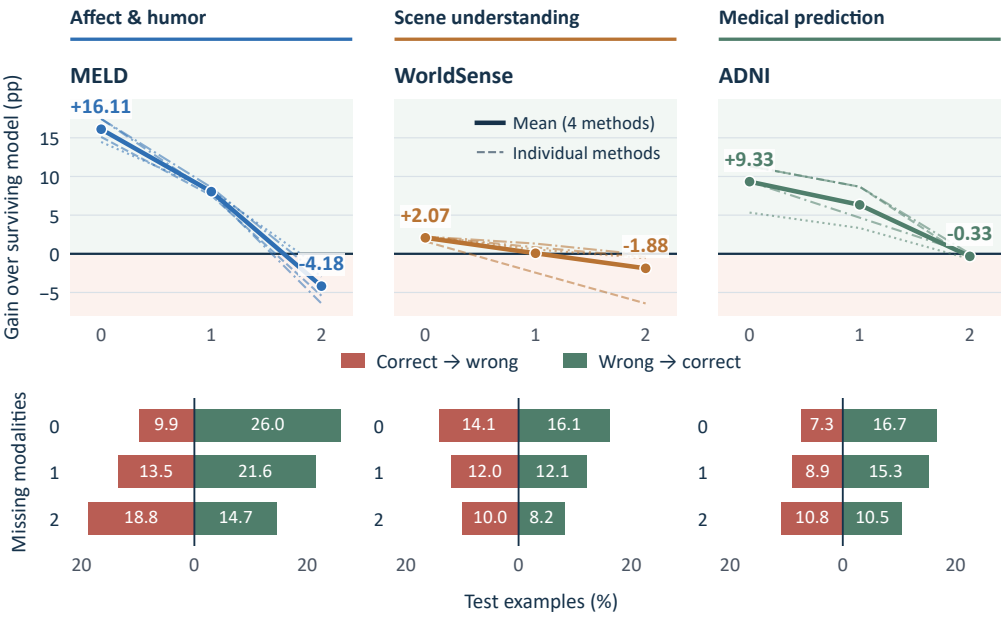

Figure 12: **Method-level gains and prediction transitions.** Solid curves show four-method mean gains; thin curves show individual methods. Bars report Harmed/Rescued shares of all test examples; Rescued minus Harmed equals ordinary-accuracy gain. Counts 0–2 denote removed modalities. These are individual datasets; anchors and collaboration checkpoints remain fixed across states.

E.2 Private-signal recoverability

We assess how accessible modality-specific features remain after collaboration. For each modality, cross-fitted ridge regression predicts its features from the other modalities; the residual serves as an operational private-signal target. A linear probe then decodes this same target from the modality-alone and full-collaboration representations. The score measures linear recoverability of that residual, rather than task accuracy.

Paired probe in Figure 1b. The paired Attn. pool diagnostic uses 64-dimensional representations and three-fold cross-validation. Scores use the coefficient of determination R^2, clipped below at zero and expressed as percentages: zero indicates no improvement over predicting the target mean, and 100 indicates perfect linear recovery. We average over maskable modalities within each dataset, then weight the eight datasets equally. The mean score is 58.52% for a modality alone and 33.47% after full collaboration; recovery decreases on every dataset. Table 42 gives the endpoints and Full/modality-alone retention ratios. These observations motivate direct access to modality features; the prediction-path and consistency-target ablations in Table 3 evaluate the corresponding design choices through task performance.

Across collaboration methods. Table 41 extends the diagnostic to Logit stack, Concat MLP, Attn. pool, and Soft MoE under Full-only and All-views training. For each modality, Minimal retains that input and any always-available context; the same residual target is decoded from Minimal and Full representations. Scores are variance-weighted, out-of-fold R^2, clipped below at zero and expressed as percentages. The residual is defined for independently maskable inputs, not individual pretrained models. For candidate-answer tasks, the correct-option representation is selected as the post-hoc probe readout. This supplementary analysis uses separate folds for Minimal and Full decoding; its endpoints are therefore reported separately from the paired probe in Figure 1b.

All four methods have lower Full recovery under both training policies. The decrease holds across all 192 modality-level comparisons: 24 maskable modalities, four methods, and two policies. Logit stack starts from lower recovery because it operates on modality logits; its smaller absolute decrease does not indicate better preservation of modality features.

Table 41: **Private-signal decodability across collaboration methods.** Probe R^2 (%), averaged equally over eight datasets. Drop is Minimal minus Full (pp); Full-only/All-views specify the training policy.

Method	Full-only training			All-views training		
	Minimal	Full	Drop	Minimal	Full	Drop
Logit stack	7.81	6.25	1.56	7.77	6.28	1.49
Concat MLP	40.95	20.91	20.03	40.45	20.45	20.00
Attn. pool	57.15	28.34	28.81	58.38	33.52	24.86
Soft MoE	56.85	23.13	33.72	57.52	27.31	30.21

E.3 Representation changes under modality removal

For the same example under states S_1 and S_2, let $\mathbf{z}_1^{\mathrm{att}}$ and $\mathbf{z}_2^{\mathrm{att}}$ be the corresponding Attn. pool shared representations. We measure their cosine distance:

$$1 - \frac{\langle \mathbf{z}_1^{\mathrm{att}}, \mathbf{z}_2^{\mathrm{att}} \rangle}{\|\mathbf{z}_1^{\mathrm{att}}\|_2 \|\mathbf{z}_2^{\mathrm{att}}\|_2}.$$

Figure 1c compares adjacent nested states differing by one modality with Full-to-Minimal states. It equally weights the six three-modality tasks (SIMS, MOSI, MELD, URF, WS, and ADNI), giving mean distances of 0.279 and 0.456, respectively. The adjacent category averages all single-removal edges, while Full-to-Minimal removes two modalities from Full. Daily has only one removal to Minimal and CirCor has three, so neither enters this exact-count comparison.

Table 42 reports all eight datasets and additionally includes equal-count states with different modality identities, such as TA/TV. Adjacent distances are smaller than Full-to-Minimal distances on every task with a multi-removal comparison; the categories coincide on Daily. These observations characterize representation changes as evidence is removed, with larger changes accompanying multi-step removals. PCC uses the connections between availability states to organize local guidance (Section 3); the task-level effect of this choice is evaluated by the pair-selection ablation in Table 3c.

Table 42: **Complete Attn. pool representation diagnostics.** Probe scores and retention are percentages. Drift is mean cosine distance for one-modality removal, Full to Minimal, and equal-count states with different modality identities. Summary rows weight datasets equally, with mean retention computed from per-dataset Full/modality-alone ratios.

	Modality-specific probe (%)			Cosine distance		
Dataset	Modality alone	Full	Retention	One removed	Full to Minimal	Same count
SIMS	69.40	40.41	58.22	0.271	0.442	0.758
MOSI	68.33	39.12	57.25	0.273	0.447	0.764
MELD	68.62	37.74	55.01	0.276	0.451	0.763
URF	66.81	34.21	51.21	0.302	0.489	0.778
Daily	57.25	45.23	79.01	0.329	0.329	1.039
WS	62.10	41.36	66.61	0.297	0.483	0.800
CirCor	30.30	12.95	42.73	0.155	0.381	0.386
ADNI	45.34	16.76	36.97	0.257	0.421	0.729
Dataset mean	58.52	33.47	55.88	0.270	0.430	0.752
Three-modality mean	–	–	–	0.279	0.456	0.765

E.4 Prediction changes across availability states

As a complementary output diagnostic, we evaluate fixed collaboration models on the diagnostic cohorts reported in Table 43. For the same example under states S_1 and S_2, we compare the model's class-probability distributions $\mathbf{q}_1$ and $\mathbf{q}_2$ using normalized Jensen–Shannon divergence,

$\mathrm{JSD}(\mathbf{q}_1, \mathbf{q}_2)/\ln 2 \in [0, 1]$. Zero means identical probability distributions. The table gives diagnostic-cohort sizes; the main evaluation split sizes appear in Table 12. This diagnostic applies to the six classification tasks; SIMS and MOSI have scalar regression outputs. Figure 13 presents these output differences alongside the separate Attn. pool representation analysis.

Adjacent pairs differ by one removed modality, and Full-to-Minimal pairs connect Full to each minimal state. Same-level JSD compares states at the largest incomplete level; the representation summary includes equal-cardinality pairs at all nonempty levels. Full-to-Minimal JSD exceeds Adjacent JSD on all five classification tasks with a multi-removal contrast; Daily's pair sets coincide. Table 43 provides the exact means, and Figure 13b shows variation across individual state pairs.

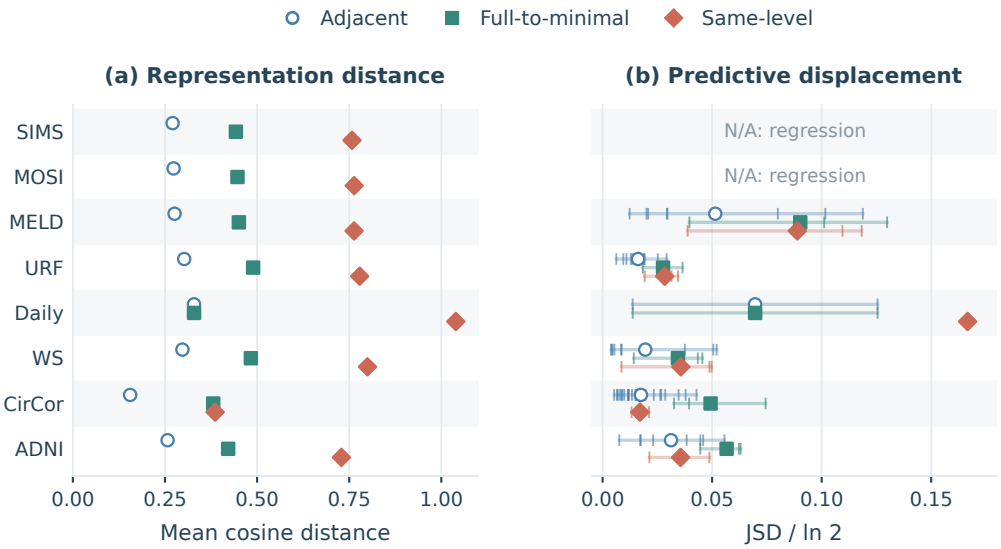

Figure 13: **Complementary representation and prediction diagnostics.** (a) Attn. pool cosine distance on eight tasks. (b) Normalized predictive JSD on the diagnostic cohorts of six classification tasks; regression JSD is N/A. Large markers show category means; ticks and spans show individual state-pair means and their range. Pair definitions appear above; Table 43 reports exact JSD values and cohort sizes.

Table 43: **Prediction differences across availability states.** Cells show $10^3 \mathrm{JSD}/\ln 2$ for fixed diagnostic models, with ratios to Adjacent in parentheses. Examples gives the number of evaluated items in each diagnostic cohort. Regression JSD is N/A; Daily's Adjacent and Full-to-minimal sets coincide.

Dataset	Examples	Adjacent	Same-level	Full-to-minimal
SIMS	4403	N/A	N/A	N/A
MOSI	2199	N/A	N/A	N/A
MELD	13704	51.43 (1.00×)	88.86 (1.73×)	90.23 (1.75×)
URF	9588	16.18 (1.00×)	28.37 (1.75×)	27.60 (1.71×)
Daily	1019	69.63 (1.00×)	166.61 (2.39×)	69.63 (1.00×)
WS	2694	19.48 (1.00×)	35.73 (1.83×)	34.41 (1.77×)
CirCor	585	17.49 (1.00×)	17.06 (0.98×)	49.30 (2.82×)
ADNI	284	31.26 (1.00×)	35.60 (1.14×)	56.61 (1.81×)

F IMPLEMENTATION AND REPRODUCIBILITY

Appendix B specifies the task splits, feature inputs, baseline interfaces, and main-run optimization settings. The released code will include preprocessing scripts, fixed split files, unimodal encoders, feature caches where redistribution is permitted, all modality-availability state definitions, hyperparameter search spaces, random seeds, and commands that reproduce every table. Modality features are projected and standardized before collaboration. Task-level optimization, regularization, early stopping, and availability-state sampling follow the corresponding per-dataset configuration.

Compatibility-graph construction. Task heads are fitted to frozen model features on the training partition, and their predictions on the training examples used for graph construction determine the compatibility weights described in Section 3.2. For classification, we use $\mathrm{sim}(\mathbf{q}_i, \mathbf{q}_j) = 1 - \frac{1}{2}\|\mathbf{q}_i - \mathbf{q}_j\|_1$, where $\mathbf{q}_i$ and $\mathbf{q}_j$ are class-probability vectors from models i and j on the same training example. For regression, we use $\mathrm{sim}(r_i, r_j) = \exp(-|r_i - r_j|)$, where r_i, r_j are the corresponding standardized scalar predictions. Using the absolute rather than squared gap yields a more gradual decay for large standardized prediction differences. The configured weight floor and symmetrization are applied. When graph-temperature scaling is enabled, it transforms an agreement weight A_{ij} to $\exp((\min(A_{ij}, 1) - 1)/\tau)$, where $\tau > 0$ is the graph temperature, before union top-k pruning. Here k counts the highest-weight neighbors selected per model; the union retains an edge if either endpoint selects it, and positive self-loops are then retained. The graph temperature acts on agreement weights, not on per-example predictive probabilities. These weights remain fixed throughout collaboration training and inference. For each availability state, the graph is restricted to available models and incoming weights are normalized with a numerical stabilizer.

F.1 PCC WEIGHTING AND DISCREPANCY CONFIGURATION

The base discrepancy underlying Equation (2) is the cosine distance between normalized common consensuses of the same example at two availability states:

$$d(S_k, S_{k+1}) = 1 - \bar{\mathbf{z}}_k^\top \bar{\mathbf{z}}_{k+1}. \tag{9}$$

Configurations may shape a base distance d as $\max(0, d - \delta)^\rho$ with a dead zone $\delta \geq 0$ and power $\rho > 0$; the base objective has $\delta = 0$ and $\rho = 1$. Edge weights normalize contributions within each child missingness level and may redistribute that level's weight across removed modalities. The chain budget K, level weights, δ, ρ, and loss coefficients belong to the per-task configuration; the code specifies the chain-sampling implementation. The stability theorem in Appendix G uses these base distances along a path.

Prediction interfaces. The examples used for compatibility estimation come from the training partition; validation and test examples are excluded. Task heads fitted on training-partition frozen features supply the predictions used to construct the graph. These predictions are not additional per-example inputs to the collaboration model. Model features are masked before and after adaptation; zero slots represent unavailable models and are accompanied by an explicit availability mask. Both prediction heads use that mask to adapt their outputs to the available modalities, while the private path retains direct access to adapted modality features. PCC acts directly on the normalized common consensus; the shared adapters also receive its gradients through the common branch. For each edge, the richer endpoint is detached and the poorer endpoint receives the direct alignment gradient. Parameters are shared across states, so these updates also affect subsequent predictions at richer states; an intermediate state can itself be the poorer endpoint of another sampled edge. Numerical normalization uses an epsilon near zero.

The attention pooling in group readout uses a learned scoring vector $\mathbf{w}$ and a normalized attention weight α_i for each available model i:

$$\mathbf{z} = \sum_i \alpha_i \mathbf{c}_i^{(L)}, \qquad \alpha_i = \frac{\exp(\mathbf{w}^\top \mathbf{c}_i^{(L)})}{\sum_j \exp(\mathbf{w}^\top \mathbf{c}_j^{(L)})}. \tag{10}$$

Both sums run only over available models; the state index is omitted, as in the common-branch equations in the main text.

G Additional Derivations

G.1 Paths between arbitrary modality subsets

For any two distinct nonempty endpoint states $S_0, S_D \subseteq \mathcal{M}$, add the modalities in $S_D \setminus S_0$ one at a time to reach $S_0 \cup S_D$, then remove the modalities in $S_0 \setminus S_D$ one at a time to reach S_D. Every intermediate state is nonempty: the addition segment contains S_0, and the removal segment contains S_D. Consecutive states differ by exactly one modality, so they form a path on the undirected version of the cover graph. Its length is

$$D = |S_D \setminus S_0| + |S_0 \setminus S_D|. \tag{11}$$

Each modality in the symmetric difference must change availability at least once along any path from S_0 to S_D; this construction therefore has the minimum number of edges. For example, TA $\rightarrow$ TAV $\rightarrow$ TV connects equal-size states with different modality identities. For identical endpoints, the representations coincide and the stability statements hold trivially with an empty path.

The cosine distance in Equation (9) is symmetric, even though PCC trains on each edge in the richer-to-poorer direction. Thus an addition step in the path uses the same distance as the corresponding removal edge. Stop-gradient directs the alignment update toward the poorer endpoint while preserving this geometric distance.

G.2 Local-to-global representation stability

Throughout this subsection, model parameters are fixed and inference is deterministic.

Theorem 1 (Local-to-global representation stability). Fix an input and a path $S_0, S_1, \ldots, S_D$ of $D \geq 1$ single-modality transitions. Let $d_k = 1 - \bar{\mathbf{z}}_{k-1}^\top \bar{\mathbf{z}}_k$ denote the cosine discrepancy between consecutive normalized common representations. If the common consensus is nonzero at every state on the path, then

$$\|\bar{\mathbf{z}}_0 - \bar{\mathbf{z}}_D\|_2 \leq \sum_{k=1}^{D} \sqrt{2d_k} \leq \sqrt{2D \sum_{k=1}^{D} d_k}. \tag{12}$$

Proof of Theorem 1. Fix an input and a path $S_0, S_1, \ldots, S_D$ with $D \geq 1$. Assume $\mathbf{z}_k \neq \mathbf{0}$ for all $k \in \{0, \ldots, D\}$, so that $\bar{\mathbf{z}}_k = \mathbf{z}_k / \|\mathbf{z}_k\|_2$ is defined and has unit norm. For each consecutive pair, the definition $d_k = d(S_{k-1}, S_k) = 1 - \bar{\mathbf{z}}_{k-1}^\top \bar{\mathbf{z}}_k$ gives $0 \leq d_k \leq 2$ and

$$\|\bar{\mathbf{z}}_{k-1} - \bar{\mathbf{z}}_k\|_2^2 = \|\bar{\mathbf{z}}_{k-1}\|_2^2 + \|\bar{\mathbf{z}}_k\|_2^2 - 2\bar{\mathbf{z}}_{k-1}^\top \bar{\mathbf{z}}_k \tag{13}$$

$$= 2d_k. \tag{14}$$

Telescoping the endpoint difference and applying the triangle inequality therefore yields

$$\|\bar{\mathbf{z}}_0 - \bar{\mathbf{z}}_D\|_2 = \left\| \sum_{k=1}^{D} (\bar{\mathbf{z}}_{k-1} - \bar{\mathbf{z}}_k) \right\|_2 \tag{15}$$

$$\leq \sum_{k=1}^{D} \|\bar{\mathbf{z}}_{k-1} - \bar{\mathbf{z}}_k\|_2 \tag{16}$$

$$= \sum_{k=1}^{D} \sqrt{2d_k}. \tag{17}$$

Applying Cauchy–Schwarz to the vectors $(1, \ldots, 1)$ and $(\sqrt{2d_1}, \ldots, \sqrt{2d_D})$ gives

$$\left(\sum_{k=1}^{D} \sqrt{2d_k} \right)^2 \leq \left(\sum_{k=1}^{D} 1^2 \right) \left(\sum_{k=1}^{D} 2d_k \right) = 2D \sum_{k=1}^{D} d_k. \tag{18}$$

□

The result holds pointwise for every path satisfying the nonzero-consensus assumption, using the base cosine distances d_k along that path. Taking the minimum over admissible paths gives the tightest pathwise bound. Both path length and accumulated local drift contribute to the bound, so a longer path with smaller discrepancies can provide tighter control.

Connection to the training objective. The bounds use the base cosine distances along the chosen path. For a fixed sample, a weighted PCC sum with positive weights bounds the sum of base distances on any included simple path after division by the smallest weight on that path. Under random chain sampling, the corresponding expected sum provides this control when every required edge has positive expected weight, using those expected weights in the conversion. The margin and power configuration in Appendix F.1 is related to the base distances by

$$\sum_{k=1}^{D} d_k \le D\delta + \sum_{k=1}^{D} \left[\max(0, d_k - \delta)^{\rho}\right]^{1/\rho}. \tag{19}$$

Thus weighted or transformed sampled objectives can be related to the same geometric bound through their weights, sampling coverage, and discrepancy settings. Stop-gradient preserves the forward distances while directing the alignment update toward the poorer state. The proof uses exact unit normalization; for an epsilon-stabilized implementation, it applies when each path consensus norm is at least the normalization floor.

G.3 Supplementary output-drift bound

Fix the model parameters and one input, and use deterministic inference. At state S_k, write the complete prediction from Equation (1) as $\widehat{\mathbf{y}}_k = f(\mathbf{p}(S_k), \mathbf{z}_k)$. Here $\mathbf{p}(S_k)$ collects independently processed direct-evidence slots in a fixed order, zero-fills unavailable slots, and includes every availability indicator used by the prediction mapping. Any subsequent mixing on the private path belongs to the prediction mapping f. Because each retained evidence slot depends only on its own input, its value is unchanged when other slots become unavailable. Thus the difference $\mathbf{p}(S_0) - \mathbf{p}(S_D)$ is confined to changed evidence slots and availability coordinates; the private prediction itself remains free to adapt to these changes.

For direct-input vectors $\mathbf{p}, \mathbf{p}'$ and raw common consensuses $\mathbf{z}, \mathbf{z}'$ in a product domain containing the compared inputs, assume the complete mapping is Lipschitz in each block:

$$\|f(\mathbf{p}, \mathbf{z}) - f(\mathbf{p}', \mathbf{z})\|_2 \le L_{\mathrm{priv}} \|\mathbf{p} - \mathbf{p}'\|_2, \tag{20}$$

$$\|f(\mathbf{p}, \mathbf{z}) - f(\mathbf{p}, \mathbf{z}')\|_2 \le L_{\mathrm{com}} \|\mathbf{z} - \mathbf{z}'\|_2. \tag{21}$$

The constant L_{priv} accounts for the private path and both heads' explicit availability conditioning while the common consensus is held fixed. The constant L_{com} includes the fixed residual scale in Equation (1). This formulation accounts separately for changes in the direct inputs and in the raw common consensus used by the prediction head.

Corollary 1 (Output drift of the complete predictor)**.** *Under the assumptions of Theorem 1 and the block-Lipschitz conditions above, the complete predictions satisfy*

$$\begin{aligned}\|\widehat{\mathbf{y}}_0 - \widehat{\mathbf{y}}_D\|_2 \le{}& L_{\mathrm{priv}} \|\mathbf{p}(S_0) - \mathbf{p}(S_D)\|_2 \\ &+ L_{\mathrm{com}} \Big[\big(\|\mathbf{z}_0\|_2 - \|\mathbf{z}_D\|_2\big)^2 + 2D \|\mathbf{z}_0\|_2 \|\mathbf{z}_D\|_2 \sum_{k=1}^{D} d_k \Big]^{1/2}.\end{aligned} \tag{22}$$

Proof of Corollary 1. Insert the intermediate prediction $f(\mathbf{p}(S_D), \mathbf{z}_0)$ and apply the block-Lipschitz conditions:

$$\begin{aligned}\|\widehat{\mathbf{y}}_0 - \widehat{\mathbf{y}}_D\|_2 &\le \|f(\mathbf{p}(S_0), \mathbf{z}_0) - f(\mathbf{p}(S_D), \mathbf{z}_0)\|_2 \\ &\quad + \|f(\mathbf{p}(S_D), \mathbf{z}_0) - f(\mathbf{p}(S_D), \mathbf{z}_D)\|_2 \\ &\le L_{\mathrm{priv}} \|\mathbf{p}(S_0) - \mathbf{p}(S_D)\|_2 + L_{\mathrm{com}} \|\mathbf{z}_0 - \mathbf{z}_D\|_2.\end{aligned} \tag{23}$$

For nonzero endpoint consensuses, their raw Euclidean distance decomposes exactly into magnitude and direction changes:

$$\begin{aligned}\|\mathbf{z}_0 - \mathbf{z}_D\|_2^2 &= (\|\mathbf{z}_0\|_2 - \|\mathbf{z}_D\|_2)^2 \\ &\quad + \|\mathbf{z}_0\|_2\|\mathbf{z}_D\|_2\|\bar{\mathbf{z}}_0 - \bar{\mathbf{z}}_D\|_2^2 \\ &\leq (\|\mathbf{z}_0\|_2 - \|\mathbf{z}_D\|_2)^2 + 2D\|\mathbf{z}_0\|_2\|\mathbf{z}_D\|_2 \sum_{k=1}^{D} d_k, \end{aligned} \tag{24}$$

where the last inequality follows from Theorem 1. Taking the square root and substituting into the preceding prediction bound proves Equation (22). □

The bound separates changes in direct evidence from changes in the magnitude and direction of the common consensus. The pathwise PCC distances enter through the directional component.

H LIMITATIONS

Missingness conditions. Our evaluation focuses on controlled removal of entire modalities with known availability states. This setting isolates the effect of missing inputs across modality combinations. Further evaluation is needed for partial corruption, temporal misalignment, and naturally occurring missingness associated with sample characteristics.

Model-pool updates. LLUNA is trained for a specified pool of frozen models. Our pool-size and replacement experiments evaluate alternative pools after training the corresponding collaboration modules. Supporting changes to the pool without retraining the full collaboration module is a direction for future work.

I BROADER IMPACT

More reliable behavior under missing modalities is valuable in healthcare, sensing, and human-centered applications, but it does not by itself ensure calibration, fairness, or safety. Performance should be audited separately across missingness mechanisms and demographic groups. Predictions made with fewer modalities should also expose the available-modality state to downstream users rather than presenting the same confidence semantics as a complete observation.