REFUSE: WIDENING MULTIMODAL FUSION VIA RECURSIVE SELECTIVE REFINEMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Standard multimodal fusion performs a single integration of all modality representations and feeds the fused result for prediction. This one-shot paradigm risks under-fusing modality information, particularly for weak modalities whose task-relevant signals can be overwhelmed by dominant modalities. To this end, we propose ReFuse (Refinement Fusion), a recursive fusion framework that iteratively diagnoses information insufficiency in the current fused representation and selectively re-fuses only modalities that still carry task-relevant information not yet captured. At the core of ReFuse is an information sufficiency diagnoser that estimates the conditional mutual information between each modality and label given the current fused representation, driving both modality selection and adaptive early termination. The fusion module is parameter-shared across all recursion steps, so ReFuse *widens* modality fusion rather than *deepening* the network. We provide theoretical analysis showing that ReFuse monotonically increases task-relevant information, achieves approximate sufficiency upon termination, and enjoys better parameter efficiency under equal parameter budgets. Experiments on multiple multimodal benchmarks demonstrate that ReFuse consistently outperforms one-shot fusion baselines and, under strict equal-parameter comparison, outperforms deeper fusion networks especially under modality imbalance.

1 INTRODUCTION

Multimodal learning aims to integrate complementary information from multiple sensory modalities (e.g., text, audio, vision) to improve prediction (Baltrušaitis et al., 2019). A dominant design paradigm is to encode each modality independently, perform a single fusion operation (attention, gating, etc.), and then feed the fused representation into a deep network for further processing (Tsai et al., 2019; Wang et al., 2025; Li et al., 2026). While effective, this *one-shot fusion* paradigm has a fundamental limitation: the fusion layer has only one opportunity to extract task-relevant information from each modality. When modalities are imbalanced, which is a common scenario in real-world applications where dominant modality carry stronger signals than weak modalities, the single fusion step is prone to under-fusing the weak modality. This is because the gradients and information of the weak modalities are overwhelmed by the dominant modality (Wei et al., 2025c;a).

Deepening the fusion network is a natural remedy, but it increases parameter count and may not address the core issue: the bottleneck is insufficient modality interaction, not insufficient nonlinear feature extraction. A deeper network still performs fusion only at specific layers, and weak modality information can be lost early and never recovered.

To this end, we propose ReFuse (Refinement Fusion), a recursive refinement fusion framework that replaces one-shot fusion with an iterative "diagnose-select-refuse" loop. Starting from an initial fused representation h_0, ReFuse at each step t: (1) *diagnoses* which modalities still carry task-relevant information not captured by h_t, by estimating the conditional mutual information $I(m_i; y \mid h_t)$; (2) *selects* the information-deficient modalities; and (3) *re-fuses* only those modalities with h_t to produce h_{t+1}. The loop terminates adaptively when no modality provides significant additional task information, or when a maximum step budget is reached. Unlike modality reconstruction probes that measure whether h_t preserves *all* information of m_i (Han et al., 2021; Lian et al., 2023; Xiong et al., 2026) which might includes noisy and redundant information, our diagnoser estimates $I(m_i; y \mid h_t)$, which measures only the *task-relevant* information that m_i can additionally provide. This is both

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.

theoretically principled (aligned with the information bottleneck objective) and practically efficient. The proposed fusion strategy ensures that all discriminative information in all modalities can be passed into the multimodal representation, effectively alleviating the modality imbalance problem and avoiding the inclusion of noisy and redundant information.

Notably, the fusion module in ReFuse is parameter-shared across all recursion steps. This means ReFuse does not deepen the network in the parameter sense. Instead, it *widens* the fusion process by increasing the number of cross-modal interactions. Under an equal-parameter budget, ReFuse can afford more interactions than a deeper network that allocates parameters to each layer independently.

In addition, we provide theoretical contributions for ReFuse: (1) **Information monotonicity and approximate sufficiency**: We prove that the task-relevant information learned by ReFuse is monotonically non-decreasing across recursion steps, and that upon termination h_T approximates a sufficient statistic within a bounded residual; (2) **Greedy selection guarantee**: Under mild submodularity conditions, the greedy modality selection achieves a $(1 - 1/e)$ approximation to the optimal subset selection; (3) **Parameter efficiency**: We show that parameter sharing gives ReFuse a better capacity-per-parameter trade-off than an equal-computation deep network.

Our contributions are summarized as follows:

- We propose ReFuse, a recursive fusion framework that iteratively diagnoses information insufficiency and selectively re-fuses deficient modalities, with adaptive early termination.
- We design an information sufficiency diagnoser based on conditional mutual information $I(m_i; y \mid h_t)$, which is theoretically principled and computationally efficient.
- We provide theoretical analysis of information monotonicity, approximate sufficiency, greedy selection guarantees, and parameter efficiency advantages.
- Experiments on multiple multimodal benchmarks demonstrate that ReFuse achieves state-of-the-art performance. ReFuse outperforms one-shot fusion and equal-parameter deeper/wider networks, especially under modality imbalance.

2 Related Work

Multimodal fusion strategies. Early fusion methods concatenate modality features and feed them to a shared network (Ngiam et al., 2011), a design recently revived by token-level early-fusion transformers that mix modalities inside one backbone from the earliest layers (Schlarmann et al., 2025; Yang et al., 2025). Tensor-based fusion (Zadeh et al., 2017) captures high-order interactions but is parameter-intensive. Attention-based fusion (Tsai et al., 2019; Mai et al., 2025) models cross-modal interactions but incurs quadratic complexity in the number of modalities. Gated fusion (Arevalo et al., 2017) uses learned gates to control modality contributions, with recent adaptive, entropy-conditioned gates further down-weighting noisy or dominant modalities (Wu et al., 2025a). However, all these methods perform fusion in a single forward pass without iteratively refining the fused representation based on diagnosed insufficiency.

Recursive and iterative multimodal architectures. Several recent works explore recursion in multimodal models. Gated Recursive Fusion (GRF) (Shihata, 2025) processes modalities sequentially with a shared context vector and GRU-inspired gating, but follows a fixed order and fuses every modality exactly once. There is no information sufficiency diagnosis, no selective re-fusion of already-processed modalities, and no adaptive termination in GRF. RecursiveVLM (Xu et al., 2026) applies recursive Transformer loops to large multimodal models, but recursion operates at the entire model level with fixed depth, and does not include modality-level information diagnosis or selective fusion. SDRFPT-Net (Zhou et al., 2025) applies recursive transformations within a single modality for spectral feature refinement, not across modalities. HIVE (Zhang et al., 2026) uses loop-based Transformers for latent reasoning with hierarchical visual cues, but does not diagnose or selectively fuse modalities. Unlike all these works, ReFuse introduces a closed-loop "diagnose–select–refuse" mechanism with modality-level conditional information estimation and adaptive termination.

Weak modality enhancement and modality imbalance. Recent works address modality imbalance through gradient rebalancing (Peng et al., 2022; Wei et al., 2025c;a; Luo et al., 2025), weak

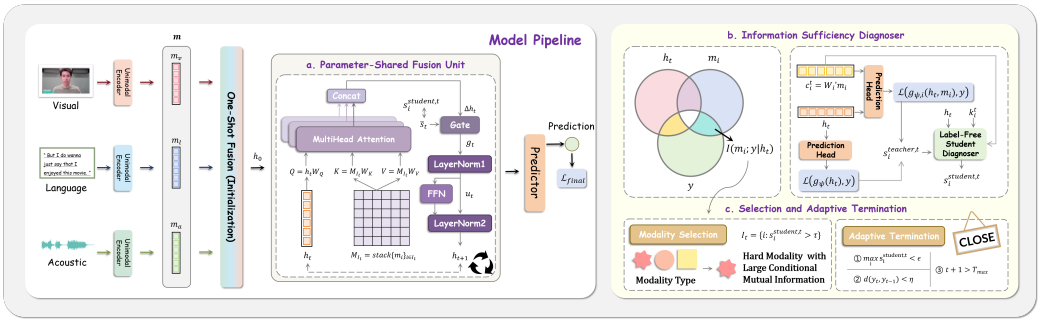


Figure 1: The pipeline of ReFuse. We use three modalities for illustration.

modality enhancement (Bi et al., 2025; He et al., 2026), and information routing (Kim et al., 2026). These methods typically apply a one-time enhancement or rebalancing before fusion. ReFuse is complementary: its recursive loop naturally gives weak modalities multiple opportunities to contribute, and the diagnoser explicitly identifies when a weak modality's information is still missing.

3 METHOD

3.1 PROBLEM FORMULATION

Let the input consist of M modalities. Unimodal encoders first map each modality into a common representation space. We denote the resulting i-th modality vector by $m_i \in \mathbb{R}^d$ and write $\mathbf{m} = (m_1, \ldots, m_M)$ for the joint tuple, with label $y \in \mathcal{Y}$. One-shot fusion model learns $f : (\mathbb{R}^d)^M \to \mathbb{R}^d$ and a predictor $g_\psi : \mathbb{R}^d \to \mathcal{Y}$, producing $h = f(\mathbf{m})$ and $\hat{y} = g_\psi(h)$ in a single forward pass.

ReFuse instead learns a recursive fusion process:

$$h_0 = f_0(\mathbf{m}), \qquad h_{t+1} = F_\theta(h_t, \{m_i\}_{i \in I_t}), \quad t = 0, 1, \ldots, T-1, \tag{1}$$

where f_0 is a one-shot fusion that initializes the state, $I_t \subseteq \{1, \ldots, M\}$ is the subset of modalities selected for re-fusion at step t, and T is an adaptively determined, sample-specific number of recursion steps bounded by $T_{\max}$. The final representation h_T is used for prediction: $\hat{y} = g_\psi(h_T)$.

ReFuse has three core components: (i) a single parameter-shared fusion unit F_θ reused at every step; (ii) an information-sufficiency diagnoser that, at state h_t, produces a scalar insufficiency score s_i^t for every modality, which is a tractable surrogate of the conditional information gain $I(m_i; y \mid h_t)$; and (iii) a selection-and-termination strategy that decides I_t and T. We detail each below.

3.2 PARAMETER-SHARED FUSION UNIT

The fusion unit F_θ takes the current representation $h_t \in \mathbb{R}^d$ and the selected modality vectors $\{m_i\}_{i \in I_t}$, and produces an updated representation. We design it as a gated cross-attention with a residual feed-forward block. First, we stack the selected modalities as keys/values:

$$M_{I_t} = \text{stack}\{m_i\}_{i \in I_t} \in \mathbb{R}^{|I_t| \times d}. \tag{2}$$

Multi-head cross-attention uses h_t as query and M_{I_t} as key/value:

$$\Delta h_t = \text{MultiHeadAttn}(Q = h_t W_Q,\ K = M_{I_t} W_K,\ V = M_{I_t} W_V). \tag{3}$$

A gate controls the update strength, conditioned on the current state, the attention output, and the largest selected insufficiency score $\bar{s}_t = [\max_{i \in I_t} s_i^{\text{student},t}]_+ = \max(0, \max_{i \in I_t} s_i^{\text{student},t})$:

$$g_t = \sigma(W_g[h_t;\ \Delta h_t;\ \bar{s}_t]) \in [0, 1]^d, \tag{4}$$

where σ is the elementwise sigmoid and $s_i^{\text{student},t}$ is defined in Section 3.3. Basing the gate on a continuous student score keeps the update differentiable and label-free at inference. A first residual LayerNorm is followed by a parameter-shared position-wise feed-forward block FFN with a second residual LayerNorm, as in a Transformer layer:

$$u_t = \text{LN}_1(h_t + g_t \odot \Delta h_t), \qquad h_{t+1} = \text{LN}_2(u_t + \text{FFN}(u_t)). \tag{5}$$

The gate (i) allows strong updates when information insufficiency is severe (i.e., when $s_i^{\text{student},t}$ is large) and fine adjustments when it is mild, and (ii) attenuates updates when a selected modality carries mostly noise. Crucially, all parameters of F_θ, including W_Q, W_K, W_V, W_g, the FFN, and the two LayerNorms, are shared across every recursion step and every modality subset, so the parameter count is independent of T. In other words, increasing the number of fusion iterations widens the computation without adding parameters (i.e., "widen fusion instead of deepening the network").

Architecture-agnostic design. We emphasize that ReFuse is a wrapper around an arbitrary fusion function rather than a specific fusion operator. The gated cross-attention in Eq. 3–5 is one instantiation of the fusion unit F_θ, but the diagnose-select-refuse loop is agnostic to this choice: F_θ can be replaced by other fusion strategies, while the diagnoser, selection, and termination machinery remain unchanged (see Section 5.3). This controls for fusion-operator capacity and isolates the gain that is attributable to the iterative mechanism itself, rather than to a stronger one-shot fusion operator.

3.3 Information Sufficiency Diagnoser

The target quantity of the diagnoser is, for each modality i and step t, the task-relevant information that m_i can still add beyond what h_t already captures, namely the conditional mutual information $I(m_i; y \mid h_t)$. When $I(m_i; y \mid h_t) \approx 0$, h_t already contains all task-relevant information from m_i (or m_i is itself task-irrelevant), and re-fusing m_i is unnecessary. A key challenge is that the ideal diagnostic signal requires the label y, which is unavailable at inference time. We address this with a teacher-student design: during training, we compute a label-dependent teacher signal and use it to supervise a lightweight student diagnoser that operates without labels at inference time.

3.3.1 Teacher Signal: Dual-Predictor Loss Difference

Directly estimating high-dimensional conditional mutual information is intractable. During training, where labels are available, we use a **dual-predictor loss difference** as a computable surrogate. Let $g_\psi : \mathbb{R}^d \to \mathcal{Y}$ be the shared task prediction head. For each modality i, a modality-specific projection $W_i' \in \mathbb{R}^{d\times d}$ (shared across recursion steps) forms the conditional feature $c_i^t = W_i' m_i$, and a conditional predictor reuses the same head g_ψ:

$$g_{\psi,i}(h_t, m_i) = g_\psi\big(\mathrm{LN}\big(h_t + c_i^t\big)\big) = g_\psi\big(\mathrm{LN}\big(h_t + W_i' m_i\big)\big)\,. \tag{6}$$

The per-sample teacher insufficiency score is

$$s_i^{\text{teacher},t} = \mathcal{L}(g_\psi(h_t), y) - \mathcal{L}(g_{\psi,i}(h_t, m_i), y)\,. \tag{7}$$

Both terms are evaluated with stopped gradients: $s_i^{\text{teacher},t}$ is a supervision target only and never back-propagates into g_ψ or the fusion unit (it is detached in the implementation). Intuitively, a positive $s_i^{\text{teacher},t}$ means that incorporating m_i reduces the prediction loss, indicating m_i carries task-relevant information not yet in h_t. A score near zero means m_i provides no additional task benefit. A negative score means m_i introduces noise and should be skipped.

Additionally, we give an information-theoretic justification for this implementation. For logarithmic loss (cross-entropy), when both predictors attain their respective Bayes risks, the expected dual-predictor loss difference equals the conditional mutual information $I(m_i; y \mid h_t)$; for squared loss it is a strictly increasing function of this quantity (with an exact closed form under Gaussian conditionals). For finite-capacity predictors estimated on finite batches, the realized score obeys the lower bound (see Appendix B for the proof):

$$\mathbb{E}_{h_t, m_i, y}\left[s_i^{\text{teacher},t}\right] \geq I(m_i; y \mid h_t) - \epsilon_{\text{approx}}, \tag{8}$$

where ϵ_{approx} collects the two predictors' excess Bayes risks together with an $O(1/\sqrt{B})$ finite-sample term for batch size B, and tends to zero as predictor capacity and batch size grow.

3.3.2 Student Diagnoser: Label-Free Network

The teacher signal requires labels and thus cannot be used at inference. We train a lightweight, label-free student diagnoser D_ϕ that reproduces the teacher score from inputs available without y, using the teacher signal as soft supervision. Specifically, for modality i the student receives three quantities, all available without the label: (1) the **current state** h_t; (2) the **projected modality**

$c_i^t = W_i' m_i$ (the same conditional projection used by the teacher in Eq. 6); (3) a scalar **prediction-shift** k_i^t that measures how much the prediction moves when modality i is tentatively added,

$$k_i^t = \begin{cases} d(p_\psi(y \mid h_t),\, p_{\psi,i}(y \mid h_t, m_i))\,, & \text{classification (a distribution divergence),} \\ |g_\psi(\mathrm{LN}(h_t + c_i^t)) - g_\psi(h_t)|\,, & \text{regression (shift of the predicted value).} \end{cases} \tag{9}$$

A large k_i^t indicates that m_i changes the current prediction and therefore plausibly carries information not yet in h_t. Under well-calibrated predictors it tracks the teacher loss difference, providing a strong label-free prior. Afterwards, one shared MLP $D_\phi : \mathbb{R}^{2d+1} \to \mathbb{R}$ is applied independently to every modality (then stacked into an M-vector of scores):

$$s_i^{\text{student},t} = D_\phi\big(\mathrm{sg}(h_t),\, c_i^t,\, \mathrm{sg}(k_i^t)\big)\,, \qquad (s_1^{\text{student},t}, \ldots, s_M^{\text{student},t}) \in \mathbb{R}^M, \tag{10}$$

where $\mathrm{sg}(\cdot)$ denotes stop-gradient. In the implementation D_ϕ is a three-linear-layer MLP. Its cost is $O(M\,d\,h)$ and is negligible next to the fusion unit. Stop-gradient on h_t and k_i^t ensures that the distillation target shapes only the diagnoser itself and the conditional projections W_i', rather than perturbing the main fusion/prediction pathway through the diagnostic loss.

Finally, we calculate the **distillation loss**. Let $a_t^j \in \{0, 1\}$ indicate whether sample j is still active at step t (Section 3.4). The student mimics the (detached) teacher, averaged over active samples, modalities and the steps actually executed:

$$\mathcal{L}_{\text{diag}} = \frac{1}{T} \sum_{t=0}^{T-1} \frac{\sum_j a_t^j \sum_{i=1}^M \left(s_i^{\text{student},t,j} - s_i^{\text{teacher},t,j} \right)^2}{M \sum_j a_t^j + \varepsilon}. \tag{11}$$

During training, modality selection and the gate use the student scores $s_i^{\text{student},t}$, so the inference pathway is exactly the one that was optimized. The teacher is used only inside $\mathcal{L}_{\text{diag}}$. At inference, $s_i^t = s_i^{\text{student},t}$ is computed from Eq. 10 with no label.

Why not reconstruction? An alternative diagnoser could measure reconstruction error $\|m_i - \hat{m}_i(h_t)\|$, and reconstruction is indeed the workhorse of many robust-learning methods, which either reconstruct a missing/weak modality from the observed or fused representations (Liu & Zhan, 2024) or impose per-modality autoencoding/latent-reconstruction objectives to ensure information completeness (Xiong et al., 2026; Nguyen et al., 2025). However, this criterion evaluates whether h_t preserves all information of m_i, including noise, redundancy, and task-irrelevant variation. Our task-conditional diagnoser is both more aligned with the learning objective and more efficient: it estimates only the task-relevant residual $I(m_i; y \mid h_t)$, and naturally filters out task-irrelevant modalities that would otherwise trigger unnecessary re-fusion.

3.4 Selection and Adaptive Termination

Modality selection. At each step t, among samples that are still active we select every modality whose student score exceeds a threshold,

$$I_t = \{i : s_i^{\text{student},t} > \tau\}, \tag{12}$$

with $\tau \geq 0$ a hyperparameter. If an active sample has no modality above τ ($I_t = \varnothing$), we force $I_t = \{\arg\max_i s_i^{\text{student},t}\}$, i.e. it re-fuses its single most-deficient modality. This guarantees at least one visible key for cross-attention and one meaningful update per active step. True stopping is decided solely by termination below. The threshold rule thus yields sample-adaptive $|I_t|$: most samples select one modality, hard samples several. A top-k alternative is compared in the ablation.

Adaptive termination. A per-sample active indicator a_t is initialized to 1 and updated every step. A sample continues ($a_{t+1} = 1$) only if all of the following hold: (1) **Information insufficiency**: $\max_i s_i^{\text{student},t} > \epsilon$; (2) **Prediction not yet stable**: $d(\hat{y}_t, \hat{y}_{t-1}) > \eta$, where for classification d is the entropy $H(p_\psi(y \mid h_t))$ and for regression it is $|g_\psi(h_t) - g_\psi(h_{t-1})|$; (3) **Within budget**: $t + 1 < T_{\max}$, a hard upper bound.

During training we force $a_t = 1$ for the first $T_{\min}$ steps ($t < T_{\min}$), overriding conditions 1–2, so the student diagnoser receives distillation signal before it is allowed to terminate a sample.

3.5 TRAINING OBJECTIVE

Writing $\mathcal{B}$ for the training batch and $a_t^j \in \{0,1\}$ for the active indicator of sample j at step t, the total loss combines five terms (auxiliary terms are means over active samples and executed steps):

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{final}} + \alpha\,\mathcal{L}_{\text{step}} + \beta\,\mathcal{L}_{\text{budget}} + \gamma\,\mathcal{L}_{\text{mono}} + \delta\,\mathcal{L}_{\text{diag}}, \qquad \mathcal{L}_{\text{budget}} = \mathbb{E}_{j\in\mathcal{B}}[T^j]/T_{\max}. \tag{13}$$

- **Final loss**: $\mathcal{L}_{\text{final}} = \mathbb{E}_{j\in\mathcal{B}}\,\mathcal{L}(g_\psi(h_T^j), y^j)$, the task loss on the final representation.
- **Step-wise deep supervision**: a task loss on the state h_t before each update, averaged over active samples and the T executed steps, $\mathcal{L}_{\text{step}} = \frac{1}{T}\sum_{t=0}^{T-1} \frac{\sum_j a_t^j \mathcal{L}(g_\psi(h_t^j), y^j)}{\sum_j a_t^j + \varepsilon}$. It supplies gradient to the shared fusion unit at every iteration and stabilizes the recursive loop.
- **Computation budget**: $\mathcal{L}_{\text{budget}}$ penalizes the expected number of steps $\mathbb{E}_{j\in\mathcal{B}}[T^j]$ (normalized by $T_{\max}$), encouraging early termination (Banino et al., 2021).
- **Monotonicity constraint**: $\mathcal{L}_{\text{mono}} = \frac{1}{T-1}\sum_{t=1}^{T-1} \frac{\sum_j a_t^j [\mathcal{L}(g_\psi(h_t^j), y^j) - \mathcal{L}(g_\psi(h_{t-1}^j), y^j)]_+}{\sum_j a_t^j + \varepsilon}$, which penalizes any active step that increases the task loss, guarding against a poorly gated update degrading the representation.
- **Diagnoser distillation**: $\mathcal{L}_{\text{diag}}$ as defined in Eq. 11, training the label-free student to match the detached teacher. δ is its weight, which is set to 1.

The training procedure is given in Algorithm 1, and the inference procedure in Algorithm 2.

4 THEORETICAL ANALYSIS

We summarize the theoretical properties of ReFuse. All formal statements and proofs are deferred to Appendix A, and the equivalence between the dual-predictor loss difference and the conditional mutual information (CMI) is established in Appendix B. The analysis makes four points:

(1) **Re-fusion is non-destructive.** Under reasonable assumptions, each re-fusion step is invertible in h_t conditionally on the modalities it injects. We retain the modalities visited in an accumulated state $S_t = (h_t, \{m_i\}_{i\in J_t})$, for which the data-processing inequality gives $I(S_t; y) \le I(S_{t+1}; y) \le I(\mathbf{m}; y)$: iteration never erases task information and can add to it (Lemma 1, Theorem 1).

(2) **Termination is approximately sufficient, and weak modalities are not lost.** When the diagnoser stops, the task information not yet captured by h_T is bounded by $M\bar{\epsilon}$ and the excess Bayes risk by $O(\sqrt{M\bar{\epsilon}})$, where $\bar{\epsilon}$ aggregates the stopping threshold with the teacher's Bayes gap and the student-teacher distillation error (Theorem 2). Moreover, every informative modality is driven below $\bar{\epsilon}$ within finitely many steps (Corollary 1). This is the second-chance mechanism: a modality that is overwhelmed by a dominant modality at initial fusion is repeatedly re-selected until its conditional gain is exhausted, which is a recovery guarantee that a single-pass model does not possess.

(3) **Selection is a near-optimal greedy step.** Viewing the conditional gain as a set function $f_t(S) = I(\{m_i\}_{i\in S}; y \mid h_t)$, conditional submodularity (which holds under conditional independence) yields the standard $(1 - 1/e)$ single-step approximation. Our threshold rule is a label-free batch relaxation of this greedy choice (Theorem 3).

(4) **Widening fusion is cheaper than deepening the network.** Because one fusion operator is reused across iterations, adding steps adds no parameters and improves the capacity-per-parameter trade-off (Theorem 4). Under an equal parameter budget, stacking T independently parameterised layers loses a $\sqrt{T}$ factor of per-layer spectral capacity, or must spend $\sim T\times$ parameters to recover it (Corollary 2); this makes "widen fusion instead of deepen the network" precise.

Sequential information-bottleneck view. ReFuse admits a *sequential* information-bottleneck (IB) interpretation (Tishby et al., 2000; Mai et al., 2023b; Wu et al., 2025b): rather than compressing all modalities in a single pass, it greedily increases sufficiency by acquiring at step t only the modalities whose estimated conditional gain exceeds τ, paying a small local compression cost instead of one joint bottleneck. Under the information-preservation condition the per-step gains telescope into $I(y; h_T)$, so the required information is pulled through several local bottlenecks, avoiding the

Table 1: The results on MOSI and MOSEI. The results with † are obtained from original papers.

Model	Venue	MOSI					MOSEI				
		Acc7↑	Acc2↑	F1↑	MAE↓	Corr↑	Acc7↑	Acc2↑	F1↑	MAE↓	Corr↑
EMOE (Fang et al., 2025)	CVPR 2025	45.2	84.8	84.8	0.723	0.790	52.5	85.0	85.0	0.542	0.760
PSA-MF (Xie et al., 2026)	AAAI 2026	46.5	86.4	86.2	0.686	0.807	55.0	86.3	86.3	0.521	0.774
DecAlign (Qian et al., 2026)	ICLR 2026	40.9	86.7	86.3	0.749	0.790	52.8	87.0	86.8	0.534	0.791
C-MIB (Mai et al., 2023b)	TMM 2023	47.7	87.8	87.8	0.662	0.835	52.7	86.9	86.8	0.542	0.784
ITHP (Xiao et al., 2024)	ICLR 2024	47.7	88.5	88.5	0.663	**0.856**	52.2	87.1	87.1	0.550	0.792
Diffusion Bridge (Lee et al., 2025)	CVPR 2025	47.3	86.9	86.8	0.649	0.839	53.1	87.1	87.0	0.531	**0.800**
GSCon (Shi et al., 2025)	TIP 2025	45.7	88.1	88.0	0.696	0.832	50.8	87.4	87.4	0.561	0.750
MCIB (Wang et al., 2026)	AAAI 2026	46.7	87.3	87.3	0.721	0.821	47.9	86.4	86.4	0.593	0.792
OMIB (Wu et al., 2025b)	ICML 2025	47.6	88.9	88.8	0.636	0.844	53.5	87.3	87.3	0.530	0.789
ReFuse	-	**49.4**	**89.5**	**89.5**	**0.622**	0.845	**55.4**	**87.7**	**87.8**	**0.512**	0.792

regime in which a single-pass bottleneck lets a dominant modality suppress a weak one. The formal sequential objective and its relation to one-shot and multimodal IB are given in Appendix A.8.

5 Experiments

We evaluate ReFuse on multimodal sentiment analysis (MSA), multimodal humor detection (MHD), multimodal sarcasm detection (MSD), and audio-visual event recognition under modality imbalance. The datasets include MOSI (Zadeh et al., 2016), MOSEI (Zadeh et al., 2018), CH-SIMS-v2 (Liu et al., 2022), UR-FUNNY (Hasan et al., 2019), MUStARD (Castro et al., 2019), and AVE (Tian et al., 2018). We introduce experimental details in Appendix E and additional results in Appendix C.

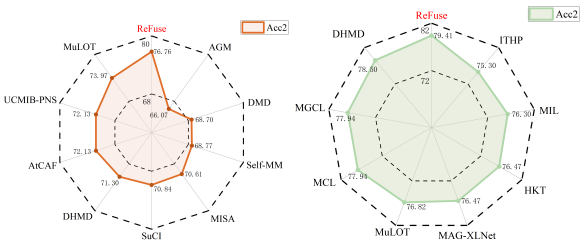
5.1 Main Results

Results on the MSA task. Table 1 reports results on MOSI and MOSEI, and Table 2 on CH-SIMS-v2. ReFuse sets a new state of the art on most of the metrics. On MOSI it improves Acc7 by 1.7 points over the strongest IB-based baselines C-MIB/ITHP, and improves Acc2/F1 by 0.6/0.7 over OMIB. On MOSEI it obtains the best results on Acc7, Acc2, F1 score and MAE. The gains are largest on CH-SIMS-v2, where ReFuse is best on all metrics (Acc5 +2.3 and Acc3 +3.4 over the runner-up). The reason is that one-shot fusion mixes every modality in a single pass: when one (language) modality dominates, the weak acoustic and visual cues are partially drowned out, and once lost they cannot be recovered by later layers. IB-based methods such as OMIB and C-MIB share this limitation. ReFuse instead detects which insufficiently-learned modality still carries task-useful information, and the loop injects exactly that modality, so the information is recovered rather than discarded. The larger gain on CH-SIMS-v2 is consistent with this view, as its acoustic and visual streams carry relatively more sentiment information and therefore benefit more from re-fusion.

Table 2: The results on CH-SIMS-v2.

Model	CH-SIMS v2					
	Acc5↑	Acc3↑	Acc2↑	F1↑	MAE↓	Corr↑
MAG-BERT (Rahman et al., 2020)	49.2	70.6	77.1	77.1	0.346	0.641
Self-MM (Yu et al., 2021)	53.5	72.7	78.7	78.6	0.315	0.691
MMIM (Han et al., 2021)	50.5	70.4	77.8	77.8	0.339	0.641
AV-MC (Liu et al., 2022)	52.1	73.2	80.6	80.7	**0.301**	**0.721**
DLF (Wang et al., 2025)	47.5	70.0	78.1	77.9	0.346	0.683
TMDC (Zhuang et al., 2026)	53.8	71.9	79.6	79.6	0.356	0.703
Diffusion Bridge (Lee et al., 2025)	52.5	70.7	78.6	79.0	0.323	0.677
MCIB (Wang et al., 2026)	47.7	72.3	80.2	78.6	0.345	0.712
ReFuse	**56.1**	**76.6**	**82.8**	**82.9**	**0.278**	**0.745**

Results on the MSD and MHD tasks. Figure 2 shows that the advantage transfers beyond sentiment. On UR-FUNNY (MHD task), ReFuse attains 76.76% accuracy, 2.79 points above the strongest baseline. On MUStARD (MSD task), it reaches 79.41%, 0.91 above the runner-up DHMD. These tasks have label semantics very different from sentiment, yet no task-specific change is made to the fusion mechanism. The consistent gain indicates that the diagnose-select-refuse loop is not tied to a particular objective: humor and sarcasm likewise hinge on non-verbal cues



(a) Results on MHD tasks. (b) Results on MSD tasks.

Figure 2: The results on UR-FUNNY and MUStARD.

Table 4: Equal fusion-parameter comparison on CMU-MOSI and CMU-MOSEI.

Dataset	Architecture	Total Parameters	Fusion Parameters	Acc-7↑	Acc-2↑	F1↑	MAE↓	Corr↑
MOSEI	Deep	186,836,352	577,845	54.2	87.0	87.0	0.517	0.798
	Wide	186,836,007	577,500	54.4	87.1	87.1	0.513	**0.801**
	ReFuse	186,826,672	568,165	**55.4**	**87.7**	**87.8**	**0.512**	0.792
MOSI	Deep	186,639, 326	394,875	47.5	87.5	87.4	0.670	0.830
	Wide	186,665,641	421,190	47.7	87.8	87.6	0.649	0.841
	ReFuse	186,658,632	414,182	**49.4**	**89.5**	**89.5**	**0.622**	**0.845**

that are easy to miss in a single fusion, and selectively re-fusing the deficient modality recovers them, providing evidence that the mechanism generalizes across multimodal tasks.

Results on the AVE dataset. Table 3 shows that ReFuse attains 72.57% accuracy, surpassing the strongest dedicated imbalance method InfoReg and ARL by about 2.7 points, the information-balanced state-of-the-art IBML by 2.4 point. Notably, every competing baseline explicitly rebalances optimization (gradient modulation, conflict resolution, or information balancing), whereas ReFuse never modifies the gradients or unimodal encoders. Its gain comes entirely from giving the weak visual modality repeated, demand-driven opportunities to contribute: the sufficiency diagnoser detects that the task-relevant visual information is still missing from h_t, and the loop selectively re-fuses the visual modality instead of letting it be overwhelmed in a single pass. The result on AVE supports the central claim that widening modality interaction helps weak modalities more than one-shot and optimization-rebalancing alternatives.

Table 3: Audio-visual event recognition accuracy (%) on the AVE dataset.

Method	Venue	Acc (%) ↑
OGM-GE (Peng et al., 2022)	CVPR 2022	65.67 ± 1.02
PMR (Fan et al., 2023)	CVPR 2023	66.09 ± 0.47
MMCosine (Xu et al., 2023)	ICASSP 2023	67.49 ± 0.47
AGM (Li et al., 2023a)	ICCV 2023	68.32 ± 0.94
InfoReg (Huang et al., 2025b)	CVPR 2025	69.89 ± 0.19
MMPareto (Wei & Hu, 2024)	ICML 2024	69.73 ± 0.62
MLA (Zhang et al., 2024a)	CVPR 2024	68.74 ± 0.47
DGL (Wei et al., 2025b)	ICCV 2025	68.82 ± 2.34
ARL (Wei et al., 2025a)	ICCV 2025	69.87 ± 0.44
IBML (Qin et al., 2026)	TPAMI 2026	$\underline{70.15 \pm 1.07}$
ReFuse	-	$\mathbf{72.57 \pm 0.21}$

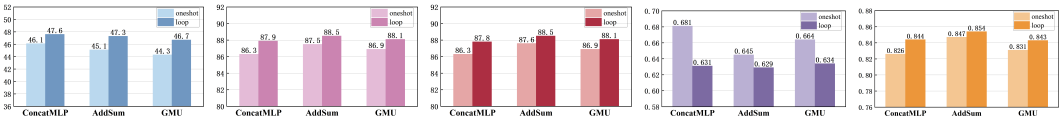
5.2 Equal-Parameter Comparison: Wider Fusion vs. Deeper Network

Table 4 compares the three ways of spending the same fusion-parameter budget. All three architectures use the same fusion-parameter budget P: **Deep** stacks T independent narrow Transformer blocks, **Wide** uses one wider Transformer block, and **ReFuse** uses one block recursively, driven by the sufficiency diagnoser. On MOSI, ReFuse is best on all five metrics, e.g. raising Acc7 from 47.7 to 49.4 and lowering MAE from 0.649 to 0.622 against the strongest baseline. On MOSEI, ReFuse is best on Acc7, Acc2, F1 and MAE, while Wide has a marginally higher correlation. Overall, allocating the budget to diagnoser-guided recursive re-fusion is more effective than either stacking independent narrow layers or using a single wider block, which supports our central claim of widening modality fusion rather than deepening the fusion network.

The three architectures spend the same fusion budget in different ways. Deep divides it among T independent blocks, which forces a narrow working width: the raw modalities are first squeezed through a small, shared projection, so information of weak modalities can be lost at this bottleneck before any of the T layers acts on it. Wide keeps a single, wider block but still fuses only once, so it has no second chance to recover a modality suppressed in that single pass. Both Deep and Wide apply one-shot fusion. ReFuse instead concentrates the budget into one full-capacity, parameter-shared block and closes a diagnoser-guided loop: it re-reads the uncompressed modality features, injects only the modalities that still carry task-useful information, and terminates per sample.

5.3 Backbone-Agnostic Validation: Gains Come from Iterative Fusion

A natural concern is whether ReFuse's gains simply arise from using a more expressive fusion operator (gated cross-attention) rather than from the iterative mechanism. To rule this out, we instantiate the fusion unit F_θ with weak operators, ranging from a plain additive sum (AddSum) to a concatenation MLP (ConcatMLP) and a gated multimodal unit (GMU). As shown in Figure 3, in every case,



(a) Acc7 (b) Acc2 (c) F1 (d) MAE (e) Corr

Figure 3: **Backbone-agnostic validation on MOSI.** For three different fusion operators (ConcatMLP, AddSum, GMU), we compare one-shot fusion with the ReFuse iterative loop.

wrapping the operator in our iterative diagnose-select-refuse loop improves all metrics over applying it once: for example, Acc7 rises by 1.5/2.2/2.4 points and MAE drops by 0.050/0.016/0.030 for ConcatMLP/AddSum/GMU, respectively. Since the only change is the loop while the operator itself is held fixed, the improvement must be attributed to the iterative, diagnoser-guided fusion mechanism rather than the expressiveness of F_θ, confirming the architecture-agnostic claim of Section 3.2.

5.4 Ablation Studies

Here we verify the contribution of each component: (1) **One-shot Fusion** uses only the initial fusion h_0, removing the recursive loop entirely. It tests whether iterative re-fusion is needed at all; (2) **No diagnoser (fixed T)** removes the teacher/student diagnosers and $\mathcal{L}_{\text{diag}}$, and fuses all modalities for a fixed number of steps. It tests the information-sufficiency diagnosis that drives the loop; (3) **No Modality Selection** keeps the diagnoser and adaptive stopping but re-fuses every modality at each step. It tests selective, modality-specific injection; (4) **Fixed Steps** keeps the diagnoser and selection but runs a fixed number of passes without early termination. It tests sufficiency-based adaptive stopping; (5) **Unshared Fusion Units** assigns an independent fusion unit to each step. It tests whether sharing one full-capacity block is preferable to deepening with separate units.

Table 5: Ablation studies on MOSI and MOSEI.

Model	MOSI			MOSEI		
	Acc7↑	**Acc2↑**	**MAE↓**	**Acc7↑**	**Acc2↑**	**MAE↓**
Unshared Fusion Units	47.3	89.3	0.626	53.2	87.0	0.526
Fixed Steps	46.4	87.9	0.640	54.7	86.2	0.516
No Modality Selection	46.4	87.0	0.645	53.6	87.1	0.521
No diagnoser (fixed T)	47.0	86.8	0.653	53.4	86.8	0.523
One-shot Fusion	45.2	86.7	0.655	54.4	86.4	0.518
ReFuse	**49.4**	**89.5**	**0.622**	**55.4**	**87.7**	**0.512**

Table 5 dissects ReFuse on MOSI and MOSEI. Removing any component degrades performance, with consistent trends across both datasets: (1) **Adaptive termination matters.** Fixing the number of fusion steps instead of early-stopping causes significant drop (Acc-7 -3.0 on MOSI; MAE rises to 0.640). This is because running the loop for every sample regardless of whether the representation is already sufficient wastes computation and, more importantly, re-injects noise from modalities that have already been fully captured; (2) **Iteration itself is essential.** Collapsing the loop to one-shot fusion costs 4.2 and 1.0 Acc7 points, confirming that a second, targeted pass over insufficiently-learned modalities recovers information lost in the single fusion; (3) **Selection and diagnosis are both needed.** Re-fusing all modalities at every step (no selection) lowers Acc7 by 3.0/1.8, and removing the sufficiency diagnoser while keeping a fixed loop lowers it by 2.4/2.0. The diagnoser directs computation toward the modalities that still carry task-useful information rather than blindly re-fusing everything; (4) Finally, using separate (unshared) fusion units at each step instead of one shared block reduces Acc7 by 2.1/2.2: **repeatedly applying a single full-capacity block is more effective than maintaining step-specific units**, which under a fixed budget must be narrower.

6 Conclusion

We proposed ReFuse, a recursive refinement fusion framework that replaces one-shot fusion with an iterative "diagnose-select-refuse" loop. ReFuse proposes an information sufficiency diagnoser to identify which modalities still carry task-relevant information not yet captured, driving both selective re-fusion and adaptive early termination. By sharing fusion parameters across recursion steps, ReFuse widens modality interaction without deepening the network. We also provide theoretical guarantees for ReFuse. Experiments on multiple benchmarks showed consistent improvements and equal-parameter comparison confirmed that our wider recursive fusion outperforms deeper fusion.

AI USE STATEMENT

In this paper, AI, specifically large language models (LLMs), were used solely for language polishing to improve clarity and fluency. All scientific content, including the experimental design, results, and conclusions, is the responsibility of the authors. No LLM was involved in the ideation, analysis, or interpretation of the research, and all LLM-assisted content was carefully reviewed and verified by the authors.

REPRODUCIBILITY STATEMENT

Code, pre-trained models, and configuration files will be released upon acceptance. All experiments use publicly available datasets and standard pre-extracted features.

REFERENCES

Hamidreza Alikhani, Anil Kanduri, Pasi Liljeberg, Amir M. Rahmani, and Nikil Dutt. DynaFuse: Dynamic fusion for resource efficient multi-modal machine learning inference. *IEEE Embedded Systems Letters*, 2023. doi: 10.1109/LES.2023.3298738.

John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A. González. Gated multimodal units for information fusion. In *ICLR Workshop*, 2017. arXiv:1702.01992.

Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pp. 59–66. IEEE, 2018.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2): 423–443, Feb 2019. ISSN 0162-8828. doi: 10.1109/TPAMI.2018.2798607.

Andrea Banino, Jan Balaguer, and Charles Blundell. PonderNet: Learning to ponder. *arXiv preprint arXiv:2107.05407*, 2021.

Peter Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. *Advances in neural information processing systems*, 30, 2017.

Lanxin Bi, Yunqi Zhang, Luyi Wang, Yake Niu, and Hui Zhao. Two challenges, one solution: Robust multimodal learning through dynamic modality recognition and enhancement. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 12855–12867, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-emnlp.689.

Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. Towards multimodal sarcasm detection (an _obviously_ perfect paper). In *ACL*, pp. 4619–4629, 2019.

Jili Chen, Yihua Zhong, Qionghao Huang, Changqin Huang, Fan Jiang, Xiaodi Huang, and Xun Wang. Ucmib-pns: Balancing sufficiency and necessity with probabilistic causality and cross-modal uncertainty in multimodal sentiment analysis. *IEEE Transactions on Affective Computing*, 17(1):92–106, 2026. doi: 10.1109/TAFFC.2025.3606964.

Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. Covarep: A collaborative voice analysis repository for speech technologies. In *ICASSP*, pp. 960–964, 2014.

J. Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019.

Siyi Du, Xinzhe Luo, Declan P. O'Regan, and Chen Qin. Inference-time dynamic modality selection for incomplete multimodal classification. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2601.22853.

Florian Eyben. Opensmile: the munich versatile and fast open-source audio feature extractor. In *ACM International Conference on Multimedia*, pp. 1459–1462, 2010.

Yi Fan, Wei Xu, Hao Wang, Jiajun Wang, and Song Guo. PMR: Prototypical modal rebalance for multimodal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20029–20038, 2023.

Yiyang Fang, Wenke Huang, Guancheng Wan, Kehua Su, and Mang Ye. Emoe: Modality-specific enhanced dynamic emotion experts. In *CVPR*, pp. 14314–14324, 2025.

Xinyu Feng, Yuming Lin, Lihua He, You Li, Liang Chang, and Ya Zhou. Knowledge-guided dynamic modality attention fusion framework for multimodal sentiment analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 14755–14766, 2024.

Dimitris Gkoumas, Qiuchi Li, C. Lioma, Yijun Yu, and Da wei Song. What makes the difference? an empirical comparison of fusion strategies for multimodal language analysis. *Information Fusion*, 66:184–197, 2021.

Wei Han, Hui Chen, and Soujanya Poria. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In *EMNLP*, pp. 9180–9192, November 2021.

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 *EMNLP*, pp. 2046–2056, 2019.

Md Kamrul Hasan, Sangwu Lee, Wasifur Rahman, Amir Zadeh, Rada Mihalcea, Louis-Philippe Morency, and Ehsan Hoque. Humor knowledge enriched transformer for understanding multimodal humor. In *AAAI*, volume 35, pp. 12972–12980, 2021.

Devamanyu Hazarika, R. Zimmermann, and Soujanya Poria. Misa: Modality-invariant and -specific representations for multimodal sentiment analysis. *ACM MM*, pp. 1122–1131, 2020.

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

Kang He, Yuzhe Ding, Xinrong Wang, Fei Li, Chong Teng, and Donghong Ji. Enhance-then-balance modality collaboration for robust multimodal sentiment analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. arXiv:2604.12518.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: decoding-enhanced bert with disentangled attention. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, 2021.

Changqin Huang, Jili Chen, Qionghao Huang, Shijin Wang, Yaxin Tu, and Xiaodi Huang. Atcaf: Attention-based causality-aware fusion network for multimodal sentiment analysis. *Information Fusion*, 114:102725, 2025a.

Chao Huang, Yake Wei, Zhiyu Yang, and Di Hu. Adaptive unimodal regulation for balanced multimodal information acquisition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25854–25863, 2025b.

Seulgi Kim, Mohit Prabhushankar, and Ghassan AlRegib. Information router for mitigating modality dominance in vision-language models. *arXiv preprint arXiv:2604.16264*, 2026.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*, 2020.

Jeong Ryong Lee, Yejee Shin, Geonhui Son, and Dosik Hwang. Diffusion bridge: Leveraging diffusion model to reduce the modality gap between text and vision for zero-shot image captioning. In *CVPR*, pp. 4050–4059, 2025.

Hong Li, Xingyu Li, Pengbo Hu, Yinuo Lei, Chunxiao Li, and Yi Zhou. Boosting multi-modal model performance with adaptive gradient modulation. In *ICCV*, pp. 22214–22224, 2023a.

Yong Li, Yuanzhi Wang, and Zhen Cui. Decoupled multimodal distilling for emotion recognition. In *CVPR*, pp. 6631–6640, 2023b.

Yong Li, Yuanzhi Wang, Yi Ding, Shiqing Zhang, Ke Lu, and Cuntai Guan. Decoupled hierarchical distillation for multimodal emotion recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Zheng Lian, Lan Chen, Licai Sun, Bin Liu, and Jianhua Tao. Gcnet: Graph completion network for incomplete multimodal learning in conversation. *IEEE Transactions on pattern analysis and machine intelligence*, 45(7):8419–8432, 2023.

Weide Liu and Huijing Zhan. Multimodal sentiment analysis with missing modality: A knowledge-transfer approach. *arXiv preprint arXiv:2401.10747*, 2024.

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 *ICMI*, pp. 247–258, 2022.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

Miaosen Luo, Yuncheng Jiang, and Sijie Mai. Towards explainable fusion and balanced learning in multimodal sentiment analysis. In *ACM MM*, pp. 1997–2006, 2025.

Sijie Mai, Ya Sun, Ying Zeng, and Haifeng Hu. Excavating multimodal correlation for representation learning. *Information Fusion*, 91:542–555, 2023a.

Sijie Mai, Ying Zeng, and Haifeng Hu. Multimodal information bottleneck: Learning minimal sufficient unimodal and multimodal representations. *IEEE Transactions on Multimedia*, 25:4121–4134, 2023b.

Sijie Mai, Ying Zeng, and Haifeng Hu. Learning from the global view: Supervised contrastive learning of multimodal representation. *Information Fusion*, 100:101920, 2023c.

Sijie Mai, Ying Zeng, Shuangjia Zheng, and Haifeng Hu. Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis. *IEEE Transactions on Affective Computing*, 14 (3):2276–2289, 2023d.

Sijie Mai, Ying Zeng, Aolin Xiong, and Haifeng Hu. Injecting multimodal information into pre-trained language model for multimodal sentiment analysis. *IEEE Transactions on Affective Computing*, 2025.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

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

Duy A. Nguyen, Abhi Kamboj, and Minh N. Do. Robult: Leveraging redundancy and modality-specific features for robust multimodal learning. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI)*, 2025.

Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. arXiv:2203.15332.

Shraman Pramanick, Aniket Roy, and Vishal M Patel. Multimodal learning using optimal transport for sarcasm and humor detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3930–3940, 2022.

Chengxuan Qian, Shuo Xing, Shawn Li, Yue Zhao, and Zhengzhong Tu. Decalign: Hierarchical cross-modal alignment for decoupled multimodal representation learning. In *ICLR*, 2026.

Yang Qin, Yanglin Feng, Yuan Sun, Dezhong Peng, Xi Peng, and Peng Hu. Deep information-balanced multimodal learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(8):9384–9396, 2026.

Chongyu Qu, Zhengyi Lu, Yuxiang Lai, Thomas Z. Li, Junchao Zhu, Junlin Guo, Juming Xiong, Yanfan Zhu, Yuechen Yang, Allen J. Luna, Kim L. Sandler, Bennett A. Landman, and Yuankai Huo. AdaFuse: Adaptive multimodal fusion for lung cancer risk prediction via reinforcement learning. *arXiv preprint arXiv:2602.00347*, 2026.

Wasifur Rahman, M. Hasan, Sangwu Lee, Amir Zadeh, Chengfeng Mao, Louis-Philippe Morency, and E. Hoque. Integrating multimodal information in large pretrained transformers. *ACL*, 2020: 2359–2369, 2020.

Christian Schlarmann, Francesco Croce, Nicolas Flammarion, and Matthias Hein. FuseLIP: Multimodal embeddings via early fusion of discrete tokens. *arXiv preprint arXiv:2506.03096*, 2025.

QingHongYa Shi, Mang Ye, Wenke Huang, Bo Du, and Xiaofen Zong. Gradient and structure consistency in multimodal emotion recognition. *IEEE Transactions on Image Processing*, 2025.

Yusuf Shihata. Gated recursive fusion: A stateful approach to scalable multimodal transformers. *arXiv preprint arXiv:2507.02985*, 2025.

Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 247–263, 2018.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *ACL*, pp. 6558–6569, July 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

Pan Wang, Qiang Zhou, Yawen Wu, Tianlong Chen, and Jingtong Hu. Dlf: Disentangled-language-focused multimodal sentiment analysis. In *AAAI*, volume 39, pp. 21180–21188, 2025.

Yihua Wang, Qi Jia, Cong Xu, Feiyu Chen, Yuhan Liu, Haotian Zhang, Liang Jin, Lu Liu, and Zhichun Wang. Conditional information bottleneck for multimodal fusion: Overcoming shortcut learning in sarcasm detection. In *AAAI*, volume 40, pp. 26652–26660, 2026.

Shicai Wei, Chunbo Luo, and Yang Luo. Improving multimodal learning via imbalanced learning. In *ICCV*, pp. 2250–2259. IEEE, 2025a.

Shuo Wei, Chaojie Luo, and Yong Luo. Boosting multimodal learning via disentangled gradient learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22879–22888, 2025b.

Yake Wei and Di Hu. MMPareto: Boosting multimodal learning with innocent unimodal assistance. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, pp. 52559–52572, 2024.

Yake Wei, Di Hu, Henghui Du, and Ji-Rong Wen. On-the-fly modulation for balanced multimodal learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(1):469–485, 2025c. doi: 10.1109/TPAMI.2024.3468315.

Han Wu, Yanming Sun, Yunhe Yang, and Derek F. Wong. Beyond simple fusion: Adaptive gated fusion for robust multimodal sentiment analysis. *arXiv preprint arXiv:2510.01677*, 2025a.

Qilong Wu, Yiyang Shao, Jun Wang, and Xiaobo Sun. Learning optimal multimodal information bottleneck representations. In *International Conference on Machine Learning*, pp. 67584–67622. PMLR, 2025b.

Xiongye Xiao, Gengshuo Liu, Gaurav Gupta, Defu Cao, Shixuan Li, Yaxing Li, Tianqing Fang, Mingxi Cheng, and Paul Bogdan. Neuro-inspired information-theoretic hierarchical perception for multimodal learning. In *ICLR*, 2024.

Heng Xie, Kang Zhu, Zhengqi Wen, Jianhua Tao, Xuefei Liu, Ruibo Fu, and Changsheng Li. Psamf: Personality-sentiment aligned multi-level fusion for multimodal sentiment analysis. In *AAAI*, 2026.

Jiayu Xiong, Jing Wang, Jun Xue, Wanlong Wang, Jianlong Kwan, Xiaosen Lyu, and Zhouqiang Jiang. Multimodal fusion via self-consistent task-gradient fields. In *International Conference on Machine Learning (ICML)*, 2026.

Ruihan Xu, Yuting Gao, Lan Wang, Jianing Li, Weihao Chen, Qingpei Guo, Ming Yang, and Shiliang Zhang. Looping back to move forward: Recursive transformers for efficient and flexible large multimodal models. *arXiv preprint arXiv:2602.09080*, 2026.

Ruoxuan Xu, Ruyang Feng, Shi-Xu Zhang, and Di Hu. MMCosine: Multi-modal cosine loss towards balanced audio-visual fine-grained learning. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023.

Zhi Xu, Dingkang Yang, Mingcheng Li, Yuzheng Wang, Zhaoyu Chen, Jiawei Chen, Jinjie Wei, and Lihua Zhang. Debiased multimodal understanding for human language sequences. In *AAAI*, volume 39, pp. 14450–14458, 2025.

Rui Yang, Lin Song, Yicheng Xiao, Runhui Huang, Yixiao Ge, Ying Shan, and Hengshuang Zhao. HaploVL: A single-transformer baseline for multi-modal understanding. *arXiv preprint arXiv:2503.14694*, 2025.

Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *AAAI*, volume 35, pp. 10790–10797, 2021.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis Philippe Morency. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6): 82–88, 11 2016.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis Philippe Morency. Tensor fusion network for multimodal sentiment analysis. In *EMNLP*, pp. 1114–1125, 2017.

Amir Zadeh, Paul Pu Liang, Jonathan Vanbriesen, Soujanya Poria, Edmund Tong, Erik Cambria, Minghai Chen, and Louis Philippe Morency. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *ACL*, pp. 2236–2246, 2018.

Xinyi Zhang, Jae Yoon, Mohit Bansal, and Huaxiu Yao. Multimodal representation learning by alternating unimodal adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27456–27466, 2024a.

Yazhou Zhang, Yang Yu, Dongming Zhao, Zuhe Li, Bo Wang, Yuexian Hou, Prayag Tiwari, and Jing Qin. Learning multitask commonness and uniqueness for multimodal sarcasm detection and sentiment analysis in conversation. *IEEE Transactions on Artificial Intelligence*, 5(3):1349–1361, 2024b. doi: 10.1109/TAI.2023.3298328.

Yiming Zhang, Qiangyu Yan, Borui Jiang, and Kai Han. Multimodal latent reasoning via hierarchical visual cues injection. *arXiv preprint arXiv:2602.05359*, 2026.

Peida Zhou, Xiaoyong Sun, Bei Sun, Runze Guo, Zhaoyang Dang, and Shaojing Su. SDRFPT-Net: A spectral dual-stream recursive fusion network for multispectral object detection. *Preprints.org*, 2025. URL `https://www.preprints.org/manuscript/202505.0884`. Preprint 202505.0884, not peer-reviewed.

Yan Zhuang, Minhao Liu, Yanru Zhang, Jiawen Deng, and Fuji Ren. Tmdc: A two-stage modality denoising and complementation framework for multimodal sentiment analysis with missing and noisy modalities. In *AAAI*, volume 40, pp. 2281–2289, 2026.

Limitations. ReFuse's student diagnoser is trained via distillation from the label-dependent teacher signal, which means its accuracy depends on the quality of the teacher signal early in training when the prediction head is poorly calibrated. This can be mitigated by warm-starting the prediction head, a curriculum that gradually increases the distillation weight δ, or the training warm-up $T_{\min}$ that keeps every sample active for its first few steps. Additionally, the thresholds τ, ϵ, and η are dataset-dependent; an adaptive threshold based on validation statistics could improve generality. The student diagnoser's input includes label-free prediction-shift features, which require M additional (batched) forward passes per step; while these are lightweight, the overhead grows linearly with the number of modalities.

Broader impact. By improving weak modality capture under imbalance, ReFuse can make multimodal systems more robust to low-quality sensors and reduce the dominance of any single modality, which is relevant for fair and inclusive AI systems.

More broadly, ReFuse reframes the central design question of multimodal fusion: instead of asking "how do we design a better one-shot fusion function?", it asks "how do we make fusion iterative and adaptive, so that every modality is fused until it has nothing task-relevant left to contribute?" We hope this perspective, i.e., widening the fusion process rather than deepening the network, inspires a shift from designing ever more elaborate single-pass operators to building closed-loop, self-diagnosing fusion systems.

A Formal Statements and Proofs of Theoretical Results

A.1 Proof of Lemma 1 (Information Preservation)

Lemma 1 (Task-Information Preservation of a Gated Residual Fusion Step)**.** *Let $J_0 = \varnothing$ and $J_t := \bigcup_{s=0}^{t-1} I_s$ be the modalities re-fused before step t, and define the* accumulated state

$$S_t := \big(h_t,\, \{m_i\}_{i\in J_t}\big), \qquad S_0 = h_0, \tag{14}$$

which augments the current fused representation with the (uncompressed) modality vectors already visited. Write $M_t := \{m_i\}_{i\in J_t}$ (modalities already accumulated) and $B_t := \{m_i\}_{i\in I_t}$ (modalities injected at this step). One fusion step of Eq. 5 consists of two sub-units, each pairing a gated residual map with a LayerNorm:

$$\begin{aligned} u_t &= \mathrm{LN}_1\big(h_t + R_1(h_t)\big), & R_1(h) &= g_t \odot \mathrm{MultiHeadAttn}(h, \{m_i\}_{i\in I_t}), \\ h_{t+1} &= \mathrm{LN}_2\big(u_t + R_2(u_t)\big), & R_2(u) &= \mathrm{FFN}(u). \end{aligned} \tag{15}$$

We make two assumptions.

(A1) Contractive residual branches. *Every residual branch is a contraction in the state variable on the relevant domain, $\|J_{R_k}(z)\|_{\mathrm{op}} \le q < 1$, enforceable by spectral normalisation (the sigmoid gate $g_t \in [0,1]^d$ already shrinks the attention branch).*

(A2) Normalisation directions carry no task information. *Write any $x \in \mathbb{R}^d$ as $x = a\,\pi_0(x) + b\,\mathbf{1}$, where $b = \frac{1}{d}\sum_j x_j$ is the global location, $a = \sqrt{\frac{1}{d}\sum_j (x_j - b)^2}$ the global scale, and $\pi_0(x) = (x - b\mathbf{1})/a$ the normalised "shape". Assume that, at each LayerNorm and conditionally on* any *subset $C \subseteq \mathbf{m}$ of modalities, the removed location/scale coordinates contain no label information once the shape is known: $(a, b) \perp y \mid \pi_0(x), C$.*

Under (A1)-(A2)*, the accumulated state is task-information non-decreasing, with the exact per-step gain*

$$I(S_{t+1}; y) - I(S_t; y) = I(B_t; y \mid h_t, M_t) \ge 0, \tag{16}$$

hence $I(S_t; y) \le I(S_{t+1}; y)$.

Proof. Fix the input modalities; all maps below act on the state variable.

Step 1: Residual maps are bijective once the injected modalities are held fixed. Define the pre-normalisation residual maps

$$G_1(h) = h + R_1(h), \qquad G_2(u) = u + R_2(u). \tag{17}$$

The attention branch R_1 reads the injected modalities B_t as its keys/values, so the operative map is $G_1(\,\cdot\,; B_t)$. Under (A1), the invertible-residual-network theorem states that $z \mapsto z + R_k(z)$ is a bijection whose inverse is the convergent Neumann series $\sum_{\ell \geq 0} (-R_k)^{\circ \ell}$. Hence, *for fixed* B_t, G_1 is bijective; $G_2 = u + \mathrm{FFN}(u)$ does not depend on any modality and is bijective outright. (The gate $g_t \neq 0$ alone is neither necessary nor sufficient: contractivity is the operative condition.) This bijectivity is conditional on the injected modalities, which is precisely why they are kept in the accumulated state rather than discarded.

Step 2: Under (A2), LayerNorm preserves task information. For fixed affine parameters (γ, β), LayerNorm computes

$$\mathrm{LN}(x) = \gamma \odot \pi_0(x) + \beta, \tag{18}$$

so $\mathrm{LN}(x)$ is in one-to-one correspondence with the shape $\pi_0(x)$. The decomposition $x = a\,\pi_0(x) + b\,\mathbf{1}$ is a bijection between x and $(\pi_0(x), a, b)$. Assumption (A2), with C being the relevant conditioning subset, gives, conditionally on C,

$$\begin{aligned} I(x; y \mid C) = I(\pi_0(x), a, b; y \mid C) &= I(\pi_0(x); y \mid C) + \underbrace{I(a, b; y \mid \pi_0(x), C)}_{=0} \\ &= I(\pi_0(x); y \mid C) = I(\mathrm{LN}(x); y \mid C). \end{aligned} \tag{19}$$

Thus LayerNorm removes only directions that are independent of the label and consequently loses no task information. (LayerNorm itself is not injective; this is why bijection is asserted only for G_1, G_2, while LayerNorm is handled by (A2).)

Step 3: One step preserves task information conditionally. For the first sub-unit, $u_t = \mathrm{LN}_1(G_1(h_t; B_t))$: G_1 is bijective for fixed B_t (Step 1) and LN_1 preserves task information (Step 2), so

$$I(u_t; y \mid M_t, B_t) = I(h_t; y \mid M_t, B_t). \tag{20}$$

For the second sub-unit, $h_{t+1} = \mathrm{LN}_2(G_2(u_t))$: G_2 is bijective and LN_2 preserves task information, hence

$$I(h_{t+1}; y \mid M_t, B_t) = I(u_t; y \mid M_t, B_t) = I(h_t; y \mid M_t, B_t). \tag{21}$$

Step 4: Accumulated state is non-decreasing. Using Eq. 21 and the chain rule,

$$\begin{aligned} I(S_{t+1}; y) = I(h_{t+1}, M_t, B_t; y) &= I(M_t, B_t; y) + I(h_{t+1}; y \mid M_t, B_t) \\ &= I(M_t, B_t; y) + I(h_t; y \mid M_t, B_t) = I(h_t, M_t, B_t; y) \\ &= I(h_t, M_t; y) + I(B_t; y \mid h_t, M_t) \\ &= I(S_t; y) + I(B_t; y \mid h_t, M_t). \end{aligned} \tag{22}$$

The last conditional mutual information is non-negative, which is Eq. 16; therefore $I(S_t; y) \leq I(S_{t+1}; y)$. □

A.2 Proof of Lemma 2 (CMI Chain Rule and Subadditivity)

Lemma 2 (Conditional Mutual Information Chain Rule, Subadditivity and Decomposition)**.** *For any random variables $A_1, \ldots, A_M, B, C$, the chain rule is always an identity:*

$$I(A_1, \ldots, A_M; B \mid C) = I(A_1; B \mid C) + \sum_{k=2}^{M} I(A_k; B \mid C, A_1, \ldots, A_{k-1}). \tag{23}$$

Conditioning does not *decrease mutual information in general The subadditivity bound*

$$I(A_1, \ldots, A_M; B \mid C) \leq \sum_{k=1}^{M} I(A_k; B \mid C) \tag{24}$$

holds under conditional submodularity*, namely $I(A_k; B \mid C, A_{1:k-1}) \leq I(A_k; B \mid C)$ for all k; a simple sufficient condition is conditional independence $A_k \perp (A_1, \ldots, A_{k-1}) \mid (B, C)$. Equivalently, defining the synergy residual*

$$\Xi_{B|C}(A_{1:M}) := \sum_{k=1}^{M} I(A_k; B \mid C) - I(A_1, \ldots, A_M; B \mid C), \tag{25}$$

Eq. 24 is exactly $\Xi_{B|C} \geq 0$; without this assumption one has the exact identity $I(A_{1:M}; B \mid C) = \sum_k I(A_k; B \mid C) - \Xi_{B|C}$. Finally, whenever C is a deterministic function of $(A_1, \ldots, A_M)$,

$$I(A_1, \ldots, A_M; B) = I(C; B) + I(A_1, \ldots, A_M; B \mid C). \tag{26}$$

Proof. **Proof of Eq. 23.** By the definition of conditional mutual information,

$$I(A_1, \ldots, A_M; B \mid C) = H(B \mid C) - H(B \mid C, A_1, \ldots, A_M). \tag{27}$$

We decompose the conditional entropy using the chain rule:

$$H(B \mid C) - H(B \mid C, A_1, \ldots, A_M) = \sum_{k=1}^{M} \big[H(B \mid C, A_1, \ldots, A_{k-1}) - H(B \mid C, A_1, \ldots, A_k)\big] \tag{28}$$

$$= \sum_{k=1}^{M} I(A_k; B \mid C, A_1, \ldots, A_{k-1}) \tag{29}$$

$$= I(A_1; B \mid C) + \sum_{k=2}^{M} I(A_k; B \mid C, A_1, \ldots, A_{k-1}), \tag{30}$$

which is Eq. 23.

Proof of Eq. 24. Unlike the chain rule, conditioning does *not necessarily* decrease mutual information in general: $I(A_k; B \mid C, A_{1:k-1})$ may exceed $I(A_k; B \mid C)$ under synergistic (XOR-type) dependence. Eq. 24 therefore requires *conditional submodularity*, namely

$$I(A_k; B \mid C, A_1, \ldots, A_{k-1}) \leq I(A_k; B \mid C) \quad \forall k, \tag{31}$$

a sufficient condition for which is conditional independence $A_k \perp (A_1, \ldots, A_{k-1}) \mid (B, C)$. Summing these inequalities and using Eq. 23 gives Eq. 24. Defining $\Xi_{B|C} := \sum_k I(A_k; B \mid C) - I(A_1, \ldots, A_M; B \mid C)$, conditional submodularity is exactly $\Xi_{B|C} \geq 0$; without it the exact identity $I(A_1, \ldots, A_M; B \mid C) = \sum_k I(A_k; B \mid C) - \Xi_{B|C}$ still holds and reduces to Eq. 24 iff $\Xi_{B|C} \geq 0$.

Proof of Eq. 26. The chain rule always gives

$$I(C, A_1, \ldots, A_M; B) = I(C; B) + I(A_1, \ldots, A_M; B \mid C). \tag{32}$$

If C is a deterministic function of $(A_1, \ldots, A_M)$, then $(C, A_1, \ldots, A_M)$ and $(A_1, \ldots, A_M)$ are informationally equivalent, so $I(C, A_1, \ldots, A_M; B) = I(A_1, \ldots, A_M; B)$; substituting yields Eq. 26. (In our application $C = h_T$ is a deterministic function of $A_{1:M} = \mathbf{m}$, which is why the identity applies.)

□

A.3 Proof of Theorem 1 (Information Monotonicity)

Theorem 1 (Information Monotonicity)**.** *Under the conditions of Lemma 1, let S_t be the accumulated state of Eq. 14. Then, for every recursion step $t = 0, 1, \ldots, T-1$,*

$$I(S_0; y) \leq I(S_t; y) \leq I(S_{t+1}; y) \leq I(m_1, \ldots, m_M; y), \tag{33}$$

where $S_0 = h_0$. Moreover h_t is a deterministic function of S_t, so $I(h_t; y) \leq I(S_t; y)$; monotonicity is asserted for the accumulated state S_t rather than for h_t alone, since the latter need not hold once the visited modalities are discarded.

Proof. We prove the two inequalities in Eq. 33.

Non-decrease: $I(S_t; y) \leq I(S_{t+1}; y)$. This is exactly Eq. 16 of Lemma 1: under (A1) the residual maps are bijective once the injected modalities are held fixed, and under (A2) each LayerNorm removes only location/scale directions independent of y. Iterating Eq. 16 also gives $I(S_0; y) \leq I(S_t; y)$.

Upper bound: $I(S_{t+1}; y) \leq I(\mathbf{m}; y)$. By construction every f_0 and F_θ is deterministic: $h_0 = f_0(\mathbf{m})$ and $h_{t+1} = F_\theta(h_t, \{m_i\}_{i \in I_t})$, so by induction h_{t+1} is a measurable function of $\mathbf{m}$; the accumulated modalities M_{t+1} are a subset of $\mathbf{m}$. Hence S_{t+1} is a deterministic function of $\mathbf{m}$,

the Markov chain $y \to \mathbf{m} \to S_{t+1}$ holds, and the data processing inequality gives $I(S_{t+1}; y) \leq I(\mathbf{m}; y)$.

Combining the two parts yields Eq. 33. □

A.4 Proof of Theorem 2 (Approximate Sufficiency)

Theorem 2 (Approximate Sufficiency)**.** *The implemented termination rule is expressed in the* learned *student score, not in the ideal conditional mutual information:*

$$\max_{1 \leq i \leq M} s_i^{\textit{student},T} \leq \epsilon. \tag{34}$$

Let ϵ_{approx} bound the teacher's Bayes gap (Proposition in Appendix B). Assume that the student–teacher distillation error is uniformly bounded: there exists $\epsilon_{\text{distill}}$ such that, for all (i, t),

$$|s_i^{\text{student},t} - s_i^{\text{teacher},t}| \leq \epsilon_{\text{distill}}.$$

Then, uniformly in (i, t),

$$I(m_i; y \mid h_t) \leq [s_i^{\textit{student},t}]_+ + \epsilon_{\text{approx}} + \epsilon_{\text{distill}}, \tag{35}$$

where $[\cdot]_+ = \max(0, \cdot)$ accounts for negative scores. Set $\bar{\epsilon} := \epsilon + \epsilon_{\text{approx}} + \epsilon_{\text{distill}}$. Because h_T is a deterministic function of $\mathbf{m}$, Eq. 26 gives $I(\mathbf{m}; y) - I(h_T; y) = I(\mathbf{m}; y \mid h_T)$. Under the conditional-submodularity condition of Lemma 2 ($\Xi_{y|h_T} \geq 0$),

$$0 \leq I(\mathbf{m}; y) - I(h_T; y) = I(\mathbf{m}; y \mid h_T) \leq \sum_{i=1}^{M} I(m_i; y \mid h_T) \leq M\bar{\epsilon}. \tag{36}$$

Without conditional submodularity the exact statement carries an additive synergy term, $I(\mathbf{m}; y) - I(h_T; y) \leq M\bar{\epsilon} + [-\Xi_{y|h_T}]_+$. For 0-1 loss, Pinsker's inequality and Jensen give the excess Bayes risk

$$R^*(h_T) - R^*(\mathbf{m}) \leq \sqrt{2\, I(\mathbf{m}; y \mid h_T)} \leq \sqrt{2M\bar{\epsilon}}, \tag{37}$$

(nats); for a general loss bounded in $[0, 1]$ the same $O(\sqrt{M\bar{\epsilon}})$ rate holds with a loss-dependent constant. Here $R^*(Z) = \inf_f \mathbb{E}[\ell(f(Z), y)]$ is the Bayes risk using feature Z.

Proof. **From learned scores to conditional mutual information.** The implemented stopping rule uses student scores, $\max_i s_i^{\text{student},T} \leq \epsilon$ (Eq. 34), not the ideal CMI. The Proposition in Appendix B bounds the teacher's Bayes gap by ϵ_{approx}, i.e. in expectation $s_i^{\text{teacher},T} \geq I(m_i; y \mid h_T) - \epsilon_{\text{approx}}$, and under the uniform distillation-error assumption we have $|s_i^{\text{student},T} - s_i^{\text{teacher},T}| \leq \epsilon_{\text{distill}}$ uniformly over (i, T). Combining,

$$I(m_i; y \mid h_T) \leq [s_i^{\text{student},T}]_+ + \epsilon_{\text{approx}} + \epsilon_{\text{distill}} \leq \bar{\epsilon}, \qquad \bar{\epsilon} := \epsilon + \epsilon_{\text{approx}} + \epsilon_{\text{distill}}. \tag{38}$$

Proof of Eq. 36. Since h_T is a deterministic function of $\mathbf{m}$, Eq. 26 with $C = h_T$, $A_i = m_i$, $B = y$ gives

$$I(\mathbf{m}; y) - I(h_T; y) = I(\mathbf{m}; y \mid h_T). \tag{39}$$

Under the conditional-submodularity condition of Lemma 2, Eq. 24 and Eq. 38 give

$$I(\mathbf{m}; y \mid h_T) \leq \sum_{i=1}^{M} I(m_i; y \mid h_T) \leq M\bar{\epsilon}. \tag{40}$$

Non-negativity of the gap follows from data processing (h_T is a function of $\mathbf{m}$), completing Eq. 36. Without conditional submodularity the exact identity $I(\mathbf{m}; y \mid h_T) = \sum_i I(m_i; y \mid h_T) - \Xi_{y|h_T}$ instead yields $I(\mathbf{m}; y) - I(h_T; y) \leq M\bar{\epsilon} + [-\Xi_{y|h_T}]_+$.

Proof of Eq. 37. Let $p^*(y \mid \mathbf{m})$ and $p^*(y \mid h_T)$ be the Bayes-optimal conditionals. For 0-1 loss, the excess Bayes risk of predicting from h_T instead of $\mathbf{m}$ is bounded by twice the averaged total-variation distance,

$$R^*(h_T) - R^*(\mathbf{m}) \leq 2\, \mathbb{E}_{h_T} \| p^*(y \mid \mathbf{m}) - p^*(y \mid h_T) \|_{\text{TV}}. \tag{41}$$

Pinsker gives $\mathrm{TV}(p,q) \le \sqrt{\frac{1}{2} D_{\mathrm{KL}}(p\|q)}$; Jensen and $I(\mathbf{m}; y \mid h_T) = \mathbb{E}_{h_T} D_{\mathrm{KL}}(p^*(y \mid \mathbf{m}) \| p^*(y \mid h_T))$ (all logarithms natural, hence nats) then yield

$$R^*(h_T) - R^*(\mathbf{m}) \le 2\sqrt{\tfrac{1}{2} I(\mathbf{m}; y \mid h_T)} = \sqrt{2\, I(\mathbf{m}; y \mid h_T)} \le \sqrt{2M\bar{\epsilon}}, \tag{42}$$

which is Eq. 37. For a general loss bounded in $[0,1]$ the same argument gives the $O(\sqrt{M\bar{\epsilon}})$ rate with a loss-dependent constant.

□

A.5 Proof of Corollary 1 (Weak Modality Capture)

Corollary 1 (Weak Modality Information Capture)**.** *For any modality m_i with $I(m_i; y) > 0$, define its capture residual relative to the accumulated state,*

$$r_i(t) := I(m_i; y \mid S_t) = I(m_i; y \mid h_t, M_t). \tag{43}$$

Then:

1. ***Monotone capture.*** *If $i \in I_t$, then m_i is a deterministic component of S_{t+1} and $r_i(t+1) = 0 \le r_i(t)$; if $i \notin I_t$, then under the conditional-submodularity condition of Lemma 2 (modalities' marginal gains diminish as the conditioning state grows), $r_i(t+1) \le r_i(t)$.*

2. ***Finite-time capture.*** *Assume the same uniform distillation-error bound as in Theorem 2: $|s_i^{\mathrm{student},t} - s_i^{\mathrm{teacher},t}| \le \epsilon_{\mathrm{distill}}$ for all (i,t). Set $\bar{\epsilon} = \epsilon + \epsilon_{\mathrm{approx}} + \epsilon_{\mathrm{distill}}$. If the selection threshold satisfies $\tau < \epsilon$, then there exists a finite T_i such that $r_i(t) < \bar{\epsilon}$ for all $t \ge T_i$; in fact a modality is captured (its residual becomes 0) at the first step at which it is selected while $r_i \ge \bar{\epsilon}$.*

Proof. **Proof of (1).** If $i \in I_t$, then $J_{t+1} = J_t \cup I_t$ contains i, so m_i is one of the uncompressed modality vectors stored in S_{t+1}; conditioned on S_{t+1} the variable m_i is deterministic, hence

$$r_i(t+1) = I(m_i; y \mid S_{t+1}) = 0 \le I(m_i; y \mid S_t) = r_i(t). \tag{44}$$

If $i \notin I_t$, then $S_{t+1} = (S_t$ information plus $B_t)$. Conditional submodularity (Lemma 2) states that a modality's marginal information does not increase as the background grows, i.e.

$$r_i(t+1) = I(m_i; y \mid S_t, B_t) \le I(m_i; y \mid S_t) = r_i(t). \tag{45}$$

This proves (1).

Proof of (2). The sequence $\{r_i(t)\}$ is non-increasing and bounded below by 0, hence it converges to some $r_i^* \ge 0$. The score bridge (Appendix B together with distillation) gives, uniformly in (i,t),

$$s_i^{\mathrm{student},t} \ge I(m_i; y \mid h_t) - \epsilon_{\mathrm{approx}} - \epsilon_{\mathrm{distill}}, \tag{46}$$

and conditional submodularity gives $r_i(t) = I(m_i; y \mid h_t, M_t) \le I(m_i; y \mid h_t)$. Assume for contradiction that $r_i^* \ge \bar{\epsilon}$. Then for all sufficiently large t, $I(m_i; y \mid h_t) \ge r_i(t) \ge \bar{\epsilon}$, and therefore

$$s_i^{\mathrm{student},t} \ge \bar{\epsilon} - \epsilon_{\mathrm{approx}} - \epsilon_{\mathrm{distill}} = \epsilon > \tau, \tag{47}$$

so $i \in I_t$ is selected at every such step; by part (1) this forces $r_i(t+1) = 0$, contradicting $r_i^* \ge \bar{\epsilon}$. Hence $r_i^* < \bar{\epsilon}$, and a finite T_i exists: the first step at which i is selected while $r_i \ge \bar{\epsilon}$ drives its residual to 0. □

A.6 Proof of Theorem 3 (Greedy Selection)

Theorem 3 (Single-Step Greedy Selection Approximation)**.** *Conditioned on the current state, define the set function $f_t : 2^{\{1,\dots,M\}} \to \mathbb{R}_{\ge 0}$*

$$f_t(S) = I(\{m_i\}_{i \in S}; y \mid h_t), \qquad S \subseteq \{1, \dots, M\}. \tag{48}$$

If f_t is monotone and submodular, i.e., for all $S \subseteq T$ and all $i \notin T$,

$$f_t(S \cup \{i\}) - f_t(S) \ge f_t(T \cup \{i\}) - f_t(T) \ge 0, \tag{49}$$

then the greedy rule that builds a k-element subset within this step *by* $i_k = \arg\max_{i \notin S_{k-1}} [f_t(S_{k-1} \cup \{i\}) - f_t(S_{k-1})]$ *satisfies*

$$f_t(S_k) \geq \left(1 - \frac{1}{e}\right) f_t(S_k^*), \tag{50}$$

where $S_k^* = \arg\max_{|S|=k} f_t(S)$ is the optimal k-element subset at state h_t.

Remark 1 (When is f_t submodular?)**.** *Under conditional independence $m_i \perp m_j \mid (y, h_t)$ for $i \neq j$ (the conditional naive-Bayes case), the modalities are independent once the label is known but are still correlated when only h_t is given, because they share y as a common cause. Consequently the joint conditional information is sub-additive, $I(\{m_i\}_{i\in S}; y \mid h_t) \leq \sum_{i\in S} I(m_i; y \mid h_t)$, and the marginal gains diminish as S grows: f_t is* submodular*, so the $(1 - 1/e)$ guarantee of Theorem 3 applies, but the greedy choice is not exactly optimal. f_t becomes* modular*—and the greedy rule within a step exact—only when the modalities contribute independent conditional information, i.e. $I(\{m_i\}_{i\in S}; y \mid h_t) = \sum_{i\in S} I(m_i; y \mid h_t)$. A sufficient condition is that the modalities are mutually conditionally independent both given h_t and given (y, h_t). Redundancy among modalities is the more common regime and likewise yields diminishing conditional marginal returns, hence (approximate) submodularity. Conversely, synergistic interactions can violate submodularity; this is captured exactly by the synergy term $\Xi_{y|h_t}$ of Lemma 2.*

Proof. We prove the Nemhauser–Wolsey–Fisher bound *within a single step*. Fix the current state h_t and suppress it, writing $f = f_t$. Every marginal below is therefore conditional on h_t. Let $S_0 = \varnothing$, let S_j be the set after j greedy picks in this step, and let $\mathrm{OPT}_k = \max_{|S|=k} f(S) = f(S_k^*)$. Write $\Delta(i \mid S) = f(S \cup \{i\}) - f(S)$.

Key marginal-gain inequality. For $j = 1, \ldots, k$,

$$f(S_j) - f(S_{j-1}) \geq \frac{\mathrm{OPT}_k - f(S_{j-1})}{k}. \tag{51}$$

To see this, add the k elements of S_k^* to S_{j-1} one at a time. Monotonicity gives $f(S_{j-1} \cup S_k^*) \geq f(S_k^*) = \mathrm{OPT}_k$; submodularity implies that each element's marginal on top of S_{j-1} is at least its marginal when added later in the sequence, so the sum of these k marginals is at least $\mathrm{OPT}_k - f(S_{j-1})$. At least one of them—and therefore the greedy maximum $f(S_j) - f(S_{j-1})$—is at least their average, proving Eq. 51.

Induction. Set $D_j = \mathrm{OPT}_k - f(S_j)$. Eq. 51 gives

$$D_j = \mathrm{OPT}_k - f(S_{j-1}) - [f(S_j) - f(S_{j-1})] \leq \left(1 - \frac{1}{k}\right)(\mathrm{OPT}_k - f(S_{j-1})) = \left(1 - \frac{1}{k}\right) D_{j-1}. \tag{52}$$

Iterating for $j = 1, \ldots, k$ with $D_0 = \mathrm{OPT}_k$,

$$D_k \leq \left(1 - \frac{1}{k}\right)^k \mathrm{OPT}_k \leq e^{-1} \mathrm{OPT}_k, \tag{53}$$

hence $f(S_k) \geq (1 - 1/e)\mathrm{OPT}_k = (1 - 1/e) f(S_k^*)$, which is Eq. 50.

Relation to the implemented rule. The bound describes the quality of choosing a subset *at one state h_t*; accumulation across recursion steps is governed by Theorems 1–2. ReFuse's threshold rule $I_t = \{i : s_i^{\text{student},t} > \tau\}$ is a label-free batch relaxation of this greedy procedure (the student score is a learned surrogate of $\Delta(i \mid S)$, accurate up to $\epsilon_{\text{approx}} + \epsilon_{\text{distill}}$), and the greedy fallback $I_t = \{\arg\max_i s_i^{\text{student},t}\}$ is exactly the $k = 1$ greedy step.

□

A.7 Proof of Theorem 4 and Corollary 2

Theorem 4 (Generalization Bound)**.** *Let $\mathcal{G} = \{g_\psi \circ F_\theta^t : t \leq T_{\max}\}$ be the hypothesis class of ReFuse, where F_θ^t denotes t iterations of the diagnose–select–refuse loop. Let the loss $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, 1]$ be L-Lipschitz in its second argument. Then, with probability at least $1 - \zeta$ over the training draw of n i.i.d. samples, for every $g \in \mathcal{G}$,*

$$R(g) \leq \hat{R}(g) + 2L \cdot \mathfrak{R}_n(\mathcal{G}) + \sqrt{\frac{\log(1/\zeta)}{2n}}, \tag{54}$$

where $R(g) = \mathbb{E}_{(X,y)\sim P}[\ell(g(X), y)]$ is the true risk, $\hat{R}(g) = \frac{1}{n}\sum_{j=1}^{n} \ell(g(X_j), y_j)$ is the empirical risk, and $\mathfrak{R}_n(\mathcal{G}) = \mathbb{E}_{\sigma,X}[\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{j=1}^{n} \sigma_j g(X_j)]$ is the Rademacher complexity (we use ζ for the confidence parameter to distinguish it from the distillation weight δ).

Corollary 2 (Parameter Efficiency of Sharing in the Equal-Parameter Regime)**.** *Let a deep model stack T independently parameterised layers $\{W_l\}_{l=1}^{T}$, while ReFuse reuses one fusion operator W for T iterations. The spectral Rademacher bound is monotone in the per-layer spectral-norm product (Bartlett et al., 2017).*

(i) Equal per-step strength. *If $\|W_l\|_2 = \|W\|_2 = s$ for every l, both products equal s^T and hence admit the same spectral bound, but the deep model uses T times as many parameters.*

(ii) Equal total Frobenius-norm budget *(the equal-capacity setting considered here), $\sum_{l=1}^{T} \|W_l\|_F^2 = \|W\|_F^2$. Using $\|W_l\|_2 \le \|W_l\|_F$ and AM-GM,*

$$\prod_{l=1}^{T} \|W_l\|_2 \;\le\; \prod_{l=1}^{T} \|W_l\|_F \;\le\; \Big(\tfrac{1}{T}\sum_{l=1}^{T} \|W_l\|_F^2\Big)^{T/2} = \Big(\tfrac{\|W\|_F}{\sqrt{T}}\Big)^{T}. \tag{55}$$

Thus, unless the single shared operator has condition number $\|W\|_F/\|W\|_2 \ge \sqrt{T}$, an independent deep net with the same total Frobenius-norm budget *cannot attain at every layer the per-step spectral strength $\|W\|_2$ that ReFuse applies on every iteration; matching that per-step strength requires T times the Frobenius budget, i.e. $\sim T\times$ parameters.*

Proof. **Proof of Theorem 4.** Let $\mathcal{F} = \{(x, y) \mapsto \ell(g(x), y) : g \in \mathcal{G}\}$ be the loss class. By the standard Rademacher uniform-convergence bound (Mohri et al., 2018, Theorem 3.1), with probability at least $1 - \zeta$,

$$\sup_{f\in\mathcal{F}} |\mathbb{E}[f] - \hat{\mathbb{E}}[f]| \le 2\mathfrak{R}_n(\mathcal{F}) + \sqrt{\frac{\log(1/\zeta)}{2n}}. \tag{56}$$

Since ℓ is L-Lipschitz in its second argument, the Lipschitz contraction lemma (Mohri et al., 2018, Lemma 4.2) gives

$$\mathfrak{R}_n(\mathcal{F}) \le L \cdot \mathfrak{R}_n(\mathcal{G}). \tag{57}$$

Substituting Eq. 57 into Eq. 56, with $\mathbb{E}[f] = R(g)$ and $\hat{\mathbb{E}}[f] = \hat{R}(g)$, yields Eq. 54.

Proof of Corollary 2. For a feedforward network with spectral-normalised weights, the Rademacher complexity is bounded by a term monotone in the product of per-layer spectral norms (Bartlett et al., 2017),

$$\mathfrak{R}_n(\mathcal{G}) \le C \cdot \Big(\prod_{l=1}^{T} \|W_l\|_2\Big) \cdot \Big(\tfrac{1}{n}\sum_{j=1}^{n} \|X_j\|_2\Big)\sqrt{\tfrac{\log d}{n}}, \tag{58}$$

where C depends only on the activations and d on the dimension; the same form applies to attention via the product of the query/key/value projection norms.

Shared operator. ReFuse applies one operator W for T iterations, so its spectral product is $\|W\|_2^T$.

Claim (1), equal per-step strength. If $\|W_l\|_2 = \|W\|_2 = s$ for every independent layer, both products equal s^T and Eq. 58 gives the same bound, but the deep model stores T independent operators (T times the parameters of ReFuse).

Claim (2), equal total budget. The equal-parameter regime imposes $\sum_{l=1}^{T} \|W_l\|_F^2 = \|W\|_F^2$ (not T times it). Using $\|W_l\|_2 \le \|W_l\|_F$ and AM–GM on the non-negative numbers $\{\|W_l\|_F^2\}$,

$$\prod_{l=1}^{T} \|W_l\|_2 \le \prod_{l=1}^{T} \|W_l\|_F \le \Big(\tfrac{1}{T}\sum_{l=1}^{T} \|W_l\|_F^2\Big)^{T/2} = \Big(\tfrac{\|W\|_F}{\sqrt{T}}\Big)^{T}. \tag{59}$$

Thus the average independent layer can spend spectral norm at most $\|W\|_F/\sqrt{T}$, whereas the shared operator keeps the full $\|W\|_2$ at every iteration. Matching $\|W\|_2$ at every layer would require $\|W\|_F/\sqrt{T} \ge \|W\|_2$, i.e. a single-operator condition number $\|W\|_F/\|W\|_2 \ge \sqrt{T}$. Absent such strong anisotropy, an independent deep net under the same total budget cannot reproduce ReFuse's per-step strength, and recovering it needs $\sum_l \|W_l\|_F^2 = T\|W\|_F^2$, i.e. $\sim T\times$ parameters. This is

Eq. 55, stated as a parameter-efficiency / capacity-per-parameter result rather than an unconditional ordering of Rademacher complexities (which need not hold under a shared budget).

□

A.8 Connection to the Sequential Information Bottleneck

ReFuse admits a clean interpretation as a *sequential* information bottleneck (IB), which connects the recursive fusion loop to the classical IB principle (Tishby et al., 2000) and to multimodal IB methods (Mai et al., 2023b; Wu et al., 2025b).

From one-shot IB to sequential IB. A one-shot multimodal IB seeks a single fused representation z by optimizing

$$\max_{z} \; I(y; z) - \beta \, I(\mathbf{m}; z), \tag{60}$$

where the first term drives sufficiency and the second penalizes compression cost (redundancy and noise). In ReFuse, information is acquired incrementally: at step t only the selected modalities I_t contribute, and the representation evolves from h_{t-1} to h_t. Writing the incremental compression cost of step t as the conditional mutual information $I(h_t; \{m_i\}_{i \in I_t} \mid h_{t-1})$, the recursive process optimizes the *sequential* objective

$$\max_{\{I_t, h_t\}} \; I(y; h_T) - \beta \sum_{t=0}^{T-1} I(h_t; \, \{m_i\}_{i \in I_t} \mid h_{t-1}) \, , \tag{61}$$

with $h_{-1} = \varnothing$ and the initial fusion at $t = 0$. The diagnoser-driven selection $I_t = \{i : s_i^{\text{student},t} > \tau\}$ is thus a greedy optimizer of the sufficiency term: by the Proposition in Appendix B, the student score tracks the conditional gain $I(m_i; y \mid h_t)$ up to $\epsilon_{\text{approx}} + \epsilon_{\text{distill}}$, so ReFuse adds at each step only modalities whose estimated gain exceeds the *selection* threshold τ, and terminates (Section 3.4) once the largest estimated gain falls below the separate stopping threshold ϵ.

Why spreading acquisition over steps is cheaper. The sequential form 61 differs from 60 in two structural ways that favor ReFuse under modality imbalance. (i) *Selective acquisition.* Because I_t excludes modalities whose conditional gain is already below τ, the compression sum in 61 never pays for re-encoding redundant or noise-dominated modalities, whereas a one-shot bottleneck must compress all modalities jointly in a single pass. (ii) *Distributed, smaller effective bottleneck.* Under the information-preservation condition of Lemma 1, the *accumulated state* S_t admits an exact telescoping of task information,

$$I(y; S_T) = I(y; h_0) + \sum_{t=0}^{T-1} I(y; \{m_i\}_{i \in I_t} \mid S_t) \, , \tag{62}$$

i.e. the accumulated sufficiency is a sum of non-negative per-step increments, each of which only has to cross a local conditional bottleneck $I(h_t; \{m_i\}_{i \in I_t} \mid S_{t-1})$. The prediction state h_T is a deterministic function of S_T, so $I(y; h_T) \le I(y; S_T)$, and Theorem 2 bounds the remaining gap by $M\bar{\epsilon}$, ensuring the compressed state used for prediction is itself approximately sufficient. Reaching a target normalized sufficiency therefore spreads acquisition over several small, local bottlenecks rather than forcing one bottleneck to extract the entire signal at once: precisely the regime in which a single-pass model lets a dominant modality suppress a weak one.

Relationship to prior multimodal IB. MIB (Mai et al., 2023b) and OMIB (Wu et al., 2025b) regularize the compression term of 60 but still form the fused representation in one pass. ReFuse is complementary: it leaves the choice of compression regularizer to the fusion unit and instead reorganizes the sufficiency term into an adaptive sequence. The two perspectives are compatible (an IB-regularized fusion unit can be used as F_θ inside the ReFuse loop), together controlling both what to compress at each step and which modality to acquire next.

B Loss Difference as Conditional Mutual Information Surrogate

Fix a recursion step t and a modality i. The teacher diagnoser uses two predictors: a *base* predictor g_ψ that sees only the current fused state h_t, and an *augmented* predictor $g_{\psi,i}$ that additionally sees modality m_i. For a loss $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$, write their conditional Bayes risks and the corresponding population risks as

$$L_t^B := \inf_g \mathbb{E}[\ell(g(h_t), y) \mid h_t], \qquad \bar{L}_t^B := \mathbb{E}[L_t^B],$$
$$L_{t,i}^A := \inf_g \mathbb{E}[\ell(g(h_t, m_i), y) \mid h_t, m_i], \qquad \bar{L}_{t,i}^A := \mathbb{E}[L_{t,i}^A]. \tag{63}$$

The ideal (Bayes-optimal) teacher score and the score realized by the trained predictors are, respectively,

$$s_i^{\star,t} := \bar{L}_t^B - \bar{L}_{t,i}^A, \qquad \hat{s}_i^{\text{teacher},t} := \mathbb{E}[\ell(g_\psi(h_t), y) - \ell(g_{\psi,i}(h_t, m_i), y)]. \tag{64}$$

Because the augmented predictor is strictly better informed, $L_{t,i}^A \le L_t^B$ pointwise and $s_i^{\star,t} \ge 0$.

Proposition 1 (Teacher loss difference as conditional information gain)**.** *With the notation above, the teacher score satisfies the following.*

(1) Classification (cross-entropy, discrete y). *At the Bayes optimum the conditional cross-entropy risk equals the conditional entropy, so*

$$s_i^{\star,t} = H(y \mid h_t) - H(y \mid h_t, m_i) = I(m_i; y \mid h_t). \tag{65}$$

(2) Regression (squared loss, continuous y). *The Bayes predictor is $\mathbb{E}[y \mid \cdot]$ and the Bayes risk is the conditional variance.*

(2a) *If (y, h_t, m_i) are jointly Gaussian (homoscedastic conditionals), writing $\sigma_{t,i}^2 := \mathrm{Var}(y \mid h_t, m_i) > 0$,*

$$s_i^{\star,t} = \mathrm{Var}(y \mid h_t) - \sigma_{t,i}^2 = \sigma_{t,i}^2 \Big(e^{2I(m_i; y \mid h_t)} - 1 \Big), \tag{66}$$

which is non-negative, equals zero iff $I(m_i; y \mid h_t) = 0$, and is strictly increasing in $I(m_i; y \mid h_t)$.

(2b) *If $y \mid h_t$ and $y \mid h_t, m_i$ are conditionally Gaussian with state-dependent (heteroscedastic) variances, then pointwise in $u = (h_t, m_i)$,*

$$\Delta_i(u) = \tfrac{1}{2\pi e} \Big(e^{2h_{\mathrm{de}}(y \mid h_t)} - e^{2h_{\mathrm{de}}(y \mid h_t, m_i)} \Big) = \tfrac{1}{2\pi e} e^{2h_{\mathrm{de}}(y \mid h_t, m_i)} \Big(e^{2\iota_i(u)} - 1 \Big), \tag{67}$$

where $h_{\mathrm{de}}(\cdot \mid \cdot)$ is differential entropy (not the fusion state) and $\iota_i(u) := h_{\mathrm{de}}(y \mid h_t) - h_{\mathrm{de}}(y \mid h_t, m_i)$ is the pointwise conditional information gain with $\mathbb{E}_u[\iota_i(u)] = I(m_i; y \mid h_t)$. For each fixed $h_{\mathrm{de}}(y \mid h_t, m_i)$, $\Delta_i(u)$ is strictly increasing in $\iota_i(u)$, and $\Delta_i(u) = 0$ almost surely iff $y \perp m_i \mid h_t$.

(3) Finite-capacity and finite-sample approximation. *Define the non-negative excess risks of the two trained predictors,*

$$\epsilon_B := \mathbb{E}[\ell(g_\psi(h_t), y)] - \bar{L}_t^B, \qquad \epsilon_A := \mathbb{E}[\ell(g_{\psi,i}(h_t, m_i), y)] - \bar{L}_{t,i}^A. \tag{68}$$

Under cross-entropy, the triangle inequality gives

$$\big| \hat{s}_i^{\text{teacher},t} - I(m_i; y \mid h_t) \big| \le \epsilon_B + \epsilon_A =: \epsilon_{\mathrm{opt}}. \tag{69}$$

If the population score is further replaced by an empirical mean over a batch of B samples and ℓ is bounded on a range of width c_ℓ, a Hoeffding bound plus a union bound over the two predictors adds, with probability at least $1 - \delta$, a statistical term $\epsilon_{\mathrm{stat}} = O\Big(c_\ell \sqrt{\log(1/\delta)/B} \Big)$. Hence

$$\big| \hat{s}_i^{\text{teacher},t} - I(m_i; y \mid h_t) \big| \le \epsilon_{\mathrm{approx}}, \qquad \epsilon_{\mathrm{approx}} := \epsilon_B + \epsilon_A + \epsilon_{\mathrm{stat}}, \tag{70}$$

with $\epsilon_B, \epsilon_A \to 0$ as the predictors approach their Bayes optima and $\epsilon_{\mathrm{stat}} \to 0$ as $B \to \infty$. For regression, the same bound holds with $I(m_i; y \mid h_t)$ replaced by the strictly increasing transform in Part (2) when the homoscedastic Gaussian assumption holds; under heteroscedasticity, the loss difference remains a monotone surrogate pointwise, as discussed

in Part (2b). In particular, Eq. 70 always implies the one-sided bound used in the main text,

$$\hat{s}_i^{\textit{teacher},t} \geq I(m_i; y \mid h_t) - \epsilon_{\text{approx}}. \tag{71}$$

Proof. Part (1). Under cross-entropy, predicting the true posterior $p^*(y \mid \cdot)$ is Bayes optimal, and the resulting conditional risk is the conditional entropy (the Gibbs inequality): $\bar{L}_t^B = H(y \mid h_t)$ and $\bar{L}_{t,i}^A = H(y \mid h_t, m_i)$. The entropy chain rule $H(y \mid h_t) = H(y \mid h_t, m_i) + I(m_i; y \mid h_t)$ gives Eq. 65.

Part (2a). For squared loss the Bayes predictor is the conditional mean and the Bayes risk the conditional variance. For jointly Gaussian variables, $h_{\text{de}}(y \mid z) = \frac{1}{2}\log(2\pi e \operatorname{Var}(y \mid z))$, hence

$$I(m_i; y \mid h_t) = h_{\text{de}}(y \mid h_t) - h_{\text{de}}(y \mid h_t, m_i) = \tfrac{1}{2}\log\frac{\operatorname{Var}(y \mid h_t)}{\sigma_{t,i}^2}, \tag{72}$$

i.e. $\operatorname{Var}(y \mid h_t) = \sigma_{t,i}^2 e^{2I(m_i;y|h_t)}$. Subtracting the two conditional variances gives Eq. 66; since $\sigma_{t,i}^2 > 0$ and $e^{2I} - 1$ is strictly increasing and null only at $I = 0$, the claims follow.

Part (2b). Pointwise, the conditional Gaussian risk equals the conditional variance $\frac{1}{2\pi e} e^{2h_{\text{de}}(y|\cdot)}$; subtracting the augmented from the base risk yields the first equality in Eq. 67, and factoring out $e^{2h_{\text{de}}(y|h_t,m_i)}$ gives the second. Its derivative with respect to ι_i is

$$\frac{\partial \Delta_i}{\partial \iota_i} = \frac{1}{\pi e} e^{2h_{\text{de}}(y|h_t,m_i)+2\iota_i(u)} > 0, \tag{73}$$

so Δ_i is strictly increasing in the pointwise gain, and because the prefactor $\frac{1}{2\pi e} e^{2h_{\text{de}}(y|h_t,m_i)}$ is positive, $\Delta_i = 0$ a.s. iff $\iota_i = 0$ a.s., i.e. $\operatorname{Var}(y \mid h_t) = \operatorname{Var}(y \mid h_t, m_i)$ a.s. The law of total variance gives

$$\operatorname{Var}(y \mid h_t) = \mathbb{E}[\operatorname{Var}(y \mid h_t, m_i) \mid h_t] + \operatorname{Var}(\mathbb{E}[y \mid h_t, m_i] \mid h_t)\,; \tag{74}$$

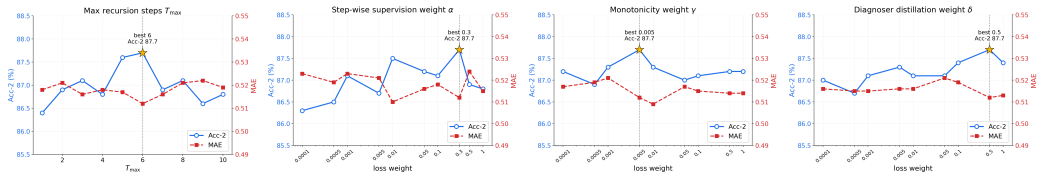
substituting equality of the two conditional variances forces $\operatorname{Var}(\mathbb{E}[y \mid h_t, m_i] \mid h_t) = 0$, so $\mathbb{E}[y \mid h_t, m_i] = \mathbb{E}[y \mid h_t]$ a.s. A Gaussian conditional law is uniquely determined by its mean and variance, and matching both is equivalent to $y \perp m_i \mid h_t$, i.e. $I(m_i; y \mid h_t) = 0$. Under heteroscedasticity $\mathbb{E}[\Delta_i]$ is not a deterministic one-to-one function of the CMI alone (the weight $e^{2h_{\text{de}}(y|h_t,m_i)}$ is state-dependent), but the pointwise monotonicity and the exact zero condition still make the loss difference a faithful ranking and thresholding surrogate of the conditional gain, which is all the diagnoser requires for modality selection and termination.

Part (3). Because $s_i^{\star,t} = I(m_i; y \mid h_t)$ for classification (Part 1),

$$\begin{aligned}\left|\hat{s}_i^{\text{teacher},t} - s_i^{\star,t}\right| &= \left|\left(\mathbb{E}\ell(g_\psi(h_t), y) - \bar{L}_t^B\right) - \left(\mathbb{E}\ell(g_{\psi,i}(h_t, m_i), y) - \bar{L}_{t,i}^A\right)\right| \\ &\leq \epsilon_B + \epsilon_A\end{aligned} \tag{75}$$

by the triangle inequality, proving Eq. 69. For bounded $\ell \in [0, c_\ell]$, Hoeffding's inequality bounds each empirical risk within $c_\ell\sqrt{\log(4/\delta)/(2B)}$ with probability at least $1 - \delta/2$. A union bound and the triangle inequality give the stated ϵ_{stat}, proving Eq. 70. The one-sided Eq. 71 follows immediately from $|\hat{s} - I| \leq \epsilon_{\text{approx}}$. For regression, replacing the entropy identity by Part (2)'s strictly increasing variance transform reproduces the same excess-risk and finite-sample argument. □

Remark 2 (Conditioning on learned features, and why not reconstruction)**.** *The score conditions on the encoded modality $m_i = e_i(x_i)$, not on the raw observation x_i. By the conditional data-processing inequality, $I(m_i; y \mid h_t) \leq I(x_i; y \mid h_t)$, approached from below as the unimodal encoder e_i becomes sufficient. The teacher thus measures the* actionable *gain that re-fusion can actually recover from the current representation. Crucially this is task information, not all information: a reconstruction criterion $\|m_i - \hat{m}_i(h_t)\|$ would instead demand that h_t preserve every component of m_i, including noise, redundancy and task-irrelevant variation, and would trigger re-fusion even for modalities that carry no signal about y. The loss-difference score is nonzero only when m_i reduces predictive uncertainty about y, which is exactly the quantity governing selection and termination.*



(a) Maximum recursion steps $T_{\max}$. (b) Step-wise supervision loss weight α. (c) Monotonicity weight γ. (d) Diagnoser-distillation loss weight δ.

Figure 4: **Hyperparameter sensitivity on MOSEI.** Each panel sweeps one key hyperparameter, reporting Acc2 (Ablue solid, left axis) and MAE (red dashed, right axis). The selected operating point is marked with a star and a vertical line, and loss weights use a logarithmic x-axis. All four hyperparameters exhibit a wide flat optimum.

C Additional Experimental Results

C.1 Hyperparameter sensitivity analysis

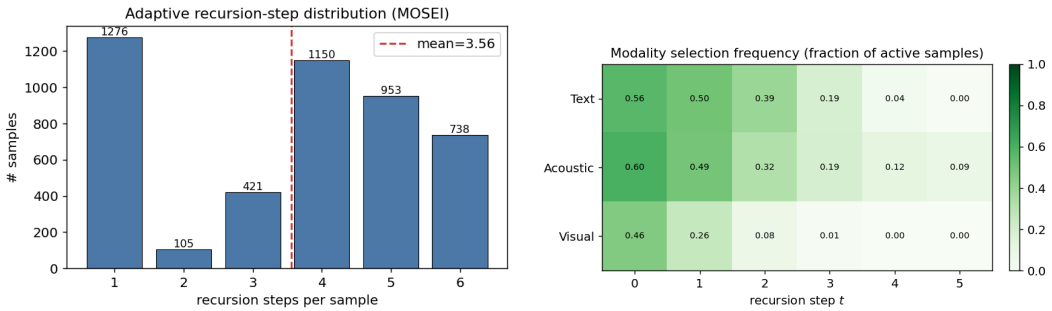
Figure 4a–4d study the four key hyperparameters on MOSEI, using wide axis ranges to expose how little the result varies.

(1) Maximum recursion steps $T_{\max}$. A single fusion pass ($T_{\max} = 1$) is the worst setting. The Acc2 of the model rises with more allowed passes and peaks at $T_{\max} = 6$, after which it plateaus and fluctuates slightly. This follows from how ReFuse uses $T_{\max}$: it is only an upper bound, while the actual depth is decided per sample by the stopping criterion. At $T_{\max} = 1$ the loop is effectively disabled and every weak modality is fused exactly once, so the information it fails to contribute is irrecoverable, which is the exact under-fusion problem we target. Enlarging the cap gives hard samples enough diagnose-refuse rounds to recover the missing task-useful information, explaining the initial rise. Once the cap exceeds what most samples need (the majority terminate within roughly 3–6 steps), the extra headroom is left unused by easy samples and only a small tail benefits, so the marginal gain vanishes. The slight dip beyond 6 is consistent with a marginally longer unrolled supervision path and single-run noise, rather than a need for deeper fixed stacks.

(2) Step-wise supervision weight α. The Acc2 of the model is low for very small α (86.3-87.1 for $\alpha \leq 10^{-3}$), forms a broad optimum over $\alpha \in [0.01, 0.3]$, and weakens slightly for $\alpha \geq 0.5$. The step-wise loss places a prediction objective on every intermediate h_t, which both makes each state immediately 'judgeable' and injects a direct gradient at each step, easing optimization when gradients flow through repeated applications of the shared block. Too small a weight leaves intermediate predictions unsupervised, and re-uses of the block receive little gradient, approaching an under-trained one-shot model. Too large a weight forces every early, deliberately incomplete h_t to already predict the label, conflicting with progressive completion and encouraging the network to fit the target with the dominant modality it already has, hardening the bias instead of absorbing more information later. A moderate weight keeps intermediate states judgeable while allowing incremental completion, which is why a wide interval (not a single point) works well.

(3) Monotonicity weight γ. The curves are almost invariant across four orders of magnitude: Acc-2 stays within a 0.8-point band and MAE within 0.509-0.521, with only a marginal preference for $\gamma \in [0.005, 0.01]$. The monotonicity term is a structural regularizer that encourages a modality's diagnosed insufficiency to decrease after it is re-fused, preventing the loop from oscillating or repeatedly selecting the same modality. It has so little effect because the main loop already produces this behavior implicitly: once a modality is injected, the updated h_t naturally lowers that modality's loss-difference score at the next diagnosis. The term therefore makes explicit a trend that would mostly occur anyway, acting as a stabilizer against noisy diagnoses. A small weight suffices to suppress oscillation, larger weights add nothing once the constraint is satisfied, and the term is never harmful because its direction agrees with the objective.

(4) Diagnoser-distillation weight δ. The Acc2 of the model is slightly lower for small δ, peaks at $\delta = 0.5$, and retreats marginally at $\delta = 1$. The teacher diagnoser is accurate but relies on information unavailable at inference, whereas the student operates only from h_t and is what actually runs at test



(a) Per-sample recursion steps on MOSEI (mean 3.56, $T_{\max} = 6$).

(b) Modality selection frequency among still-active samples at each step.

Figure 5: **Adaptive, selective fusion.** The number of fusion passes is not fixed: easy samples terminate after a single pass while harder ones recurse more (a). Selection is sample- and step-specific rather than all-modal, and becomes increasingly rare as the representation becomes sufficient (b).

time. Distillation loss trains the student's insufficiency scores and selection decisions to match the teacher. With too small a weight the student under-fits the teacher, so at inference it mis-identifies the deficient modality and stops too early or too late. As δ approaches 0.5, the student faithfully reproduces the teacher and the closed-loop behavior seen during training transfers to inference. Pushing δ to 1 is slightly worse because the distillation term then competes on equal footing with the task loss and the shared representation over-accommodates the auxiliary target.

Generally, all four hyperparameters show a wide flat optimum, and the trends themselves corroborate the design: a single pass is insufficient, yet blindly allowing more fixed depth brings no gain, while the auxiliary objectives help only at moderate weights. ReFuse therefore does not depend on brittle, finely tuned settings and is robust to different values of the hyperparameters.

C.2 Visualization of adaptive selective fusion

Figure 5 summarizes the adaptive behavior of ReFuse: (1) **The depth is sample-dependent.** The number of passes ranges from 1 to $T_{\max}$, with an average of only 3.56 and 27.5% of samples terminating right after the initial fusion. This spread reflects the heterogeneity of the data: easy samples, for which the dominant text already determines the label, stop early, whereas harder samples that rely on non-verbal cues recurse more. Because the mean is well below the cap of 6, the upper bound is rarely reached and the loop does not waste passes on samples that need none; (2) **The fusion is also selective.** At each step only a subset of modalities is re-injected, and selection is biased toward the weak acoustic and visual modalities in the early steps. As their information is recovered, the selection frequency decreases monotonically toward zero. The visual modality, for example, is almost never selected after step 3. This is exactly the behavior the diagnoser is designed to produce: a modality with a high insufficiency score is chosen, and after re-fusion its score at the next step drops. Re-fusing every modality at every step would instead add redundant computation and let the dominant modality be reinforced repeatedly; selecting only the deficient one avoids both.

C.3 Visualization on prediction convergence

Figure 6 shows how the prediction evolves along the loop. Most of the adjustment takes place in the first two to three steps, where the missing weak-modality information is being added and the decision is corrected; accordingly $|g(h_t) - g(h_T)|$ drops by roughly an order of magnitude, from $\sim 2 \times 10^{-2}$ to below 10^{-3}, and the trajectories then become nearly flat. This indicates that once h_t contains sufficient information, further re-fusion barely changes the output. The convergence pattern also validates the stopping point. Prediction stabilization and the diagnoser's termination criterion occur together: the loop stops when the insufficiency scores fall below the threshold, which is precisely when the prediction no longer moves. It therefore neither stops too early (which would leave a still-changing, under-fused decision) nor keeps iterating after the output has saturated. A small tail of

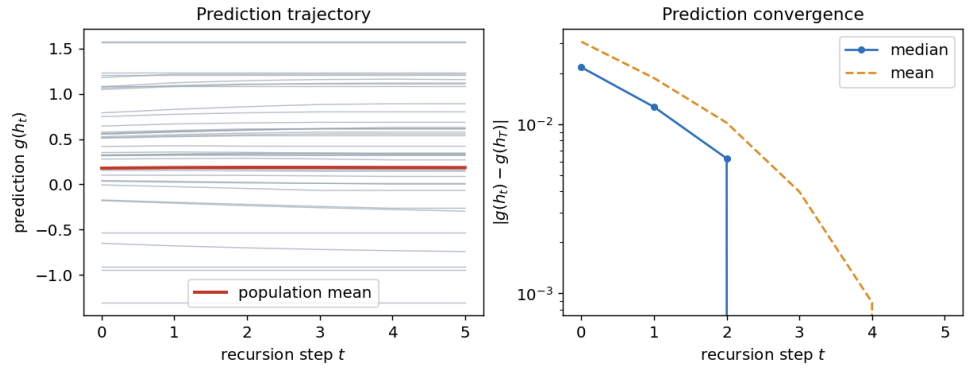


Figure 6: Prediction trajectories (left) and the median/mean change $|g(h_t) - g(h_T)|$ (right) on MOSEI. Predictions are refined in early steps and stabilize afterwards.

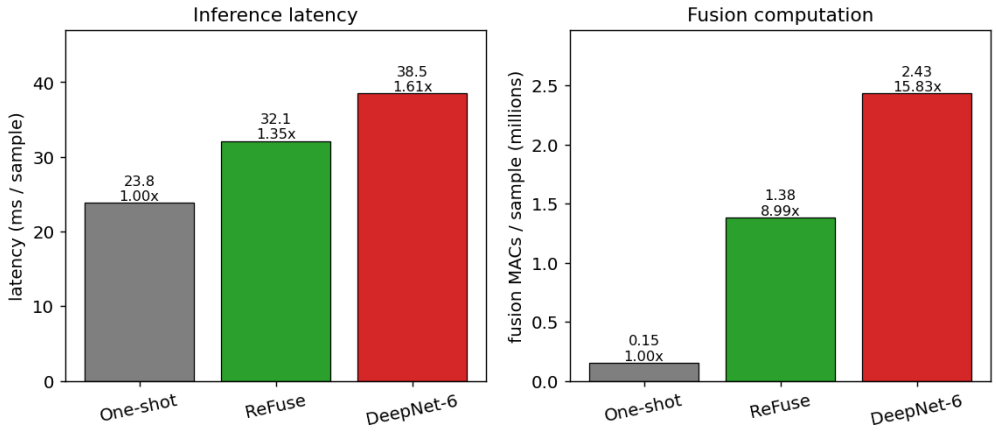


Figure 7: **Efficiency vs. a fixed-depth fusion network on MOSEI.** End-to-end latency (left) and fusion-only MACs (right) per sample, measured over 60 samples (median latency; the DeBERTa backbone is shared across configurations and dominates end-to-end latency). DeepNet-6 runs the parameter-shared fusion operator for six passes without early termination, while ReFuse terminates adaptively.

hard samples shows slightly larger late-step changes, consistent with their using more passes, but the aggregate behavior is rapid refinement followed by a stable plateau.

C.4 Efficiency Analysis

Figures 7-8 examine whether simply running the fusion operator for more passes is a good use of computation. Against a six-layer deep fusion network (Figure 7), ReFuse attains a slightly higher correlation with only $0.83\times$ the latency and $0.57\times$ the fusion MACs. The depth sweep (Figure 8) makes the reason explicit: pushing the fixed-depth operator from 4 to 10 passes does not raise the correlation but slowly lowers it from 0.792 to 0.784, while latency and fusion MACs grow (near-)linearly. Every fixed-depth operating point therefore sits above and to the right of ReFuse (more expensive yet no better), so ReFuse is Pareto-dominant. Two details explain the size of the saving. First, a fixed stack applies the same number of passes to every sample, including the many easy ones that need only one, whereas ReFuse spends passes on demand. Second, the DeBERTa backbone is shared across configurations and dominates end-to-end latency, so the fusion-only saving ($0.57\times$ MACs) is partly diluted at the full-model level ($0.83\times$ latency): on the fusion sub-network itself the advantage is larger. ReFuse thus uses the diagnoser to fuse only the modalities that still carry task-useful information and stops at the saturation point, obtaining a strictly better accuracy-compute

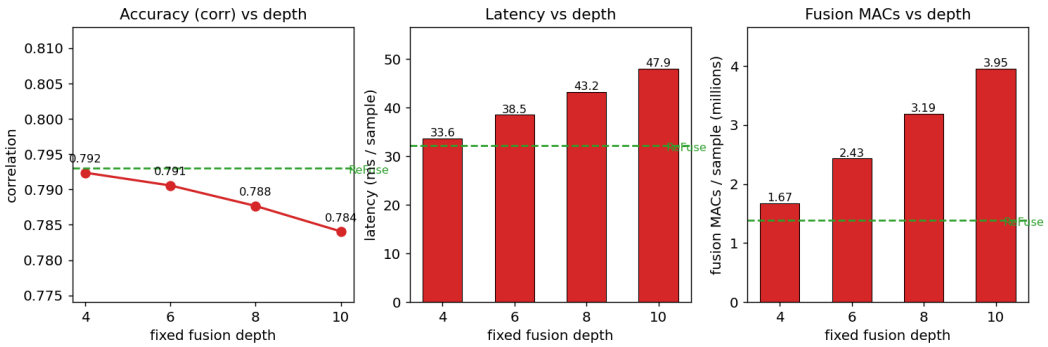


Figure 8: **Depth sweep on MOSEI.** Correlation (left), latency (middle) and fusion MACs (right) when the parameter-shared fusion operator is run for a fixed $4, 6, 8, 10$ passes; the green dashed line marks ReFuse's adaptive operating point. Correlation does not improve: it slightly degrades with depth, whereas latency and fusion MACs grow (near-)linearly.

Table 6: Model complexity analysis on the MUStARD and MOSI datasets.

MUStARD			MOSI			
Model	Acc2	Parameters (M)	Model	Acc2	Parameters (M)	FLOPs (G)
HKT Hasan et al. (2021)	76.47	17.10	EMOE Fang et al. (2025)	84.8	317.30	**10.02**
MCL Mai et al. (2023a)	<u>77.94</u>	<u>13.83</u>	GSCon Shi et al. (2025)	<u>88.1</u>	248.52	12.22
MGCL Mai et al. (2023c)	<u>77.94</u>	14.28	MCIB Wang et al. (2026)	86.4	<u>220.43</u>	13.21
ReFuse	**79.41**	**13.12**	ReFuse	**89.5**	**186.66**	<u>10.09</u>

trade-off. Results for deep networks with independent (non-shared) blocks under a matched parameter budget are reported in Section 5.2.

C.5 Model Complexity Analysis

We compare the model complexity of ReFuse with competitive baselines on the MUStARD (Castro et al., 2019) and MOSI (Zadeh et al., 2016) datasets. On MUStARD, Table 6 shows that ReFuse attains the highest binary accuracy while using the fewest parameters, i.e. about 23% fewer than HKT and also fewer than MCL and MGCL. ReFuse therefore performs better with a smaller model, so its gain does not come from increasing the parameter count. On MOSI, we additionally report FLOPs. ReFuse again achieves the best accuracy and the fewest parameters, which is roughly 41% fewer than EMOE and well below GSCon and MCIB. Its FLOPs are close to the lowest and clearly below GSCon and MCIB. In other words, ReFuse obtains the highest accuracy with a smaller model and competitive computation. The reason is cross-step parameter sharing: a single full-capacity fusion block is applied repeatedly, so letting the model iterate over insufficiently-learned modalities adds fusion passes but almost no extra parameters. This is consistent with our central idea of widening the fusion rather than deepening the network, and it shows that the improvement is bought through better use of one shared block rather than a larger or heavier model.

D Algorithm Details

D.1 Algorithm Pipeline

The training and inference pipelines are shown in Algorithm 1 and Algorithm 2.

D.2 Modality Encoders

ReFuse places no restriction on how each raw modality is encoded: fixed, task-agnostic unimodal encoders first map every modality into the common d-dimensional space of Section 3.1, and only the resulting vectors m_i enter the recursive fusion loop. We describe the encoder instantiation shared by all tasks. Following common practice (Xiao et al., 2024; Hasan et al., 2021; Wang et al., 2025),

Algorithm 1 ReFuse Training (one batch; sg = stop-gradient)

Require: batch $\{(\mathbf{m}^{(j)}, y^{(j)})\}$, weights $\alpha, \beta, \gamma, \delta$, thresholds τ, ϵ, η, bounds $T_{\min}, T_{\max}$
Ensure: F_θ, g_ψ, D_ϕ, projections $\{W'_i\}$

$h_0^{(j)} \leftarrow \mathrm{LN}\big(f_0(\mathrm{Concat}(m_1^{(j)}, \ldots, m_M^{(j)}))\big)$; $a_0^{(j)} \leftarrow 1$
for $t = 0$ **to** $T_{\max} - 1$ **do**
  $\hat{y}_t^{(j)} \leftarrow g_\psi(h_t^{(j)})$, $\ell_t^{(j)} \leftarrow \mathcal{L}(\hat{y}_t^{(j)}, y^{(j)})$ {loss on the current state}
  for $i = 1$ **to** M **do**
    $c_i^{t,j} \leftarrow W'_i m_i^{(j)}$; $\hat{y}_{t,i}^{(j)} \leftarrow g_\psi(\mathrm{LN}(h_t^{(j)} + c_i^{t,j}))$
    $k_i^{t,j} \leftarrow \mathrm{shift}(\hat{y}_t^{(j)}, \hat{y}_{t,i}^{(j)})$ {Eq. 9, detached}
    $s_i^{\mathrm{tea},t,j} \leftarrow \mathrm{sg}\left(\ell_t^{(j)} - \mathcal{L}(\hat{y}_{t,i}^{(j)}, y^{(j)})\right)$
  end for
  $s_i^{\mathrm{stu},t,j} \leftarrow D_\phi(\mathrm{sg}(h_t^{(j)}), c_i^{t,j}, \mathrm{sg}(k_i^{t,j}))\ \forall i$
  Accumulate $\mathcal{L}_{\text{diag}}$ (Eq. 11), $\mathcal{L}_{\text{step}} \mathrel{+}= a_t\, \ell_t$, and $\mathcal{L}_{\text{mono}} \mathrel{+}= a_t[\ell_t - \ell_{t-1}]_+$ (with $\ell_{-1} := \ell_0$)
  $\bar{s}_t^j \leftarrow \max_i s_i^{\mathrm{stu},t,j}$
  if training and $t < T_{\min}$ **then**
    $a_{t+1}^j \leftarrow a_t^j$ {warm-up: keep active}
  else
    $a_{t+1}^j \leftarrow a_t^j \mathbb{1}[\bar{s}_t^j > \epsilon] \mathbb{1}[d(\hat{y}_t, \hat{y}_{t-1}) > \eta]$
  end if
  if no sample is active **then**
    break
  end if
  $I_t^j \leftarrow \{i : s_i^{\mathrm{stu},t,j} > \tau\}$; if $I_t^j = \varnothing$ for an active sample then $I_t^j \leftarrow \{\arg\max_i s_i^{\mathrm{stu},t,j}\}$ {greedy fallback}
  $\Delta h_t \leftarrow \mathrm{MultiHeadAttn}(h_t, \{m_i\}_{i \in I_t})$, $g_t \leftarrow \sigma(W_g[h_t; \Delta h_t; [\max_{i \in I_t} s_i^{\mathrm{stu},t}]_+])$
  $u_t \leftarrow \mathrm{LN}_1(h_t + g_t \odot \Delta h_t)$; $h_{t+1} \leftarrow \mathrm{where}(a_{t+1}, \mathrm{LN}_2(u_t + \mathrm{FFN}(u_t)), h_t)$
end for
Normalise $\mathcal{L}_{\text{step}}, \mathcal{L}_{\text{mono}}, \mathcal{L}_{\text{diag}}$ by executed steps; $\mathcal{L}_{\text{budget}} \leftarrow \mathbb{E}[T]/T_{\max}$
$\mathcal{L} \leftarrow \mathcal{L}(g_\psi(h_T), y) + \alpha\mathcal{L}_{\text{step}} + \beta\mathcal{L}_{\text{budget}} + \gamma\mathcal{L}_{\text{mono}} + \delta\mathcal{L}_{\text{diag}}$; backpropagate and update

Algorithm 2 ReFuse Inference (Label-Free)

Require: modalities $\mathbf{m} = (m_1, \ldots, m_M)$, trained $F_\theta, g_\psi, D_\phi, \{W'_i\}$, thresholds τ, ϵ, η, bound $T_{\max}$
Ensure: prediction $\hat{y}$

$h_0 \leftarrow \mathrm{LN}(f_0(\mathrm{Concat}(m_1, \ldots, m_M)))$; active←true
for $t = 0$ **to** $T_{\max} - 1$ **do**
  for $i = 1$ **to** M **do**
    $c_i^t \leftarrow W'_i m_i$; $k_i^t \leftarrow \mathrm{shift}(g_\psi(h_t),\ g_\psi(\mathrm{LN}(h_t + c_i^t)))$
    $s_i^t \leftarrow D_\phi(h_t, c_i^t, k_i^t)$
  end for
  if $\max_i s_i^t \le \epsilon$ **or** $d(\hat{y}_t, \hat{y}_{t-1}) \le \eta$ **then**
    break {sufficient / stable}
  end if
  $I_t \leftarrow \{i : s_i^t > \tau\}$; if $I_t = \varnothing$ then $I_t \leftarrow \{\arg\max_i s_i^t\}$
  $\Delta h_t \leftarrow \mathrm{MultiHeadAttn}(h_t, \{m_i\}_{i \in I_t})$; $g_t \leftarrow \sigma(W_g[h_t; \Delta h_t; [\max_{i \in I_t} s_i^t]_+])$
  $u_t \leftarrow \mathrm{LN}_1(h_t + g_t \odot \Delta h_t)$; $h_{t+1} \leftarrow \mathrm{LN}_2(u_t + \mathrm{FFN}(u_t))$
end for
return $\hat{y} = g_\psi(h_T)$

the language stream is encoded by a pre-trained language model (PLM) (He et al., 2021; Lan et al., 2020) and then projected to the common dimension:

$$\begin{aligned} \tilde{u}_l &= \mathrm{PLM}(u_l; \theta_l) \in \mathbb{R}^{T_l \times d_l}, \\ m_l^{\mathrm{seq}} &= \tilde{u}_l W_{\mathrm{proj}} + b_{\mathrm{proj}} \in \mathbb{R}^{T_l \times d}, \end{aligned} \tag{76}$$

where u_l is the input token sequence of length T_l, and the trainable $W_{\text{proj}} \in \mathbb{R}^{d_l \times d}$, $b_{\text{proj}} \in \mathbb{R}^{1 \times d}$ align the PLM output width with the common dimension d.

For MSA, the acoustic and visual streams use a temporal convolution followed by a Transformer encoder (Vaswani et al., 2017) ($i \in \{a, v\}$, kernel size $K_i = 3$):

$$\begin{aligned} \hat{u}_i &= \text{Conv1D}(u_i; K_i) \in \mathbb{R}^{T_i \times d}, \\ m_i^{\text{seq}} &= \text{Transformer}(\hat{u}_i; \theta_i) \in \mathbb{R}^{T_i \times d}. \end{aligned} \tag{77}$$

On CH-SIMS-v2 the released features are per-utterance vectors rather than sequences (Liu et al., 2022), so for the visual and acoustic modalities we replace the convolutional Transformer stack by a multi-layer perceptron that maps directly into $\mathbb{R}^d$.

For MHD and MSD, each example consists of a target punchline and its preceding context; concatenating the context and punchline feature streams along time yields the unimodal sequence u_i ($i \in \mathcal{M} = \{a, v, l\}$), which is then encoded by a Transformer and a temporal convolution in the reverse order of Eq. 77 ($i \in \{a, v\}$):

$$\begin{aligned} \hat{u}_i &= \text{Transformer}(u_i; \theta_i) \in \mathbb{R}^{T_i \times d_i}, \\ m_i^{\text{seq}} &= \text{Conv1D}(\hat{u}_i; K_i) \in \mathbb{R}^{T_i \times d}. \end{aligned} \tag{78}$$

Irrespective of the task, each encoded sequence is layer-normalized and averaged over time so that every modality is represented by one common-space vector:

$$m_i = \text{MeanPooling}\big(\text{LN}(m_i^{\text{seq}})\big) \in \mathbb{R}^d. \tag{79}$$

These m_i are exactly the modality vectors consumed by ReFuse: they form the initial state $h_0 = f_0(\mathbf{m})$ and supply the conditional features $c_i^t = W_i' m_i$ used by both the fusion unit and the sufficiency diagnoser (Section 3).

For the AVE (Tian et al., 2018) dataset, We keep the modality-specific backbones of the imbalance-learning benchmark (Qin et al., 2026) unchanged: a ResNet18 encoder for the visual stream (three randomly sampled frames per iteration) and a single-channel ResNet18 for the audio stream, and only replace their one-shot fusion module with the ReFuse diagnose-select-refuse loop.

E Experiment Settings

E.1 Dataset Composition

(1) **MOSI** (Zadeh et al., 2016): MOSI is a standard MSA corpus containing more than 2,000 opinion video clips. Annotators assign each clip a sentiment intensity in $[-3, +3]$, where the two endpoints stand for the strongest negative and strongest positive polarity, respectively.

(2) **MOSEI** (Zadeh et al., 2018): As a larger-scale MSA benchmark, MOSEI gathers over 22,000 clips from more than 1,000 YouTube speakers covering 250 topics, sampled so as to span heterogeneous subjects and presentation styles. Each clip carries two kinds of labels: one of six emotion categories, and a sentiment-intensity score in $[-3, +3]$. We use the latter to evaluate ReFuse on the MSA task, which follows the same scoring convention as MOSI.

(3) **CH-SIMS-v2** (Liu et al., 2022): This Chinese MSA benchmark is sourced from interviews, talk shows and films so as to approximate realistic human-computer interaction; clips that fail an acoustic/visual quality filter are discarded. It adopts a 9:2:3 train/validation/test split, with 2,722 training utterances (921 negative, 433 weakly negative, 232 neutral, 318 weakly positive and 818 positive), 647 validation utterances and 1,034 test utterances.

(4) **UR-FUNNY** (Hasan et al., 2019): Built for multimodal humor detection (MHD), UR-FUNNY is collected from TED talks delivered by 1,741 speakers. Every target (punchline) utterance covers the language, acoustic and visual modalities and is modelled jointly with its preceding context utterances. Punchlines are localised through the transcript "laughter" cue that marks audience laughter, whereas a target utterance without a following laughter cue is treated as a negative sample. The corpus is partitioned into 7,614 training, 980 validation and 994 test examples; in line with prior work (Hasan et al., 2021; Mai et al., 2023a;c) we use version 2.

(5) **MUStARD** (Castro et al., 2019): Compiled from four TV series—Friends, The Big Bang Theory, The Golden Girls and Sarcasmaholics—MUStARD targets multimodal sarcasm detection (MSD). Its 690 utterances are manually labelled as sarcastic or non-sarcastic, and every target utterance is accompanied by the preceding dialogue turns (context utterances) that supply conversational evidence.

(6) **AVE** (Tian et al., 2018): AVE is an audio-visual event recognition benchmark of in-the-wild YouTube videos spanning 28 event categories (e.g. dog barking, guitar playing, racing). Each sample consists of an audio track and a visual track with temporal audio-visual event annotations. Following the unified protocol of Qin et al. (2026), we use the 3,312 / 402 / 401 training / validation / test split and treat event recognition as an audio-visual classification problem with two modalities. AVE exhibits severe modality imbalance: on this split the visual stream is the weak modality, so a single fusion pass can under-use the visual signal.

E.2 Assessment Criteria

We evaluate ReFuse and every baseline on MOSI and MOSEI with five standard metrics. (1) Acc7: seven-class accuracy, obtained after rounding both predictions and ground-truth scores to the nearest integer in $\{-3, \ldots, +3\}$. (2) Acc2: binary accuracy that distinguishes positive from negative sentiment. (3) F1: the harmonic mean of precision and recall for this binary sentiment task. Neutral utterances are excluded from both Acc2 and F1. (4) MAE: the mean absolute deviation of predictions from the ground-truth scores. (5) Corr: the Pearson correlation between predicted and ground-truth scores, reflecting the strength and direction of their linear association.

For CH-SIMS-v2 we report Acc5, Acc3, Acc2, F1, MAE and Corr, in keeping with established protocol (Liu et al., 2022; Feng et al., 2024).

For MHD and MSD we follow prior work (Hasan et al., 2021; Mai et al., 2023a;c) and report binary classification accuracy, i.e. whether a sample is correctly recognised as humorous (MHD) or sarcastic (MSD).

For the AVE (Tian et al., 2018) dataset, we report top-1 accuracy (%).

E.3 Feature Extraction Strategy

We extract unimodal features with the pipelines below.

(1) Visual modality: On the MSA benchmarks we obtain a broad visual descriptor set (facial action units and landmarks among them) with Facet[1], as in prior work (Mai et al., 2023d; Xiao et al., 2024). For CH-SIMS-v2 we follow the standard protocol (Feng et al., 2024; Liu et al., 2022) and use OpenFace (Baltrusaitis et al., 2018), which yields 68 facial landmarks, 17 action units, head pose and orientation, and eye-gaze direction. On MHD and MSD, consistent with (Hasan et al., 2021; Mai et al., 2023a), OpenFace 2 (Baltrusaitis et al., 2018) provides facial action units together with rigid and non-rigid facial-shape parameters. On the audio-visual event benchmark AVE, which has no language stream, frames uniformly sampled from each video are encoded by an ImageNet-pretrained ResNet-18 (He et al., 2016). Temporal pooling of the frame features yields a 512-dimensional visual descriptor per clip.

(2) Acoustic modality: On CH-SIMS-v2 the acoustic stream is encoded by the 25-dimensional eGeMAPS set of low-level descriptors (LLDs), computed with OpenSmile (Eyben, 2010) at a 16 kHz sampling rate. On every other benchmark we instead use COVAREP (Degottex et al., 2014), whose acoustic sequence comprises 12 Mel-frequency cepstral coefficients, pitch tracking, speech polarity and related cues. For the AVE dataset, the waveform is converted to a log-Mel spectrogram, which is treated as a single-channel image and fed to a ResNet-18 with a one-channel input stem (**?**), producing a 512-dimensional acoustic descriptor.

(3) Language modality: On MOSI and MOSEI we follow leading methods (Xiao et al., 2024; Fang et al., 2025) and encode text with DeBERTa (He et al., 2021); on CH-SIMS-v2 we use BERT (Devlin et al., 2019) as in (Feng et al., 2024; Liu et al., 2022); on MHD and MSD we follow (Mai et al., 2023c;a) and use ALBERT (Lan et al., 2020).

[1] iMotions 2017. https://imotions.com/

Table 7: The feature dimensionality of various datasets.

Dataset	Language	Acoustic	Visual
MOSI	768	74	47
MOSI (OOD)	768	74	47
MOSEI	768	74	35
CH-SIMS-v2	768	25	177
UR-FUNNY	768	60	36
MUStARD	768	60	36
AVE	—	512	512

Table 8: Hyperparameter settings of ReFuse per dataset, using the notation of Section 3.

	MOSI	MOSEI	CH-SIMS-v2	UR-FUNNY	MUStARD	AVE
Batch size	64	48	32	48	64	64
Learning rate	10^{-5}	10^{-5}	5×10^{-4}	1×10^{-6}	2×10^{-6}	10^{-3}
Common dimension d	150	150	100	50	50	256
Max recursion steps $T_{\max}$	4	6	4	6	4	4
Training warm-up $T_{\min}$	2	2	2	2	2	2
Selection threshold τ	0.02	0.001	0.001	0.02	0.02	0.01
Stopping threshold ϵ	0.01	0.001	0.005	0.01	0.01	0.01
Prediction-stability η	0	0	0	0	0	0
Step-supervision weight α	1	0.5	0.3	1.0	0.3	0.5
Budget weight β	0.001	0.1	0.01	0.001	0.001	0.01
Monotonicity weight γ	0.001	0.005	0.5	0.1	0.1	1.0
Diagnoser-distillation weight δ	0.01	0.5	0.5	1.0	0.1	0.1

Table 7 lists the resulting feature dimensionalities.

E.4 Experimental Details

We implement ReFuse in PyTorch and run all experiments on a single NVIDIA RTX 3090 GPU under CUDA 11.4 and PyTorch 1.13.1.

Following (Gkoumas et al., 2021), we first define a feasible range for every hyperparameter and then run a 50-iteration random grid search on the validation set, retaining the configuration that attains the best validation loss. With the chosen configuration we retrain from scratch under five different random seeds and report the mean over multiple runs. The selected configurations are summarized in Table 8, and all models are optimized with AdamW (Loshchilov & Hutter, 2019). The shared fusion unit has multiple attention heads and a feed-forward dimension of $2d$. The student diagnoser D_ϕ is a three-linear-layer MLP (two GELU hidden layers). The conditional projections W_i' are $d \times d$.

E.5 Baselines

The baselines for MSA include:

(1) Information-Theoretic Hierarchical Perception (**ITHP**) (Xiao et al., 2024): Grounded in the information bottleneck, ITHP designates one modality as the core and treats the remaining modalities as detectors along an information pathway, filtering and refining the transmitted information stage by stage.

(2) Self-Supervised Multi-task Multimodal sentiment analysis network (**Self-MM**) (Yu et al., 2021): Self-MM derives unimodal sentiment labels from the multimodal sample label in a self-supervised manner, encouraging each modality to learn a more discriminative representation.

(3) Disentangled-Language-Focused Model (**DLF**) (Wang et al., 2025): A disentanglement module separates modality-shared from modality-specific content, and four geometric constraints further suppress redundancy in favour of language-centric features; a language-targeted attractor then enriches the language representation with complementary cues drawn from the other modalities.

(4) Gradient and Structure Consistency (**GSCon**) (Shi et al., 2025): GSCon enforces two kinds of consistency: aligning per-modality gradient directions for balanced optimization, and preserving the cross-modal spatial structure of samples to reduce interaction noise induced by multimodal alignment.

(5) **Diffusion Bridge** (Lee et al., 2025): This method closes the modality gap directly with denoising diffusion probabilistic models.

(6) Enhanced Dynamic Emotion Experts (**EMOE**) (Fang et al., 2025): EMOE combines a mixture of modality experts whose contribution is adjusted dynamically per input with a unimodal distillation scheme that preserves each modality's predictive ability inside the fused representation.

(7) **DecAlign** (Qian et al., 2026): DecAlign proposes a hierarchical cross-modal alignment framework that splits multimodal features into modality-heterogeneous and modality-homogeneous components, applying dedicated strategies to capture modality-specific characteristics and shared semantics, respectively.

(8) Optimal Multimodal Information Bottleneck (**OMIB**) (Wu et al., 2025b): OMIB sets the compression coefficient of each modality adaptively, so as to retain all task-relevant information.

(9) Acoustic Visual Mix-up Consistent (**AV-MC**) (Liu et al., 2022): AV-MC augments acoustic and visual representations through modality mix-up, strengthening their contribution to sentiment prediction.

(10) Multimodal Adaptation Gate BERT/ALBERT (**MAG-BERT/MAG-ALBERT**) (Rahman et al., 2020): A dedicated multimodal adaptation gate injects visual and acoustic signals directly into BERT/ALBERT.

(11) **MISA** (Hazarika et al., 2020): Each unimodal input is factorised into a modality-invariant and a modality-specific component, which are then fused to form the final prediction.

(12) MultiModal InfoMax (**MMIM**) (Han et al., 2021): MMIM enhances representation learning by maximising mutual information both within unimodal features and between hierarchical multimodal representations and their unimodal counterparts.

(13) Personality-Sentiment Aligned Multi-level Fusion (**PSA-MF**) (Xie et al., 2026): PSA-MF incorporates personality traits during feature extraction to obtain personalised sentiment representations, and combines modalities through a hierarchical fusion scheme.

(14) Multimodal Conditional Information Bottleneck (**MCIB**) (Wang et al., 2026): MCIB imposes a pairwise conditional IB over modality pairs, minimising $I(x_p; b)$ while maximising $I(b; y \mid x_a)$ to remove cross-modal redundancy and preserve complementary information for sarcasm detection.

(15) Complete Multimodal Information Bottleneck (**C-MIB**) (Mai et al., 2023b): Built on the information bottleneck, C-MIB copes with noisy modalities by removing redundancy and noise from both unimodal and multimodal features.

(16) Two-stage Modality Denoising and Complementation (**TMDC**) (Zhuang et al., 2026): TMDC proceeds in two phases: an intra-modality denoising stage extracts clean modality-specific and shared representations from complete data for robustness, and an inter-modality complementation stage then leverages these representations to compensate for missing modalities.

The additional baselines for MHD and MSD include:

(1) Multimodal Global Contrastive Learning (**MGCL**) (Mai et al., 2023c): MGCL applies supervised contrastive learning to multimodal representations, constructing positive and negative pairs through a suite of augmentation techniques.

(2) Multimodal Correlation Learning (**MCL**) (Mai et al., 2023a): MCL imposes a supervised correlation objective that preserves modality-specific properties while yielding a more discriminative joint embedding space.

(3) Multimodal Multitask Interaction Learning (**MIL**) (Zhang et al., 2024b): MIL jointly detects sarcasm and sentiment, using cross-modal target attention to align textual content with the visual and acoustic signals, together with an interaction module that captures both shared and task-specific patterns.

(4) Decoupled Multimodal Distillation (**DMD**) (Li et al., 2023b): DMD decomposes unimodal features into modality-irrelevant and modality-exclusive subspaces and transfers knowledge across them with a graph distillation unit, enabling adaptive cross-modal transfer.

(5) Humor Knowledge Enriched Transformer (**HKT**) (Hasan et al., 2021): HKT feeds humour-specific external knowledge into the language modality to resolve ambiguity and capture subtle sentiment cues.

(6) Adaptive Gradient Modulation (**AGM**) (Li et al., 2023a): AGM modulates gradients adaptively to strengthen multimodal models under different fusion strategies; it introduces a unimodal concept for competition-free states and new metrics of modality competition to analyse the conflicts that arise during fusion.

(7) Uncertain Cross-Modal Information Bottleneck and Probability of Necessity and Sufficiency (**UCMIB-PNS**) (Chen et al., 2026): UCMIB-PNS pairs a variational-IB UCMIB module that curbs redundancy during fusion with a counterfactual PNS estimator that measures necessity and sufficiency, and reweights modalities to balance the two quantities.

(8) Multimodal Learning using Optimal Transport (**MuLOT**) (Pramanick et al., 2022): MuLOT uses self-attention to model intra-modal correspondences and optimal transport to model cross-modal correspondences, applying multimodal attention fusion to capture interdependence among modalities.

(9) Multimodal Adaptation Gate XLNet (**MAG-XLNet**) (Rahman et al., 2020): The Multimodal Adaptation Gate (MAG) conditions the internal representations of BERT and XLNet on the visual and acoustic streams, enabling them to ingest non-linguistic modalities during fine-tuning.

(10) Subject Causal Intervention module (**SuCI**) (Xu et al., 2025): SuCI is a lightweight causal-intervention module that removes the confounding effect of subjects treated as unobserved confounders, so that training relies on genuine causal effects.

(11) Attention-based Causality-Aware Fusion (**AtCAF**) (Huang et al., 2025a): AtCAF builds a front-door Causality-Aware Text Debiasing Module (CATDM) for causality-aware text representations, and a Counterfactual Cross-modal Attention (CCoAt) module that injects causal cues into multi-modal fusion to improve representation aggregation.

Following the unified single-backbone benchmark of Qin et al. (2026), we compare ReFuse with ten methods covering vanilla fusion and the dominant lines of modality-imbalance learning on the audio-visual event recognition task:

(1) On-the-fly Gradient Modulation (**OGM-GE**) (Peng et al., 2022): OGM-GE tracks each modality's optimization progress through the ratio of unimodal ensemble logits and on-the-fly down-weights the dominant modality's gradients so that the weak modality keeps pace.

(2) Prototypical Modal Rebalance (**PMR**) (Fan et al., 2023): PMR builds per-modality prototypes and applies a prototypical rebalancing objective that aligns the optimization pace of the different modalities.

(3) Multi-Modal Cosine loss (**MMCosine**) (Xu et al., 2023): MMCosine measures the angular gap between each modality's prediction and the label with a cosine loss to dynamically re-weight modality contributions for balanced audio-visual learning.

(4) Adaptive Gradient Modulation (**AGM**) (Li et al., 2023a): AGM adaptively modulates gradients across fusion strategies, introducing a competition-free unimodal state and modality-competition metrics to resolve optimization conflicts.

(5) Information Regulation (**InfoReg**) (Huang et al., 2025b): InfoReg adaptively regularizes unimodal information acquisition so that the task-relevant information of each modality is retained in a balanced amount.

(6) **MMPareto** (Wei & Hu, 2024): MMPareto casts multimodal learning as a Pareto multi-objective problem, letting weak modalities assist the joint model with "innocent" updates that do not hurt the strong one.

(7) Multimodal Learning with Alternating unimodal adaptation (**MLA**) (Zhang et al., 2024a): MLA alternates unimodal adaptation during joint training, iteratively aligning each modality to the joint representation to alleviate modality dominance.

(8) Disentangled Gradient Learning (**DGL**) (Wei et al., 2025b): DGL disentangles each modality's gradient into shared and conflicting components and re-weights them to suppress cross-modal gradient conflict.

(9) Adaptive Rebalancing Learning (**ARL**) (Wei et al., 2025a): ARL explicitly models modality imbalance and adaptively re-balances the optimization of the strong and weak modalities during training.

(10) Information-Balanced Multimodal Learning (**IBML**) (Qin et al., 2026): grounded in a multimodal information-balance criterion, IBML maximizes a tractable lower bound for balanced complementary-information retention and modulates per-modality task complexity, representing the state of the art on this benchmark.

All baseline numbers are taken from Qin et al. (2026) under identical backbones and splits for a fair comparison.

F Additional Related Works

Dynamic modality selection. DyMo (Du et al., 2026) performs inference-time dynamic selection of recovered modalities for incomplete multimodal classification, iteratively selecting the modality with the highest estimated task-relevant information gain. However, DyMo targets the missing modality setting, operates at inference time as a post-hoc selection procedure, and selects among reconstructed modalities rather than iteratively refining a fused representation. AdaFuse (Qu et al., 2026) uses reinforcement learning for patient-specific modality selection in medical prediction, but is a sequential decision process without a fused representation that is recursively refined. DynaFuse (Alikhani et al., 2023) dynamically adjusts modality weights based on runtime quality, but does not iterate fusion. ReFuse differs in that it is a training-time fusion architecture that recursively refines the representation itself, with the diagnoser driving both selection and termination.

Information bottleneck for multimodal learning. The information bottleneck (IB) principle (Tishby et al., 2000) seeks a representation that is maximally informative about the target while compressed with respect to the input. It has been extended to multimodal settings: MIB (Mai et al., 2023b) learns minimal sufficient unimodal and multimodal representations by regularizing both, and OMIB (Wu et al., 2025b) dynamically adjusts per-modality compression coefficients. These methods focus on the compression side of the bottleneck: removing redundancy and noise in a single representation-learning pass. ReFuse instead targets the *sufficiency* side: rather than compressing a one-shot fused representation, it iteratively checks which modalities still carry task-relevant information absent from the current representation and re-fuses them. As shown in Section A.8, ReFuse can be interpreted as a *sequential* information bottleneck that spreads information acquisition over multiple selective steps.