Alignment is Not Enough: Subset-Conditioned Boundary Compensation under Missing Modalities

Zesen Cai[1] **Lisi Mo**[1] **Yifan Fang**[1] **Yandong Yan**[2]
Keren Shi[1] **Daming Shi**[3] **Ruiting Dai**[1]

[1]University of Electronic Science and Technology of China
[2]Key Laboratory of High Confidence Software Technologies,
School of Computer Science, Peking University
[3]Department of Automation, Tsinghua University
zesen.cai@std.uestc.edu.cn, morris@uestc.edu.cn, 2023090908007@std.uestc.edu.cn,
2501111974@stu.pku.edu.cn, skr@alu.uestc.edu.cn, shidm18@tsinghua.org.cn,
rtdai@uestc.edu.cn

Abstract

Missing modalities reduce observations to arbitrary subsets, challenging robust multimodal inference. Existing methods recover missing views or align cross-subset representations, implicitly equating feature completeness with decision accuracy—leaving subset-specific decision boundary variation structurally unaddressed. Empirical analysis reveals that misclassification patterns are highly subset-specific, stable across initializations, and concentrated near decision boundaries—a persistent boundary misalignment distinct from information attenuation, which we define as Subset-Conditioned Boundary Shift (SCBS). To address this, we propose Subset-Conditioned Boundary Compensation (SCBC): a modality-factorized violation memory accumulates historical records of subset-specific margin violations, and a subset-adaptive head converts them into targeted logit corrections at inference, with a direction-consistency loss enforcing alignment during training. We prove that these corrections, built entirely from forward-pass statistics without explicitly computing loss gradients, satisfy the descent condition of the surrogate loss, to our knowledge the first such guarantee in incomplete multimodal learning. Across five benchmarks, SCBC achieves state-of-the-art-level performance, averaging a 3.4 Macro-F1 point gain under random missingness and peaking at 7.4 points on the IEMOCAP-4 $\{v, a\}$ setting. Source code is available at https://github.com/Js183/SCBC.

1 Introduction

Real-world multimodal data is rarely complete: sensor failures, privacy constraints, or bandwidth limitations frequently render certain modalities unavailable, requiring models to maintain reliable predictions under arbitrary modality combinations (observed subsets)—a challenge pervasive across multimodal classification, sentiment analysis, and rumor detection.

Existing works address this along two lines—recovering missing content through reconstruction or generation [1–4], or promoting cross-subset feature alignment through robust fusion and correlation-style regularization [5–9]. Both implicitly assume that feature-level semantic completeness translates into decision-level accuracy. This assumption need not hold: a single shared decision boundary, optimized through a global objective, receives no explicit constraint on the conflicting local manifold geometries induced by different modality subsets, so subset-specific boundary misalignments can persist even when feature alignment is largely achieved. Methods that introduce subset-aware

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

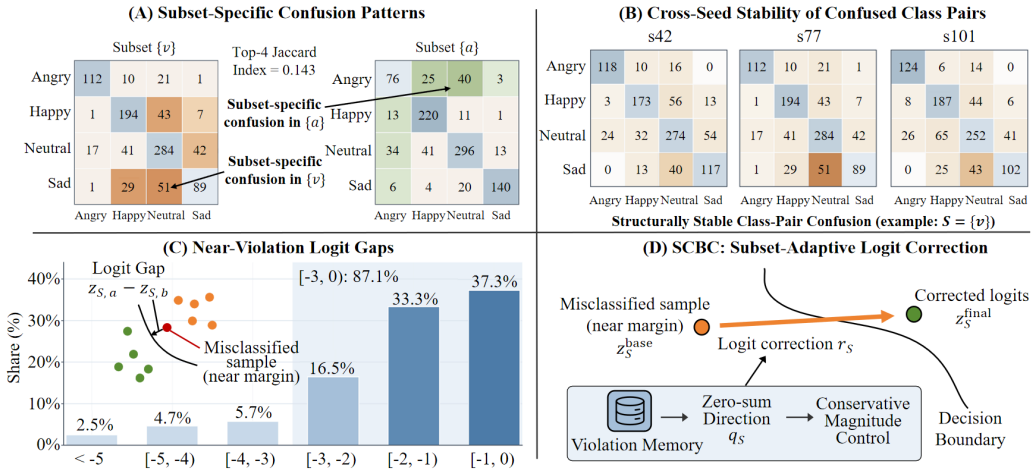

Figure 1: **Empirical diagnosis of subset-conditioned boundary shift.** (a) subset-specific variation in confused class-pair patterns between $\{v\}$ and $\{a\}$ (subset exclusivity), (b) persistent high-risk class-pair errors across random seeds (structural stability), (c) concentration of negative logit gaps near the decision boundary (boundary proximity), and (d) the SCBC correction pathway from violation memory to conservative logit displacement.

guidance [8–11] modulate all class outputs jointly through the classifier matrix, leaving subset-specific decision boundary variation structurally unaddressed.

Subset-Conditioned Boundary Shift (SCBS). Different modality subsets naturally induce different optimal decision boundaries—a consequence of distributional heterogeneity. This paper identifies a qualitatively distinct failure mode. If misclassification arose primarily from information attenuation, failure patterns would be near-random: confused class pairs would fluctuate with training randomness, and similar confusion patterns would recur across subsets sharing the same signal. Figure 1 diagnoses this on IEMOCAP-4 and reveals three complementary patterns.

Subset exclusivity. The top-4 confused class-pair sets under $\{v\}$ and $\{a\}$ overlap minimally (Jaccard index 0.143); since intrinsic class discriminability does not vary with the observed subset, this specificity points to a geometric cause rather than an overall representational deficit.

Structural stability. The failure pattern is highly stable across random initializations (mean cross-seed correlation 0.875, significantly above permutation-test baselines), ruling out optimization noise.

Boundary proximity. 87.1% of misclassified samples have violation margins concentrated in $[-3, 0)$; critically, errors cluster on the specific class pairs identified above rather than scattering uniformly, inconsistent with information attenuation, which would produce high-entropy unstructured failures. This concentration also empirically supports the local-linearity condition underlying our theoretical analysis (Assumption 1(A3) in §3.3).

We formally define this phenomenon as **Subset-Conditioned Boundary Shift (SCBS)**; see §2 for a precise characterization.

Since SCBS manifests as a localized logit-level misalignment, a natural remedy is a conservative, directed displacement of the model's output. While it is standard that a perturbation forming a positive inner product with the negative loss gradient reduces expected loss, the non-trivial contribution lies in proving that $\mathbf{q}_S$, built purely from forward-pass signals without explicitly computing loss gradients, satisfies the descent condition of the surrogate loss—a connection that does not follow from standard optimization arguments. The proof relies on the *near-violation condition*: violation margins concentrate near the target threshold so that the surrogate loss activation slope is approximately constant on the active set, empirically supported by Figure 1(c). To our knowledge, this is the first such result in incomplete multimodal learning. Grounded in this guarantee, we propose **Subset-Conditioned Boundary Compensation (SCBC)**, which accumulates historical margin violations in a modality-factorized memory at the (modality, class-pair) granularity and converts retrieved profiles into conservative logit corrections via a subset-adaptive displacement head with adaptive gating. A

direction-consistency loss enforces the descent condition end-to-end during training; Figure 1(d) illustrates the full correction pathway.

Contributions. (*i*) We identify and define SCBS—a failure mode characterized by subset exclusivity, structural stability, and boundary proximity, orthogonal to information attenuation. (*ii*) We prove (Lemma 3) that $\mathbf{q}_S$, built from forward-pass statistics without explicitly computing loss gradients, satisfies the descent condition of the surrogate loss under the near-violation assumption, independent of the training objective and not derivable from standard descent arguments; Proposition 2 provides a complementary empirical proxy verifiable during training. (*iii*) We propose SCBC, instantiating the above principle as a modality-factorized violation memory with a sample-adaptive gating displacement head, jointly trained with a direction-consistency loss. (*iv*) Across five benchmarks under both protocols, SCBC outperforms strong baselines in the large majority of settings, averaging **3.4 Macro-F1 points** under random missingness and peaking at **7.4 points** on IEMOCAP-4 $\{v, a\}$, while preserving full-modality performance.

Related Work. Recovery-based methods [1–4] reconstruct absent views before fusion, but performance degrades under severe or selective missingness. Non-recovery methods improve feature robustness through multi-view correlation, distillation, or prompting [5–9, 11–18], yet leave subset-specific decision boundary variation structurally unaddressed. Decision-aware methods [8, 10, 11] introduce subset-aware guidance at the output level but modulate all class outputs jointly through the classifier matrix, without tracking per-(subset, class-pair) violation history or decoupling correction direction from step size—gaps that SCBC directly addresses. A detailed discussion is in Appendix A.

2 Preliminaries

Notation and Problem Setup. Let $\mathcal{M} = \{1, \ldots, M\}$ be the universe of modalities. For a sample i drawn from a data distribution $\mathcal{D}$, let $S_i \subseteq \mathcal{M}$ denote its observed subset of modalities, $x_i = \{x_i^{(m)}\}_{m \in S_i}$ be the input, and $y_i \in \{1, \ldots, C\}$ be the true label. The multimodal model maps the current observation to a logit vector $\mathbf{z}_S \in \mathbb{R}^C$, and the standard cross-entropy loss is $\mathcal{L}_i(\mathbf{z}_S) = -\log \mathrm{softmax}(\mathbf{z}_S)_{y_i}$. All expectations $\mathbb{E}[\cdot]$ and probabilities $\mathbb{P}(\cdot)$ are evaluated over $\mathcal{D}$, and we write $\epsilon > 0$ for a generic small numerical-stability constant used throughout.

Formalization of SCBS. Motivated by the empirical observations in Fig. 1, we formalize boundary failure intensity at the class-pair level rather than at the sample level, as this granularity directly captures the structured, subset-specific confusion identified in §1. For an observed subset S, ground-truth class a, and competing class $b \neq a$, the expected class-pair margin violation is

$$V_S(a, b) = \mathbb{E}\Big[\big(\tau_m - (z_{S,a} - z_{S,b})\big)_+ \,\Big|\, y_i = a,\, S_i = S\Big], \tag{1}$$

where $\tau_m > 0$ is the target margin, $z_{S,c}$ denotes the c-th component of $\mathbf{z}_S$, and $(x)_+ = \max(x, 0)$. A triplet (S, a, b) constitutes an instance of SCBS if all of the following hold: (i) **persistence** (the set $\mathcal{E} = \{(S, a, b) : V_S(a, b) > 0\}$ is stable across random seeds); (ii) **subset-specificity** (there exists $S' \neq S$ such that $V_{S'}(a, b) = 0$ while $V_S(a, b) > 0$); and (iii) **near-violation** (on violating samples, the logit gap $z_{S,a} - z_{S,b}$ concentrates around the target margin τ_m, as empirically supported by Fig. 1(c) and formalized as Assumption 1(A3)).

The *excess violation intensity* matrix is $\mathbf{\Delta}_S \in \mathbb{R}^{C \times C}_{\geq 0}$ with entries $\Delta_S(a, b) = (V_S(a, b) - V_{\mathcal{M}}(a, b))_+$, where $V_{\mathcal{M}}$ is the violation under full observation. This matrix is the core estimation target of our compensation mechanism.

Logit Perturbation Perspective. To mitigate SCBS, we seek an additive perturbation $\mathbf{r} \in \mathbb{R}^C$ such that injecting it into the logit output reduces the expected loss.

Definition 1 (Gradient-Related Condition). A perturbation $\mathbf{r} \in \mathbb{R}^C$ is gradient-related to $\mathcal{L}_i$ under an event $\mathcal{C}$ with $\mathbb{P}(\mathcal{C}) > 0$ if there exists $\kappa > 0$ such that, almost surely under $\mathcal{C}$,

$$\langle \nabla_{\mathbf{z}} \mathcal{L}_i(\mathbf{z}_S), \mathbf{r} \rangle \leq -\kappa \, \|\mathbf{r}\|_2^2. \tag{2}$$

In §3.3, $\mathcal{C}$ is instantiated as $\mathcal{E}_S = \{\|\mathbf{q}_S\|_2 > \delta\}$.

Lemma 1 (Expected Loss Descent). Let $\mathcal{L}_i$ be the cross-entropy loss. The Hessian of $\mathcal{L}_i$ in the logit space has spectral norm at most $1/2$, so $\mathcal{L}_i$ is L-smooth with $L = 1/2$ (Appendix B). If $\mathbf{r}$ is

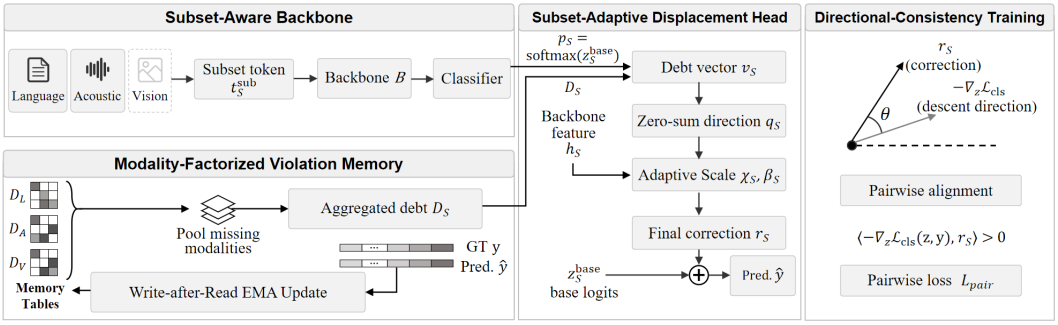

Figure 2: **Overall framework of SCBC.** The subset-aware backbone produces base logits and subset features; the violation memory stores historical excess violations relative to the full-observed reference view; and the displacement head converts retrieved debt profiles into conservative subset-adaptive logit correction.

gradient-related to $\mathcal{L}_i$ under $\mathcal{C}$ with constant $\kappa > 0$, then for any step size $\alpha \in (0, \min\{1, 4\kappa\}]$,

$$\mathbb{E}[\mathcal{L}_i(\mathbf{z}_S + \alpha\mathbf{r}) \mid \mathcal{C}] \leq \mathbb{E}[\mathcal{L}_i(\mathbf{z}_S) \mid \mathcal{C}] - \alpha\big(\kappa - \tfrac{\alpha}{4}\big)\mathbb{E}[\|\mathbf{r}\|_2^2 \mid \mathcal{C}]. \tag{3}$$

The derivation follows the classical descent lemma [19] applied pointwise with $L = 1/2$, followed by conditional expectation. Proof in Appendix B.

Lemma 1 reduces expected-loss descent to constructing $\mathbf{r}$ that satisfies Definition 1. Building on this, §3 treats the network output $\mathbf{z}_S$ as the base prediction and constructs an adaptive compensation displacement $\mathbf{r}_S$ to serve as this valid perturbation.

3 Method

SCBC Framework. We propose Subset-Conditioned Boundary Compensation (**SCBC**), shifting the perspective from feature-level recovery and alignment to decision-space boundary compensation. For an active subset S, SCBC extracts a base prediction $\mathbf{z}_S^{\text{base}}$ and applies a subset-specific displacement to compensate. As depicted in Figure 2, the architecture comprises three modules: a *subset-aware backbone* (§3.1) for base predictions, a *modality-factorized violation memory* (§3.2) accumulating historical shift priors, and a *subset-adaptive displacement head* (§3.3) mapping these priors to targeted logit corrections at inference time.

Train-Inference Protocol. We use a dual-view protocol: a *masked view* processes $\mathbf{s} \in \{0, 1\}^M$, while a *reference view* processes the same sample with the maximal available mask $\mathbf{s}^{\uparrow}$ (all modalities present for that sample) and is never additionally masked. The reference logits provide recoverable-margin targets and excess-violation updates; at inference, only the masked branch runs, using stored subset-conditioned priors.

3.1 Subset-Aware Backbone

For the linearized analysis in §3.3, $\mathbf{z}_S^{\text{base}}$ should ignore unobserved modalities and remain C^2-smooth in the observed ones. We enforce this at fusion time; non-dialogue datasets use the tokenized encoder below, while dialogue datasets follow the same masked/reference-view contract with dataset-specific backbones (Appendix C).

We inject the observation subset as an explicit summary token. For each modality m, tokens are produced by the modality-specific tokenizer $\phi_m(\cdot)$ combined with modality and positional embeddings; the subset state is encoded by a subset-summary token aggregating presence/absence embeddings:

$$\mathbf{T}_m = \phi_m(x^{(m)}) + \mathbf{e}_m^{\text{mod}} + \mathbf{E}_m^{\text{pos}}, \qquad \mathbf{t}_S^{\text{sub}} = \sum_{m=1}^{M} \big(s_m \mathbf{e}_m^{\text{pres}} + (1 - s_m)\mathbf{e}_m^{\text{miss}}\big), \tag{4}$$

where s_m indicates observation, $\mathbf{e}_m^{\text{pres}}, \mathbf{e}_m^{\text{miss}}$ are presence/absence embeddings, and $\mathbf{e}_m^{\text{mod}}, \mathbf{E}_m^{\text{pos}}$ are modality/position embeddings. Unobserved tokens are zeroed and masked with $-\infty$ attention scores.

The encoder processes $[\mathbf{x}_{\text{cls}}; \mathbf{t}_S^{\text{sub}}; \mathbf{T}_1; \ldots; \mathbf{T}_M]$, and the base logits are

$$\mathbf{h}_S = W_h[\mathbf{x}_{\text{cls}}^L; \mathbf{x}_{\text{sub}}^L], \qquad \mathbf{z}_S^{\text{base}} = W_c\, \mathbf{h}_S, \tag{5}$$

where $\mathbf{h}_S$ is fed to the displacement head.

3.2 Modality-Factorized Violation Memory

A direct subset-indexed table of $\mathbf{\Delta}_S$ incurs $\mathcal{O}(2^M)$ storage and suffers severe sparsity on rare subsets. We instead parameterize $\mathbf{\Delta}_S$ by M *modality-level tables* $\mathbf{D}_m \in \mathbb{R}_{\geq 0}^{C\times C}$, each tracking the violation attributable to the absence of modality m, and reconstruct the subset-level estimate by linear aggregation:

$$\mathbf{D}_S = \frac{1}{|\mathcal{M} \setminus S| + \epsilon} \sum_{m \notin S} \mathbf{D}_m, \tag{6}$$

where ϵ handles $S = \mathcal{M}$. This factorization reduces storage from $\mathcal{O}(2^M)$ to $\mathcal{O}(M)$; higher-order subset interactions are handled by the adaptive gating in §3.3; we empirically compare this factorized memory against a full subset-indexed alternative in the ablation study.

Excess violation signal. Using the class-pair indicator with bandwidth $\nu > 0$,

$$\phi_{ab}(\mathbf{z}) = \frac{\exp(z_b/\tau_m)}{\sum_{c\neq a}\exp(z_c/\tau_m)} \cdot \sigma\left(\frac{\tau_m - (z_a - z_b)}{\nu}\right), \tag{7}$$

which activates when class b is both plausible and near the target margin. For sample i with true class a and mask S_i, the per-modality excess signal is

$$e_i(b; m) = \left(\phi_{ab}(\mathbf{z}_{S,i}^{\text{base}}) - \phi_{ab}(\mathbf{z}_i^{\text{ref}})\right)_+, \quad m \notin S_i, \tag{8}$$

where $\mathbf{z}_i^{\text{ref}}$ is stop-gradient reference logits. The positive part isolates missingness-induced excess violation; with multiple missing modalities, the signal is evenly split across the missing set.

Write-after-read EMA with shrinkage. Let $\widehat{E}_t(m, a, b)$ be the batch mean of the split excess signal and $\widehat{N}_t(m, a)$ the batch-normalized support for class a when modality m is missing. We maintain EMA estimates and stabilize sparse entries with James–Stein shrinkage [20]:

$$\begin{aligned} V_t(m,a,b) &= \mu V_{t-1}(m,a,b) + (1-\mu)\widehat{E}_t(m,a,b), \\ N_t(m,a) &= \mu_N N_{t-1}(m,a) + (1-\mu_N)\widehat{N}_t(m,a), \\ D_t(m,a,b) &= V_t(m,a,b) \cdot \tfrac{N_t(m,a)}{N_t(m,a)+\nu_s}. \end{aligned} \tag{9}$$

Each iteration first reads historical $\mathbf{D}_{S_i}$ for correction, then updates V_t, N_t with detached statistics for subsequent iterations; simultaneous read-write would create a moving-target effect that undermines the independence assumption of Proposition 2. Appendix B gives the steady-state result.

3.3 Subset-Adaptive Displacement Head

Given $\mathbf{D}_S$ and the posterior $\mathbf{p}_S = \text{softmax}(\mathbf{z}_S^{\text{base}})$, the displacement head constructs a signed direction and modulates its magnitude. The debt vector and its zero-sum projection are

$$\mathbf{v}_S = \text{Diag}(\mathbf{p}_S)\mathbf{D}_S\mathbf{1} - \mathbf{D}_S^\top \mathbf{p}_S, \qquad \mathbf{q}_S = \Pi_0(\mathbf{v}_S), \tag{10}$$

where Π_0 projects onto $\{\mathbf{r} \in \mathbb{R}^C : \mathbf{1}^\top \mathbf{r} = 0\}$, removing the softmax shift-invariant component so that $\mathbf{q}_S$ lies in the same zero-sum subspace as $\nabla_{\mathbf{z}}\mathcal{L}_i$. The two terms encode posterior-weighted outgoing and incoming violation.

Assumption 1(Local Regularity). There exists $\rho_0 > 0$ such that on the ball $B(\mathbf{z}_S^{\text{base}}, \rho_0)$: **(A1)** the backbone satisfies the smoothness conditions of §3.1; **(A2)** $\mathbf{D}_S$ is a shrinkage-consistent estimator of $\mathbf{\Delta}_S$ up to the modality-factorization bias (Appendix B); **(A3)** (*near-violation*) on entries with $D_S(a,b) > 0$, the slope $\sigma'(\tau_m - (z_{S,a}^{\text{base}} - z_{S,b}^{\text{base}})) \approx s_0$ for some $s_0 > 0$, empirically supported by Fig. 1(c); this is the operative condition in the proof, while persistence and subset-specificity ensure $\mathbf{D}_S$ accumulates a reliable signal rather than noise; **(A4)** $\mathcal{L}_i$ is locally strictly convex on the zero-sum

subspace in the considered finite-logit neighborhood; **(A5)** on the active event $\mathcal{E}_S = \{\|\mathbf{q}_S\|_2 > \delta\}$, the true-label debt row dominates non-true-label rows in the posterior-weighted sense, so that $\mathbf{q}_S$ points away from historically confusing competitors (formalized in Appendix B).

Theorem 1 (Linearized surrogate descent direction). Define the excess-violation-weighted surrogate

$$\mathcal{R}_S(\mathbf{r}) = \sum_{a \neq b} p_{S,a} D_S(a,b) \cdot \mathrm{softplus}\big(\tau_m - [(z^{\mathrm{base}}_{S,a} + r_a) - (z^{\mathrm{base}}_{S,b} + r_b)]\big). \tag{11}$$

Under Assumption 1, conditioned on the non-vanishing event $\mathcal{E}_S = \{\|\mathbf{q}_S\|_2 > \delta\}$ for some $\delta > 0$, the linearized steepest-descent direction of $\mathcal{R}_S$ on the zero-sum unit sphere coincides with $\mathbf{q}_S/\|\mathbf{q}_S\|_2$ up to the near-violation approximation of (A3). Proof in Appendix B.

Lemma 3 (Positive alignment with the classification risk). Under Assumption 1, including the local risk-compatibility clause (A5), there exists $c_0 > 0$ such that on $\mathcal{E}_S$

$$\langle -\nabla_{\mathbf{z}} \mathcal{L}_i(\mathbf{z}^{\mathrm{base}}_S), \mathbf{q}_S \rangle \geq c_0 \|\mathbf{q}_S\|_2^2 \quad \text{a.s. on } \mathcal{E}_S. \tag{12}$$

Consequently,

$$\mathbb{E}\big[\langle -\nabla_{\mathbf{z}} \mathcal{L}_i(\mathbf{z}^{\mathrm{base}}_S), \mathbf{q}_S \rangle \mid \mathcal{E}_S\big] \geq c_0 \, \mathbb{E}\big[\|\mathbf{q}_S\|_2^2 \mid \mathcal{E}_S\big]. \tag{13}$$

Combining this with Lemma 1, the unnormalized debt perturbation $\mathbf{q}_S$ satisfies the gradient-related condition with $\kappa = c_0$. Therefore, for any $\alpha \in (0, \min\{1, 4c_0\})$, the displacement $\mathbf{z}^{\mathrm{base}}_S + \alpha \mathbf{q}_S$ yields strict expected descent on $\mathcal{E}_S$. If the normalized unit direction $\mathbf{q}_S/\|\mathbf{q}_S\|_2$ is used instead, then on $\mathcal{E}_S$ the corresponding constant is $c_0\delta$, and the admissible interval becomes $\alpha \in (0, \min\{1, 4c_0\delta\})$. Proof in Appendix B.

Concentration-scaled conservative correction. Lemma 3 motivates shrinking the raw displacement whenever the retrieved debt profile is diffuse or unreliable. In our implementation, the conservative controller is the concentration term

$$\chi_S = (1 - \mathcal{H}(\bar{\mathbf{u}}_S)/\log C)^{\omega}, \qquad \bar{\mathbf{u}}_S = \frac{|\mathbf{v}_S|}{\|\mathbf{v}_S\|_1 + \epsilon}, \tag{14}$$

where $\mathcal{H}(\cdot)$ is Shannon entropy and $\omega \geq 1$ sharpens concentration. A context vector

$$\mathbf{c}_S = [\mathbf{h}_S; \mathbf{s}; \mathbf{q}_S; |\mathbf{v}_S|; \log(\|\mathbf{v}_S\|_1 + \epsilon)] \tag{15}$$

is then fed to a small MLP to produce a per-class step modulator bounded by $\beta_{\max} \in (0, 1]$:

$$\boldsymbol{\beta}_S = \beta_{\max} \cdot \tanh(\mathrm{softplus}(\mathrm{MLP}_{\mathrm{step}}(\mathbf{c}_S))) \in [0, \beta_{\max}]^C. \tag{16}$$

The final correction is

$$\mathbf{r}_S = \Pi_0\big(\lambda_d \cdot \chi_S \cdot (\boldsymbol{\beta}_S \odot \mathbf{q}_S)\big), \qquad \mathbf{z}^{\mathrm{final}}_S = \mathbf{z}^{\mathrm{base}}_S + \mathbf{r}_S, \tag{17}$$

where $\lambda_d > 0$ is a global scale and $\odot$ is element-wise multiplication.

Proposition 1 (Bounded, safe-step compensation). The displacement satisfies $\|\mathbf{r}_S\|_\infty \leq 2\lambda_d \chi_S \beta_{\max} \|\mathbf{q}_S\|_\infty$ (Appendix B), and $\mathbf{r}_S = \mathbf{0}$ whenever $\chi_S = 0$ or $\mathbf{q}_S = \mathbf{0}$. For the unnormalized debt-direction step $\alpha_S \mathbf{q}_S$, $\alpha_S \leq 2\sqrt{C}\lambda_d \beta_{\max}$; hence $\lambda_d \beta_{\max} \leq \min\{1, 4c_0\}/(2\sqrt{C})$ is sufficient for the Lemma 1 safe-step interval, where $c_0 > 0$ is the alignment constant from Lemma 3. Proof in Appendix B.

3.4 Directional-Consistency Loss and Training

While Lemma 3 establishes the descent guarantee analytically, we introduce a direction-consistency loss to maintain this property throughout training and provide an empirically verifiable proxy via Proposition 2. Let $m^{\mathrm{mask}}_i(b) = z^{\mathrm{base}}_{S,i,y_i} - z^{\mathrm{base}}_{S,i,b}$ and $m^{\mathrm{ref}}_i(b) = z^{\mathrm{ref}}_{i,y_i} - z^{\mathrm{ref}}_{i,b}$ denote masked- and reference-view margins. We define class-pair weights

$$w_i(b) = \frac{D_{S_i}(y_i,b) + \epsilon_w/(C-1)}{\sum_{c \neq y_i} D_{S_i}(y_i,c) + \epsilon_w}, \tag{18}$$

where $\epsilon_w > 0$ is a Laplace smoothing constant.

The pairwise loss prioritizes debt-weighted class pairs:

$$\mathcal{L}_{\mathrm{pair}} = \tfrac{1}{B} \textstyle\sum_i \sum_{b \neq y_i} w_i(b) \cdot \mathrm{softplus}(\tau_m - (z^{\mathrm{final}}_{i,y_i} - z^{\mathrm{final}}_{i,b})). \tag{19}$$

Table 1: **IEMOCAP-4 results under fixed and random missingness.** The table reports F1 scores. Methods are listed by rows; the two column groups correspond to fixed-missing modality subsets and random-missing rates R_m.

Method	Fixed Missingness (F1)							Random Missingness (F1)							
	L	**V**	**A**	**LV**	**LA**	**VA**	**LVA**	**0.0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**
GCNet [2]	76.1	61.6	63.5	77.4	79.1	65.4	82.7	78.4	77.5	77.3	76.2	75.1	73.8	71.9	71.4
MPLMM [13]	69.3	57.0	59.7	74.5	75.4	67.4	72.0	72.0	70.8	67.9	67.4	66.6	66.2	66.0	65.7
IMDer [3]	67.2	70.1	70.2	72.6	71.7	72.3	75.4	75.4	72.8	71.6	70.8	70.8	68.2	67.0	65.3
DiCMoR [4]	68.2	68.8	68.7	72.1	76.5	75.2	76.5	76.5	73.7	72.5	71.6	69.0	68.5	68.0	67.5
LNLN [12]	74.4	69.9	74.6	75.5	75.4	74.9	77.4	77.4	76.3	75.7	75.4	74.0	72.9	72.6	70.2
HME [9]	78.1	70.8	**76.8**	78.9	80.0	76.3	80.6	80.6	80.3	80.0	77.8	77.3	77.1	76.1	76.0
SCBC	**78.9**	**71.6**	74.6	**82.9**	**81.5**	**83.7**	**85.0**	**85.0**	**84.4**	**83.7**	**82.1**	**79.8**	**79.6**	**78.9**	**78.5**

Table 2: **Main results on CMU-MOSEI under the fixed-missing protocol.** Each entry is Acc_2/F1/Acc_7.

L / V / A	$\{a\}$ ✘ ✘ ○	$\{v\}$ ✘ ○ ✘	$\{l\}$ ○ ✘ ✘	$\{va\}$ ✘ ○ ○	$\{la\}$ ○ ✘ ○	$\{lv\}$ ○ ○ ✘	$\{lva\}$ ○ ○ ○
CCA [5]	54.8 / 55.0 / 39.5	50.7 / 47.2 / 38.3	78.2 / 77.5 / 46.8	58.9 / 57.8 / 40.1	79.7 / 78.6 / 46.3	78.9 / 79.3 / 46.0	79.8 / 79.6 / 47.3
DCCA [6]	62.0 / 50.2 / 41.1	61.9 / 55.7 / 41.3	79.7 / 79.5 / 47.0	63.4 / 56.9 / 41.5	79.5 / 79.2 / 46.7	80.3 / 79.7 / 46.6	80.7 / 80.9 / 47.7
DCCAE [7]	61.4 / 53.8 / 40.9	61.1 / 57.2 / 40.1	78.5 / 78.7 / 46.7	62.7 / 59.2 / 41.6	80.0 / 80.0 / 47.4	80.4 / 80.4 / 47.1	81.2 / 81.2 / 48.2
CRA [1]	51.5 / 52.3 / 39.8	48.4 / 44.3 / 37.5	75.2 / 75.0 / 45.8	55.7 / 54.2 / 39.8	76.0 / 75.5 / 46.2	76.5 / 76.2 / 45.7	76.4 / 76.8 / 46.8
MCTN [21]	62.7 / 54.5 / 41.4	62.6 / 57.1 / 41.6	82.6 / 82.8 / 50.2	63.7 / 62.7 / 42.1	83.5 / 83.3 / 50.7	83.2 / 83.2 / 50.4	84.2 / 84.2 / 51.2
MMIN [22]	58.9 / 59.5 / 40.4	59.3 / 60.0 / 40.7	82.3 / 82.4 / 51.4	63.5 / 61.9 / 41.8	83.7 / 83.3 / 52.0	83.8 / 83.4 / 51.2	84.3 / 84.2 / 52.4
GCNet [2]	60.2 / 60.3 / 41.1	61.9 / 61.6 / 41.7	83.0 / 83.2 / 51.2	64.1 / 57.2 / 42.0	84.3 / 84.4 / 51.3	84.3 / 84.4 / 51.1	85.2 / 85.1 / 51.5
IMDer [3]	63.8 / 60.6 / 41.7	63.9 / 63.6 / **42.6**	84.5 / 84.5 / 52.5	64.9 / 63.5 / 42.8	85.1 / 85.1 / 53.1	85.0 / 85.0 / 53.1	85.1 / 85.1 / 53.4
LNLN [12]	62.9 / 61.8 / 41.4	64.5 / 60.8 / 42.1	84.1 / 84.4 / 52.7	64.8 / 62.7 / 42.3	85.6 / 85.4 / 52.2	85.8 / 85.5 / 52.4	85.5 / 85.2 / 53.3
TF-Mamba [16]	62.9 / 48.5 / 41.4	61.8 / 60.3 / 37.4	83.7 / 83.6 / 50.0	60.7 / 61.2 / 36.4	83.1 / 82.7 / 50.4	83.3 / 83.2 / 49.6	83.8 / 83.7 / 52.3
HME [9]	64.1 / 59.7 / 41.5	62.9 / 62.9 / 42.5	85.4 / 85.4 / 52.8	65.4 / 63.7 / 42.6	85.6 / 85.4 / 53.0	85.7 / 85.7 / 53.3	86.2 / 86.2 / 53.3
SCBC (Ours)	**70.5 / 68.8 / 42.2**	**67.3 / 66.3 / 42.6**	**86.7 / 86.6 / 53.5**	**69.8 / 68.7 / 43.5**	**86.9 / 86.8 / 54.5**	**86.8 / 86.7 / 54.6**	**86.8 / 86.7 / 55.0**

The complete objective combines task supervision on the base and corrected logits with the pairwise alignment term:

$$\mathcal{L} = \lambda_{\text{cls}}\text{CE}(\mathbf{z}_S^{\text{final}}, y) + \lambda_b\text{CE}(\mathbf{z}_S^{\text{base}}, y) + \lambda_p\mathcal{L}_{\text{pair}}. \tag{20}$$

The base-view CE anchors the backbone and keeps $\mathbf{D}_S$ historical; dataset-specific coefficients are specified in the corresponding code configurations.

Proposition 2 (Empirical proxy for the gradient-related condition). Under Assumption 1, if optimization achieves $\mathcal{L}_{\text{pair}} < \mathcal{L}_{\text{pair}}|_{\mathbf{r}_S=\mathbf{0}}$, then on $\mathcal{E}_{S_i}$,

$$\mathbb{E}\big[\langle \nabla_{\mathbf{z}}\mathcal{L}_i(\mathbf{z}_{S_i}^{\text{base}}), \mathbf{r}_{S_i}\rangle \mid \mathcal{E}_{S_i}\big] \le -c_0\,\mathbb{E}\big[\|\mathbf{r}_{S_i}\|_2^2 \mid \mathcal{E}_{S_i}\big], \tag{21}$$

corresponding to the expected analogue of Definition 1. Protocol and empirical verification are given in Appendix B.

4 Experiments

4.1 Experimental Setup

Datasets and protocols. We evaluate on five benchmarks. CMU-MOSEI [23] and IEMOCAP [24] follow fixed missingness over seven subsets (l/v/a = language/vision/audio); Weibo-17 [25] and PHEME [26, 27] follow random missingness; Hateful Memes [28] follows SyP [11] at rates $\{50\%, 70\%, 90\%\}$ for three missing cases. We report Acc_2/F1/Acc_7, Macro-F1, Acc, and AUROC respectively; the main results are averaged over 5 runs with mean$\pm$std tables in the appendix.

Implementation details. We follow the encoders used by the corresponding protocols: wav2vec-large [29], DeBERTa-large [30], and MA-Net [31] for CMU-MOSEI/IEMOCAP; BERT [32] and ResNet-34 [33] for Weibo-17/PHEME, with PHEME additionally using comments; and CLIP ViT-B/16 [34] for Hateful Memes. All experiments run on a single NVIDIA A100 GPU with AdamW [35]. The dataset-specific hyperparameter settings are specified in the corresponding code configurations; further implementation details are given in Appendix C.

Baselines. We compare against GCNet [2], IMDer [3], DiCMoR [4], LNLN [12], HME [9], and MPLMM [13] on IEMOCAP-4/CMU-MOSEI; TriSPrompt [15], CAFE [36], GameOn [37], CLKD

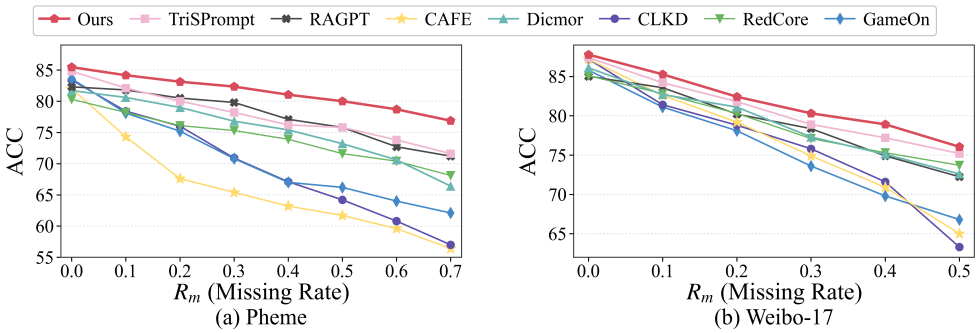

Figure 3: **Accuracy curves under the random-missing protocol.** Left: PHEME with random missing rate $R_m \in [0.0, 0.7]$. Right: Weibo-17 with $R_m \in [0.0, 0.5]$. SCBC (red) maintains a widening lead as R_m increases, supporting the view that subset-conditioned boundary shift is more severe at higher missingness rates.

Table 3: **Main results on Hateful Memes under random missingness.** Test-set AUROC. Following SyP [11], we report three missing cases × three missing rates. The green superscripts are SCBC's improvements over the runner-up.

Method	Missing text			Missing image			Both missing		
	50%	**70%**	**90%**	**50%**	**70%**	**90%**	**50%**	**70%**	**90%**
CoOp [38]	60.6	60.7	60.0	62.4	62.7	61.5	64.9	64.8	64.3
MMP [17]	60.3	61.1	57.2	62.4	63.2	61.5	65.8	65.0	63.3
MaPLe [39]	60.9	61.3	60.7	63.1	63.1	61.9	65.5	65.2	64.9
DePT [40]	61.9	61.6	61.1	63.9	63.5	62.4	65.9	65.5	65.4
DCP [18]	62.3	62.8	62.1	64.5	64.1	63.9	66.0	66.1	66.8
SyP [11]	68.3	68.9	69.7	66.8	67.0	**64.5**	68.2	68.4	68.9
HME [9]	73.5	72.8	73.4	68.9	65.1	62.3	68.7	67.1	66.2
SCBC	**77.4**$^{+3.9}$	**77.3**$^{+4.5}$	**77.5**$^{+4.1}$	**72.0**$^{+3.1}$	**67.7**$^{+0.7}$	**64.5**$^{+0.0}$	**74.8**$^{+6.1}$	**73.5**$^{+5.1}$	**71.9**$^{+3.0}$

[14], and RedCore [8] on Weibo-17/PHEME; and CoOp [38], MMP [17], MaPLe [39], DePT [40], DCP [18], SyP [11], and HME [9] on Hateful Memes (Appendix C).

4.2 Main Results

IEMOCAP-4 and CMU-MOSEI. Tables 1 and 2 summarize the main results on IEMOCAP-4 and CMU-MOSEI. On IEMOCAP-4, SCBC obtains the best F1 on six of seven fixed subsets and all random-missing rates, with the only fixed-subset exception at audio-only (74.6 vs. HME 76.8). On CMU-MOSEI, it is strongest across the seven subsets, including 6.4 and 4.4 Acc_2 gains on the language-free $\{a\}$ and $\{va\}$ settings. Additional mean±std results are provided in Appendix D.

Random missingness on Weibo-17 and PHEME. Figure 3 shows that SCBC keeps a consistent lead on both Weibo-17 and PHEME, with larger gaps at higher missing rates; Appendix D gives the corresponding mean±std tables.

Random missingness on Hateful Memes. Table 3 shows that SCBC performs strongly under random missingness, with the largest gains under both-missing and missing-text cases (up to 6.1 and 4.5 points). Appendix D reports the full SCBC Acc/F1/AUROC breakdown.

4.3 Ablation Studies

Component-level ablation (Table 4) averages IEMOCAP-4 Macro-F1 over the six incomplete subsets $\{l, v, a, lv, la, va\}$. Removing the reference-view update, displacement head, or both lowers the mean by 1.3, 1.1, and 1.6 F1, respectively, supporting both correction components.

Training-side objectives. Table 5 reports the IEMOCAP-4 training-objective ablation with $\{a\}$. Removing the pair loss and the base loss reduces Macro-F1 by 1.8 and 1.5 points, respectively, suggesting that pairwise debt supervision and the base-loss anchor are both useful.

Table 4: **Component-level ablation on IEMOCAP-4.** Mean Macro-F1 over six missing subsets; Δ is relative to Full SCBC.

Variant	Mean Macro-F1	ΔF1
Full SCBC	79.0	–
w/o reference-view update	77.7	-1.3
w/o displacement head	77.9	-1.1
Backbone only	77.4	-1.6

Table 5: **Training-objective ablation on IEMOCAP-4 with $\{a\}$.** Macro-F1; Δ is relative to Full SCBC.

Variant	Macro-F1	ΔF1
Full SCBC	**74.6**	–
w/o pair loss (λ_p=0)	72.8	-1.8
w/o base loss (λ_b=0)	73.1	-1.5

Table 6: **Cross-dataset transfer.** Average CMU-MOSEI→CMU-MOSI transfer scores; per-subset results are in Appendix D.

Method	Acc_2	F1	Acc_7
IMDer	72.7	68.5	30.9
HME	71.0	67.2	29.0
SCBC	**73.6**	**73.7**	**33.8**

Table 7: **Factorized vs. full subset-indexed memory.** Scores are averaged over $R_m \in \{0.0, \ldots, 0.7\}$.

Variant	IEMOCAP	Mem.	MOSEI	Mem.
Factorized	**81.5**	**192 B**	50.6	**588 B**
Full subset table	**81.5**	512 B	**50.7**	1.53 KiB

Cross-Dataset Generalization. Trained on CMU-MOSEI and evaluated on CMU-MOSI without fine-tuning, SCBC achieves the best average transfer result (Table 6), leading on all language-containing subsets; audio- and vision-only subsets show comparable performance to baselines, consistent with sparser violation signals under single-modality transfer. Per-subset results are in Appendix D.

Memory parameterization. Table 7 shows that the modality-factorized memory preserves the behavior of a full subset-indexed table while avoiding subset-specific storage growth.

4.4 Analysis and Verification

We empirically verify the direction-alignment properties established in Lemma 3 and Proposition 2 on five representative settings: IEMOCAP-4 $\{l\}$, $\{v\}$, $\{a\}$, Weibo-17 at $R_m = 0.5$, and PHEME at $R_m = 0.7$. Averaged across them, the non-vanishing event $\mathcal{E}_S$ occurs on 98.21% of test samples, debt-weighted sign alignment reaches 94.58%, and the active-sample fraction with $\langle \nabla_{\mathbf{z}} \mathcal{L}_i(\mathbf{z}_S^{\text{base}}), \mathbf{q}_S \rangle < 0$ reaches 85.22%. We further verify that the near-violation condition (Assumption 1(A3)) holds broadly across all five benchmarks, with a mean of 82.6% of misclassified samples having negative logit gaps in $[-3, 0)$; Appendix D provides the per-benchmark breakdown. Figure 4 confirms that the factorized memory converges rapidly, with aggregate drift entering a stable regime within the first few epochs on IEMOCAP-4. Additional visualizations, sensitivity, and efficiency analyses are reported in Appendix D. SCBC adds $\mathcal{O}(MC^2)$ memory and at most 0.0035 ms inference latency per sample over the base backbone on an A100 GPU.

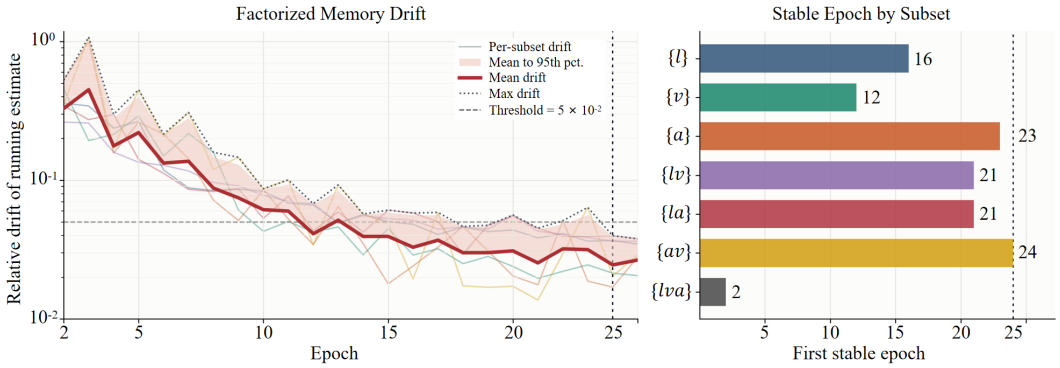

Figure 4: **Convergence of the factorized violation memory.** We track the relative drift of the running memory estimate $R^{(e)} = \frac{1}{e}\sum_{k=1}^{e} D^{(k)}$ on IEMOCAP-4 fixed-missing settings, where each $D^{(e)}$ is the complete modality-factorized memory tensor. The left panel reports per-subset drift curves and aggregate envelopes; the right panel reports the first epoch after which each subset remains below the low-drift threshold 5×10^{-2}.

5 Conclusion

We presented SCBC, a framework that shifts the perspective on missing-modality robustness from feature-level recovery and alignment to decision boundary compensation. Empirically, subset-conditioned boundary failures prove persistent, subset-specific, and boundary-proximal—a structural finding that motivates a modality-factorized violation memory and a subset-adaptive displacement head converting accumulated violations into targeted logit corrections at inference. We prove these corrections, constructed from forward-pass statistics without explicitly computing loss gradients, satisfy the descent condition of the surrogate loss, with a direction-consistency loss enforcing gradient alignment throughout training. Experiments across five benchmarks confirm consistent gains that widen as missingness increases. **Limitations.** SCBC's statistical nature may limit precision in extreme cases of rarely observed subset-class-pair combinations. Future work will explore online adaptation and extensions to broader multimodal tasks.

Acknowledgments and Disclosure of Funding

The authors declare no competing interests.

References

[1] Luan Tran, Xiaoming Liu, Jiayu Zhou, and Rong Jin. Missing modalities imputation via cascaded residual autoencoder. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1405–1414, 2017.

[2] 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. doi: 10.1109/TPAMI.2023.3234553.

[3] Yuanzhi Wang, Yong Li, and Zhen Cui. Incomplete multimodality-diffused emotion recognition. In *Advances in Neural Information Processing Systems*, volume 36, pages 17117–17128, 2023.

[4] Yuanzhi Wang, Zhen Cui, and Yong Li. Distribution-consistent modal recovering for incomplete multimodal learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22025–22034, 2023.

[5] Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28(3–4):321–377, 1936. doi: 10.1093/biomet/28.3-4.321.

[6] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1247–1255, 2013.

[7] Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes. On deep multi-view representation learning. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1083–1092, 2015.

[8] Jun Sun, Xinxin Zhang, Shoukang Han, Yu-Ping Ruan, and Taihao Li. Redcore: Relative advantage aware cross-modal representation learning for missing modalities with imbalanced missing rates. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 15173–15182, 2024. doi: 10.1609/aaai.v38i13.29440.

[9] Yan Zhuang, Minhao Liu, Wei Bai, Yanru Zhang, Wei Li, Jiawen Deng, and Fuji Ren. Hyper-modality enhancement for multimodal sentiment analysis with missing modalities. In *Advances in Neural Information Processing Systems*, 2025.

[10] Ronghao Lin, Qiaolin He, Sijie Mai, Ying Zeng, Aolin Xiong, Li Huang, Yap-Peng Tan, and Haifeng Hu. Cyin: Cyclic informative latent space for bridging complete and incomplete multimodal learning. In *Advances in Neural Information Processing Systems*, 2025.

[11] Zhihui Zhang, Luanyuan Dai, Qika Lin, Yunfeng Diao, Guangyin Jin, Yufei Guo, Jing Zhang, and Xiaoshuai Hao. Synergistic prompting for robust visual recognition with missing modalities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1881–1890, 2025.

[12] Haoyu Zhang, Wenbin Wang, and Tianshu Yu. Towards robust multimodal sentiment analysis with incomplete data. In *Advances in Neural Information Processing Systems*, volume 37, pages 55943–55974, 2024. doi: 10.52202/079017-1779.

[13] Zirun Guo, Tao Jin, and Zhou Zhao. Multimodal prompt learning with missing modalities for sentiment analysis and emotion recognition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1726–1736, 2024. doi: 10.18653/v1/2024.acl-long.94.

[14] Fan Xu, Pinyun Fu, Qi Huang, Bowei Zou, AiTi Aw, and Mingwen Wang. Leveraging contrastive learning and knowledge distillation for incomplete modality rumor detection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13492–13503, 2023. doi: 10.18653/v1/2023.findings-emnlp.900.

[15] Jiajun Chen, Yangyang Wu, Xiaoye Miao, Mengying Zhu, and Meng Xi. Trisprompt: A hierarchical soft prompt model for multimodal rumor detection with incomplete modalities. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 21687–21699, 2025. doi: 10.18653/v1/2025.findings-emnlp.1183.

[16] Xiang Li, Xianfu Cheng, Dezhuang Miao, Xiaoming Zhang, and Zhoujun Li. TF-mamba: Text-enhanced fusion mamba with missing modalities for robust multimodal sentiment analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 11252–11267, 2025. doi: 10.18653/v1/2025.findings-emnlp.602.

[17] Yi-Lun Lee, Yi-Hsuan Tsai, Wei-Chen Chiu, and Chen-Yu Lee. Multimodal prompting with missing modalities for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14943–14952, 2023.

[18] Lianyu Hu, Tongkai Shi, Wei Feng, Fanhua Shang, and Liang Wan. Deep correlated prompting for visual recognition with missing modalities. In *Advances in Neural Information Processing Systems*, volume 37, pages 67446–67466, 2024. doi: 10.52202/079017-2153.

[19] Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87 of *Applied Optimization*. Springer, Boston, MA, 2004. doi: 10.1007/978-1-4419-8853-9.

[20] William James and Charles Stein. Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 361–379. University of California Press, 1961.

[21] Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabás Póczos. Found in translation: Learning robust joint representations by cyclic translations between modalities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6892–6899, 2019. doi: 10.1609/aaai.v33i01.33016892.

[22] Jinming Zhao, Ruichen Li, and Qin Jin. Missing modality imagination network for emotion recognition with uncertain missing modalities. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2608–2618, 2021. doi: 10.18653/v1/2021.acl-long.203.

[23] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246, 2018. doi: 10.18653/v1/P18-1208.

[24] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359, 2008. doi: 10.1007/s10579-008-9076-6.

[25] Zhiwei Jin, Juan Cao, Yongdong Zhang, Jianshe Zhou, and Qi Tian. Novel visual and statistical image features for microblogs news verification. *IEEE Transactions on Multimedia*, 19(3): 598–608, 2017. doi: 10.1109/TMM.2016.2617078.

[26] Arkaitz Zubiaga, Geraldine Wong Sak Hoi, Maria Liakata, and Rob Procter. Pheme dataset of rumours and non-rumours, 2016.

[27] Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PLOS ONE*, 11(3):e0150989, 2016. doi: 10.1371/journal.pone.0150989.

[28] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. In *Advances in Neural Information Processing Systems*, volume 33, pages 2611–2624, 2020.

[29] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised pre-training for speech recognition. In *Interspeech 2019*, pages 3465–3469, 2019. doi: 10.21437/Interspeech.2019-1873.

[30] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced BERT with disentangled attention. In *International Conference on Learning Representations*, 2021.

[31] Zengqun Zhao, Qingshan Liu, and Shanmin Wang. Learning deep global multi-scale and local attention features for facial expression recognition in the wild. *IEEE Transactions on Image Processing*, 30:6544–6556, 2021. doi: 10.1109/TIP.2021.3093397.

[32] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4171–4186, 2019. doi: 10.18653/v1/N19-1423.

[33] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. doi: 10.1109/CVPR.2016.90.

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763, 2021.

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

[36] Yixuan Chen, Dongsheng Li, Peng Zhang, Jie Sui, Qin Lv, Lu Tun, and Li Shang. Cross-modal ambiguity learning for multimodal fake news detection. In *Proceedings of the ACM Web Conference 2022*, pages 2897–2905, 2022. doi: 10.1145/3485447.3511968.

[37] Mudit Dhawan, Shakshi Sharma, Aditya Kadam, Rajesh Sharma, and Ponnurangam Kumaraguru. Game-on: Graph attention network based multimodal fusion for fake news detection. *Social Network Analysis and Mining*, 14(1):114, 2024. doi: 10.1007/s13278-024-01271-4.

[38] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. doi: 10.1007/s11263-022-01653-1.

[39] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. MaPLe: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19113–19122, 2023.

[40] Ji Zhang, Shihan Wu, Lianli Gao, Heng Tao Shen, and Jingkuan Song. DePT: Decoupled prompt tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12924–12933, 2024.

[41] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017.

[42] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes].

   Justification: Please refer to the Abstract and Section 1 Introduction, which describe the SCBS diagnosis, SCBC framework, and evaluation scope.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes].

   Justification: Please refer to the limitations paragraph in the Conclusion, which discusses the main assumptions and limitations.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes].

Justification: Please refer to the Method section and Appendix proofs, where the assumptions, theorem and lemma statements, and derivations are provided.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes].

   Justification: Please refer to the Experimental Setup and Appendix implementation details; the source code is available at `https://github.com/Js183/SCBC`.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes].

Justification: The source code is available at `https://github.com/Js183/SCBC`, and the datasets used in the paper are public benchmarks cited in the manuscript.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes].

Justification: Please refer to the Experimental Setup and Appendix implementation details for the evaluation protocols, data splits, feature preprocessing, optimization procedures, and evaluation metrics.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes].

Justification: The main experiments use 5 independent runs, and the appendix reports mean$\pm$std results for detailed SCBC missingness tables.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes].

   Justification: Please refer to the Experimental Setup and Appendix efficiency analysis, which report the GPU and representative timing/memory results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes].

   Justification: The work uses established benchmarks, does not collect new human-subject data, and follows the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No].

    Justification: The paper focuses on the formulation, theoretical analysis, and benchmark evaluation of missing-modality classification. A dedicated analysis of downstream societal impacts is outside its scope.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A].

    Justification: The paper does not release high-risk pretrained models, image generators, or scraped datasets.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes].

    Justification: Existing datasets, feature backbones, and baselines used in the paper are properly cited, and their licenses and terms of use are respected.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes].

    Justification: The paper releases the SCBC source code at `https://github.com/Js183/SCBC`; no new dataset is introduced.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A].

    Justification: The paper does not involve crowdsourcing or human-subject research.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A].

    Justification: The paper does not involve crowdsourcing or human-subject research.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A].

    Justification: The core method does not use LLMs as an important, original, or non-standard component.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.

A Related Work

Missing-modality multimodal learning. Existing approaches divide into recovery-based and non-recovery strategies. **Recovery-based methods** reconstruct absent views before fusion via autoencoding [1], cyclic translation [21], graph completion [2], diffusion [3], or distribution-consistent generation [4]. However, their performance is inherently tied to reconstruction fidelity, and errors can compound under severe missingness. **Non-recovery methods** instead learn subset-robust representations through multi-view correlation [5–7], relative-advantage learning [8], cyclic latent alignment [10], or robust incomplete-data training [9, 12]. While these improve feature sufficiency, they leave subset-specific decision failures largely implicit.

Unified subset modeling. A growing line of work trains a single backbone to handle arbitrary modality subsets, avoiding per-pattern expert networks. MPLMM [13], CLKD [14], TriSPrompt [15], TF-Mamba [16], MMP [17], DCP [18], and SyP [11] adapt via prompts, distillation, or latent-space fusion under partial observations. SCBC is compatible with this paradigm but targets the subset-conditioned geometry of class competition rather than fused-representation quality alone.

Decision-aware robustness. RedCore [8], CyIN [10], and SyP [11] introduce adaptive weighting or subset-aware guidance at the decision level. Unlike these methods, SCBC explicitly tracks which class pairs persistently violate margins under each observed subset and decouples a historically grounded correction direction from a sample-adaptive step size, directly targeting the boundary-level failures documented in §2.

B Proofs and Derivations for SCBC

This appendix provides the detailed derivations behind the first-order repair-direction result and the two technical consequences quoted in the main text.

B.1 Derivation of the First-Order Repair Direction

Fix a subset S and abbreviate $\mathbf{z} = \mathbf{z}_S^{\text{base}}$, $\mathbf{D} = \mathbf{D}_S$, and $\mathbf{p} = \mathrm{softmax}(\mathbf{z})$. Let

$$\mathbf{W} = \mathrm{Diag}(\mathbf{p})\mathbf{D}, \qquad \mathbf{M}(\mathbf{r}) = (\mathbf{z}+\mathbf{r})\mathbf{1}^\top - \mathbf{1}(\mathbf{z}+\mathbf{r})^\top.$$

The local boundary functional can then be written as

$$\mathcal{R}_S(\mathbf{r}) = \big\langle \mathbf{W},\, \mathrm{softplus}\big(\tau_m \mathbf{1}\mathbf{1}^\top - \mathbf{M}(\mathbf{r})\big)\big\rangle_F .$$

Because softplus is smooth, its first-order expansion at $\mathbf{r} = \mathbf{0}$ gives

$$\mathrm{softplus}\big(\tau_m \mathbf{1}\mathbf{1}^\top - \mathbf{M}(\mathbf{r})\big) = \mathrm{softplus}\big(\tau_m \mathbf{1}\mathbf{1}^\top - \mathbf{M}(\mathbf{0})\big) - \boldsymbol{\Omega} \odot \big(\mathbf{M}(\mathbf{r}) - \mathbf{M}(\mathbf{0})\big) + O\big(\|\mathbf{r}\|_2^2\big),$$

where $\boldsymbol{\Omega}_{ab} = \mathrm{sigmoid}(\tau_m - (z_a - z_b))$. The exact first-order term contains the elementwise factor $\boldsymbol{\Omega}$. Since the debt entries are constructed from a near-margin kernel of the same sign and local support, we absorb this positive local weighting into an effective debt matrix $\widetilde{\mathbf{D}}$. Under the near-violation clause (A3), the active weights vary locally and preserve the induced signed direction up to the stated approximation. For notational simplicity, we write $\widetilde{\mathbf{D}}$ and its induced weight matrix again as $\mathbf{D}$ and $\mathbf{W}$. Therefore

$$\mathcal{R}_S(\mathbf{r}) = \mathcal{R}_S(\mathbf{0}) - \langle \mathbf{W},\, \mathbf{M}(\mathbf{r}) - \mathbf{M}(\mathbf{0})\rangle_F + O\big(\|\mathbf{r}\|_2^2\big).$$

Using $\mathbf{M}(\mathbf{r}) - \mathbf{M}(\mathbf{0}) = \mathbf{r}\mathbf{1}^\top - \mathbf{1}\mathbf{r}^\top$ and Frobenius trace identities,

$$\big\langle \mathbf{W},\, \mathbf{r}\mathbf{1}^\top - \mathbf{1}\mathbf{r}^\top\big\rangle_F = \mathbf{r}^\top\big(\mathbf{W}\mathbf{1} - \mathbf{W}^\top\mathbf{1}\big) = \big\langle \big(\mathbf{W} - \mathbf{W}^\top\big)\mathbf{1}, \mathbf{r}\big\rangle .$$

Since $\mathbf{W} - \mathbf{W}^\top = \mathrm{Diag}(\mathbf{p})\mathbf{D} - \mathbf{D}^\top \mathrm{Diag}(\mathbf{p}) = \mathcal{A}_S$, we obtain

$$\mathcal{R}_S(\mathbf{r}) = \mathcal{R}_S(\mathbf{0}) - \langle \mathcal{A}_S\mathbf{1}, \mathbf{r}\rangle + O\big(\|\mathbf{r}\|_2^2\big).$$

Since admissible residuals satisfy $\mathbf{1}^\top\mathbf{r} = 0$, only the centered component matters:

$$\nabla_{\mathbf{r}}\mathcal{R}_S(\mathbf{0}) \equiv -\Pi_0(\mathcal{A}_S\mathbf{1}).$$

Under a fixed norm budget $\|\mathbf{r}\|_2 \le \kappa$, Cauchy–Schwarz implies

$$\max_{\mathbf{1}^\top\mathbf{r}=0,\ \|\mathbf{r}\|_2\le\kappa} \Big\langle \Pi_0(\mathcal{A}_S\mathbf{1}), \mathbf{r}\Big\rangle = \kappa\, \|\Pi_0(\mathcal{A}_S\mathbf{1})\|_2 .$$

Hence the normalized centered direction is the steepest first-order repair direction.

B.2 Smoothness Bound for Lemma 1

For a single example with label y, the cross-entropy loss in logit space is

$$\ell(\mathbf{z}, y) = -\log \mathrm{softmax}(\mathbf{z})_y = \log \sum_{c=1}^{C} \exp(z_c) - z_y.$$

Let $\mathbf{p} = \mathrm{softmax}(\mathbf{z})$. Then

$$\nabla_{\mathbf{z}} \ell(\mathbf{z}, y) = \mathbf{p} - \mathbf{e}_y, \qquad \nabla^2_{\mathbf{z}} \ell(\mathbf{z}, y) = \mathrm{Diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^\top.$$

For any $\mathbf{u} \in \mathbb{R}^C$,

$$\mathbf{u}^\top \nabla^2_{\mathbf{z}} \ell(\mathbf{z}, y)\mathbf{u} = \sum_{c=1}^{C} p_c u_c^2 - \left(\sum_{c=1}^{C} p_c u_c \right)^2 = \mathrm{Var}_{c\sim\mathbf{p}}(u_c).$$

By Popoviciu's variance bound,

$$\mathrm{Var}_{c\sim\mathbf{p}}(u_c) \le \frac{(\max_c u_c - \min_c u_c)^2}{4}.$$

Moreover,

$$\max_c u_c - \min_c u_c \le \sqrt{2}\|\mathbf{u}\|_2,$$

because the range is attained by two coordinates and

$$(u_i - u_j)^2 \le 2(u_i^2 + u_j^2) \le 2\|\mathbf{u}\|_2^2.$$

Therefore,

$$\mathbf{u}^\top \nabla^2_{\mathbf{z}} \ell(\mathbf{z}, y)\mathbf{u} \le \frac{1}{2}\|\mathbf{u}\|_2^2.$$

Since this holds for all $\mathbf{u}$, we have the uniform Loewner bound

$$\nabla^2_{\mathbf{z}} \ell(\mathbf{z}, y) \preceq \frac{1}{2} I.$$

Thus the softmax cross-entropy is L-smooth in Euclidean logit space with $L = 1/2$. The constant $1/2$ is tight in logit space. In the binary case with $\mathbf{p} = (1/2, 1/2)$,

$$\nabla^2_{\mathbf{z}} \ell = \begin{pmatrix} 1/4 & -1/4 \\ -1/4 & 1/4 \end{pmatrix},$$

whose eigenvalues are 0 and $1/2$. Consequently, for any displacement $\mathbf{u}$,

$$\ell(\mathbf{z} + \mathbf{u}, y) \le \ell(\mathbf{z}, y) + \langle \nabla_{\mathbf{z}} \ell(\mathbf{z}, y), \mathbf{u} \rangle + \frac{1}{4}\|\mathbf{u}\|_2^2.$$

B.3 Proof of Lemma 1

By Appendix B.2, the cross-entropy loss is L-smooth in logit space with $L = 1/2$. Hence, for any perturbation $\mathbf{r}$ and any $\alpha > 0$, the descent lemma gives

$$\mathcal{L}_i(\mathbf{z}_S + \alpha\mathbf{r}) \le \mathcal{L}_i(\mathbf{z}_S) + \alpha\langle \nabla_{\mathbf{z}} \mathcal{L}_i(\mathbf{z}_S), \mathbf{r} \rangle + \frac{L\alpha^2}{2}\|\mathbf{r}\|_2^2.$$

Substituting $L = 1/2$ yields

$$\mathcal{L}_i(\mathbf{z}_S + \alpha\mathbf{r}) \le \mathcal{L}_i(\mathbf{z}_S) + \alpha\langle \nabla_{\mathbf{z}} \mathcal{L}_i(\mathbf{z}_S), \mathbf{r} \rangle + \frac{\alpha^2}{4}\|\mathbf{r}\|_2^2.$$

Under condition $\mathcal{C}$, the gradient-related condition gives, almost surely,

$$\langle \nabla_{\mathbf{z}} \mathcal{L}_i(\mathbf{z}_S), \mathbf{r} \rangle \le -\kappa\|\mathbf{r}\|_2^2.$$

Therefore, almost surely on $\mathcal{C}$,

$$\mathcal{L}_i(\mathbf{z}_S + \alpha\mathbf{r}) \le \mathcal{L}_i(\mathbf{z}_S) - \alpha\left(\kappa - \frac{\alpha}{4}\right)\|\mathbf{r}\|_2^2.$$

Taking conditional expectation over $\mathcal{C}$ gives

$$\mathbb{E}[\mathcal{L}_i(\mathbf{z}_S + \alpha\mathbf{r}) \mid \mathcal{C}] \le \mathbb{E}[\mathcal{L}_i(\mathbf{z}_S) \mid \mathcal{C}] - \alpha\left(\kappa - \frac{\alpha}{4}\right)\mathbb{E}[\|\mathbf{r}\|_2^2 \mid \mathcal{C}].$$

Thus any $\alpha \in (0, \min\{1, 4\kappa\}]$ ensures non-increase of the conditional expected loss. If $\alpha \in (0, \min\{1, 4\kappa\})$ and $\mathbb{E}[\|\mathbf{r}\|_2^2 \mid \mathcal{C}] > 0$, the decrease is strict.

B.4 Derivation of the Support-Calibrated Debt Limit

For clarity, we state the consistency calculation for a conceptual subset-indexed version; the implemented modality-factorized memory follows by applying the same argument to each modality-level table (m, a, b) and aggregating through Eq. (6).

Fix a triplet (S, a, b) and abbreviate

$$\bar{\Delta}_{S,a,b} = \mathbb{E}[\widehat{\mathbf{E}}_t(S, a, b)], \qquad \bar{n}_{S,a} = \mathbb{E}[\widehat{\mathbf{N}}_t(S, a)], \qquad \bar{\xi}_S = \mathbb{E}[\widehat{\xi}_t(S)].$$

Under unbiased batch estimation, $\bar{\Delta}_{S,a,b} = \Delta_S(a, b)$. Let

$$\bar{V}_t = \mathbb{E}[\mathbf{V}_t(S, a, b)], \qquad \bar{N}_t = \mathbb{E}[\mathbf{N}_t(S, a)], \qquad \bar{\xi}_t = \mathbb{E}[\xi_t(S)].$$

Taking expectations on the EMA exposure and support recursions gives the affine recursions

$$\bar{\xi}_t = \mu_N\, \bar{\xi}_{t-1} + (1-\mu_N)\bar{\xi}_S, \qquad \bar{V}_t = \mu\, \bar{V}_{t-1} + (1-\mu)\bar{\Delta}_{S,a,b}, \qquad \bar{N}_t = \mu_N\, \bar{N}_{t-1} + (1-\mu_N)\bar{n}_{S,a}.$$

Thus $\bar{\xi}_t \to \bar{\xi}_S$, $\bar{V}_t \to \bar{\Delta}_{S,a,b}$, and $\bar{N}_t \to \bar{n}_{S,a}$ as $t \to \infty$.

Now define the asymptotic exposure calibration factor

$$\bar{\gamma}_S = \mathrm{clip}\left(\frac{\max_{S'} \bar{\xi}_{S'}}{\bar{\xi}_S + \epsilon_s}, 1, \beta_{\max}\right).$$

The calibrated-support definition implies that the effective support converges in mean to $\bar{\gamma}_S \bar{n}_{S,a}$. Applying the continuous mapping theorem to the debt-ratio update then yields

$$\mathbb{E}[\mathbf{D}_t(S, a, b)] \longrightarrow \frac{\bar{\gamma}_S\, \bar{n}_{S,a}}{\bar{\gamma}_S\, \bar{n}_{S,a} + \nu}\, \bar{\Delta}_{S,a,b} = \frac{\bar{\gamma}_S\, \bar{n}_{S,a}}{\bar{\gamma}_S\, \bar{n}_{S,a} + \nu}\, \Delta_S(a, b), \qquad t \to \infty,$$

which is the support-shrunk population limit quoted in the main text. Hence the external debt statistic converges to a support-shrunk version of the target debt rather than to an unregularized empirical confusion table.

B.5 Derivation of the Conservative Activation Bound

Rewrite the residual correction as

$$\mathbf{r}_S = \Pi_0(\mathbf{v}_S), \qquad \mathbf{v}_S = \lambda_d\, \chi_S\, \big(\boldsymbol{\beta}_S \odot \mathbf{q}_S\big).$$

If $\chi_S = 0$, then $\mathbf{v}_S = \mathbf{0}$ and hence $\mathbf{r}_S = \mathbf{0}$. If $\mathbf{q}_S = \mathbf{0}$, the same conclusion holds. In both cases,

$$\mathbf{z}_S^{\mathrm{final}} = \mathbf{z}_S^{\mathrm{base}} + \mathbf{r}_S = \mathbf{z}_S^{\mathrm{base}}.$$

For the norm bound, observe that for any $\mathbf{v} \in \mathbb{R}^C$,

$$\|\Pi_0(\mathbf{v})\|_\infty = \left\|\mathbf{v} - \frac{1}{C}(\mathbf{1}^\top \mathbf{v})\mathbf{1}\right\|_\infty \le 2\|\mathbf{v}\|_\infty.$$

Applying the projector bound and the elementwise constraint $\boldsymbol{\beta}_S \in [0, \beta_{\max}]^C$ gives

$$\|\mathbf{r}_S\|_\infty \le 2\|\mathbf{v}_S\|_\infty \le 2\lambda_d\, \chi_S\, \|\boldsymbol{\beta}_S \odot \mathbf{q}_S\|_\infty \le 2\lambda_d\, \chi_S\, \beta_{\max}\, \|\mathbf{q}_S\|_\infty,$$

which is the conservative bound quoted in the main text. Since $\chi_S \le 1$ and $\|\mathbf{q}_S\|_\infty \le \|\mathbf{q}_S\|_2$, it also implies

$$\|\mathbf{r}_S\|_2 \le \sqrt{C}\|\mathbf{r}_S\|_\infty \le 2\sqrt{C}\lambda_d \beta_{\max}\|\mathbf{q}_S\|_2.$$

Thus the scalar envelope for an unnormalized debt-direction step $\alpha_S \mathbf{q}_S$ satisfies $\alpha_S \le 2\sqrt{C}\lambda_d \beta_{\max}$. The sufficient safe-step condition

$$\lambda_d \beta_{\max} \le \frac{\min\{1, 4c_0\}}{2\sqrt{C}}$$

therefore places this envelope inside the admissible interval from Lemma 3 and Lemma 1. This shows that the correction branch is contractive with respect to the learned amplitude and becomes silent in the degenerate cases listed in the main text.

B.6 Proof of Lemma 3

For completeness, we spell out the local risk-compatibility clause used in Lemma 3. This clause formalizes the regime targeted by the pairwise training loss and empirically verified in Appendix D.10 and Appendix B.7.

For a debt matrix $\mathbf{D}_S$, set $D_S(a,a) = 0$ and define the row-induced signed vector $\mathbf{T}_a(\mathbf{D}_S) \in \mathbb{R}^C$ by

$$[\mathbf{T}_a(\mathbf{D}_S)]_c = \begin{cases} \sum_{b \neq a} D_S(a,b), & c = a, \\ -D_S(a,c), & c \neq a. \end{cases}$$

Thus $\mathbf{1}^\top \mathbf{T}_a(\mathbf{D}_S) = 0$. The vector $\mathbf{T}_a(\mathbf{D}_S)$ increases the logit of class a and decreases the logits of its historically confusing competitors.

We assume that on the active event $\mathcal{E}_S = \{\|\mathbf{q}_S\|_2 > \delta\}$, for the true label $a = y_i$, the following local compatibility conditions hold. First, the debt entries are bounded:

$$0 \le D_S(c,d) \le D_{\max} \quad \text{for all } c, d$$

for some $D_{\max} > 0$. This is automatic for the EMA debt book because the update signal is bounded and the EMA shrinkage factors lie in $[0,1]$.

Second, the active true-label row contains at least one near-margin competitor: whenever $\mathbf{T}_a(\mathbf{D}_S) \neq \mathbf{0}$, there exists $b \neq a$ such that

$$D_S(a,b) > 0, \qquad z^{\text{base}}_{S,a} - z^{\text{base}}_{S,b} \le \tau_m + \varepsilon_m.$$

Define

$$m_0 = \frac{1}{1 + \exp(\tau_m + \varepsilon_m)}.$$

Third, the posterior-weighted contribution of non-true-label debt rows is dominated by the true-label row:

$$\left\| \sum_{c \neq a} p_{S,c} \mathbf{T}_c(\mathbf{D}_S) \right\|_2 \le \eta_{\text{row}}\, p_{S,a} \|\mathbf{T}_a(\mathbf{D}_S)\|_2$$

for some

$$0 \le \eta_{\text{row}} < \frac{m_0}{2}.$$

Let $\mathbf{z} = \mathbf{z}^{\text{base}}_S$, $\mathbf{p} = \mathbf{p}_S = \mathrm{softmax}(\mathbf{z})$, $\mathbf{D} = \mathbf{D}_S$, and let the true label be $a = y_i$. For cross-entropy,

$$\nabla_{\mathbf{z}} \mathcal{L}_i(\mathbf{z}) = \mathbf{p} - \mathbf{e}_a, \qquad -\nabla_{\mathbf{z}} \mathcal{L}_i(\mathbf{z}) = \mathbf{e}_a - \mathbf{p}.$$

Write $\mathbf{g}_i = \mathbf{e}_a - \mathbf{p}$. By definition,

$$\mathbf{v}_S = \mathrm{Diag}(\mathbf{p})\mathbf{D}\mathbf{1} - \mathbf{D}^\top \mathbf{p}.$$

Since

$$\mathbf{1}^\top \mathbf{v}_S = \mathbf{p}^\top \mathbf{D}\mathbf{1} - \mathbf{1}^\top \mathbf{D}^\top \mathbf{p} = \mathbf{p}^\top \mathbf{D}\mathbf{1} - \mathbf{p}^\top \mathbf{D}\mathbf{1} = 0,$$

the vector $\mathbf{v}_S$ already lies in the zero-sum subspace. Hence $\mathbf{q}_S = \Pi_0(\mathbf{v}_S) = \mathbf{v}_S$.

Using the row-induced signed vectors $\mathbf{T}_c(\mathbf{D})$, we can rewrite

$$\mathbf{q}_S = \sum_{c=1}^{C} p_c \mathbf{T}_c(\mathbf{D}) = p_a \mathbf{T}_a(\mathbf{D}) + \sum_{c \neq a} p_c \mathbf{T}_c(\mathbf{D}).$$

Define the nuisance component

$$\mathbf{n}_a = \sum_{c \neq a} p_c \mathbf{T}_c(\mathbf{D}).$$

Then $\mathbf{q}_S = p_a \mathbf{T}_a(\mathbf{D}) + \mathbf{n}_a$.

We first lower-bound the alignment of the true-label debt row. Let

$$d_a = \sum_{b \neq a} D(a,b).$$

By the definition of $\mathbf{T}_a(\mathbf{D})$,

$$\langle \mathbf{g}_i, \mathbf{T}_a(\mathbf{D}) \rangle = (1 - p_a) d_a + \sum_{b \neq a} p_b D(a, b).$$

Therefore,

$$\langle \mathbf{g}_i, \mathbf{T}_a(\mathbf{D}) \rangle \geq (1 - p_a) d_a.$$

Moreover,

$$\|\mathbf{T}_a(\mathbf{D})\|_2^2 = d_a^2 + \sum_{b \neq a} D(a, b)^2 \leq d_a^2 + \left(\sum_{b \neq a} D(a, b) \right)^2 = 2 d_a^2,$$

so

$$d_a \geq \frac{1}{\sqrt{2}} \|\mathbf{T}_a(\mathbf{D})\|_2.$$

By local near-violation compatibility, there exists $b \neq a$ with $D(a, b) > 0$ and

$$z_a - z_b \leq \tau_m + \varepsilon_m.$$

Since

$$\frac{p_b}{p_a} = \exp(z_b - z_a) \geq \exp[-(\tau_m + \varepsilon_m)],$$

we have

$$p_a + p_b \geq p_a \left(1 + \exp[-(\tau_m + \varepsilon_m)]\right).$$

Because $p_a + p_b \leq 1$,

$$p_a \leq \frac{1}{1 + \exp[-(\tau_m + \varepsilon_m)]}.$$

Equivalently,

$$1 - p_a \geq \frac{1}{1 + \exp(\tau_m + \varepsilon_m)} = m_0.$$

Combining the previous bounds gives

$$\langle \mathbf{g}_i, \mathbf{T}_a(\mathbf{D}) \rangle \geq \frac{m_0}{\sqrt{2}} \|\mathbf{T}_a(\mathbf{D})\|_2.$$

Now decompose the alignment with $\mathbf{q}_S$:

$$\langle \mathbf{g}_i, \mathbf{q}_S \rangle = p_a \langle \mathbf{g}_i, \mathbf{T}_a(\mathbf{D}) \rangle + \langle \mathbf{g}_i, \mathbf{n}_a \rangle.$$

Using Cauchy–Schwarz and $\|\mathbf{g}_i\|_2 = \|\mathbf{e}_a - \mathbf{p}\|_2 \leq \sqrt{2}$,

$$\langle \mathbf{g}_i, \mathbf{q}_S \rangle \geq p_a \frac{m_0}{\sqrt{2}} \|\mathbf{T}_a(\mathbf{D})\|_2 - \sqrt{2} \|\mathbf{n}_a\|_2.$$

By the row-dominance clause,

$$\|\mathbf{n}_a\|_2 \leq \eta_{\text{row}} p_a \|\mathbf{T}_a(\mathbf{D})\|_2.$$

Therefore,

$$\langle \mathbf{g}_i, \mathbf{q}_S \rangle \geq \left(\frac{m_0}{\sqrt{2}} - \sqrt{2} \eta_{\text{row}} \right) p_a \|\mathbf{T}_a(\mathbf{D})\|_2.$$

Since $\eta_{\text{row}} < m_0/2$, the coefficient is strictly positive.

Next, again by row dominance,

$$\|\mathbf{q}_S\|_2 \leq p_a \|\mathbf{T}_a(\mathbf{D})\|_2 + \|\mathbf{n}_a\|_2 \leq (1 + \eta_{\text{row}}) p_a \|\mathbf{T}_a(\mathbf{D})\|_2.$$

Hence

$$p_a \|\mathbf{T}_a(\mathbf{D})\|_2 \geq \frac{\|\mathbf{q}_S\|_2}{1 + \eta_{\text{row}}}.$$

Consequently,

$$\langle \mathbf{g}_i, \mathbf{q}_S \rangle \geq \gamma_0 \|\mathbf{q}_S\|_2,$$

where

$$\gamma_0 = \frac{\frac{m_0}{\sqrt{2}} - \sqrt{2}\eta_{\text{row}}}{1 + \eta_{\text{row}}} > 0.$$

It remains to convert this linear lower bound into the quadratic form used in Lemma 3. Since $0 \le D(c,d) \le D_{\max}$, for every class c,

$$\|\mathbf{T}_c(\mathbf{D})\|_2^2 = \left(\sum_{d \neq c} D(c,d)\right)^2 + \sum_{d \neq c} D(c,d)^2 \le (C-1)^2 D_{\max}^2 + (C-1) D_{\max}^2 = C(C-1) D_{\max}^2.$$

Thus

$$\|\mathbf{T}_c(\mathbf{D})\|_2 \le \sqrt{C(C-1)} D_{\max}.$$

Using $\mathbf{q}_S = \sum_c p_c \mathbf{T}_c(\mathbf{D})$ and $\sum_c p_c = 1$,

$$\|\mathbf{q}_S\|_2 \le \sum_c p_c \|\mathbf{T}_c(\mathbf{D})\|_2 \le \sqrt{C(C-1)} D_{\max} =: Q_{\max}.$$

Therefore,

$$\|\mathbf{q}_S\|_2 \ge \frac{1}{Q_{\max}} \|\mathbf{q}_S\|_2^2.$$

Combining this with the linear alignment bound gives

$$\langle \mathbf{g}_i, \mathbf{q}_S \rangle \ge \frac{\gamma_0}{Q_{\max}} \|\mathbf{q}_S\|_2^2.$$

Since $\mathbf{g}_i = -\nabla_{\mathbf{z}} \mathcal{L}_i(\mathbf{z}_S^{\text{base}})$, we obtain

$$\langle -\nabla_{\mathbf{z}} \mathcal{L}_i(\mathbf{z}_S^{\text{base}}), \mathbf{q}_S \rangle \ge c_0 \|\mathbf{q}_S\|_2^2,$$

where

$$c_0 = \frac{\gamma_0}{Q_{\max}} = \frac{\frac{m_0}{\sqrt{2}} - \sqrt{2}\eta_{\text{row}}}{(1 + \eta_{\text{row}})\sqrt{C(C-1)} D_{\max}} > 0.$$

Taking conditional expectation over $\mathcal{E}_S$ yields

$$\mathbb{E}\big[\langle -\nabla_{\mathbf{z}} \mathcal{L}_i(\mathbf{z}_S^{\text{base}}), \mathbf{q}_S \rangle \mid \mathcal{E}_S\big] \ge c_0 \mathbb{E}[\|\mathbf{q}_S\|_2^2 \mid \mathcal{E}_S],$$

which proves Lemma 3.

B.7 Empirical Verification Protocol for Proposition 2

Proposition 2 turns the training-side pairwise objective into an observable proxy for the gradient-related condition. For a fixed checkpoint, we first compare the debt-weighted pairwise loss at the realized logits and at the no-correction baseline:

$$\Delta_{\text{pair}} = \frac{1}{|\mathcal{T}|} \sum_{i \in \mathcal{T}} \Big(\ell_i^{\text{pair}}(\mathbf{z}_{S_i}^{\text{final}}) - \ell_i^{\text{pair}}(\mathbf{z}_{S_i}^{\text{base}}) \Big),$$

where $\mathcal{T}$ is the training split and ℓ_i^{pair} uses the same debt row, off-diagonal support, margin threshold τ_m, and minimum-debt fallback rule as Eq. (19). Negative Δ_{pair} means that the realized correction decreases the debt-weighted surrogate that Proposition 2 uses as its premise.

We then measure the first-order alignment of the correction direction with the base classification gradient. Writing

$$\mathbf{g}_i = \nabla_{\mathbf{z}} \mathcal{L}_i(\mathbf{z}_{S_i}^{\text{base}}), \qquad \mathbf{r}_{S_i} = \mathbf{z}_{S_i}^{\text{final}} - \mathbf{z}_{S_i}^{\text{base}},$$

the direct proposition-level statistic is the active-sample fraction

$$\text{Frac}_r = \frac{1}{|\mathcal{T}_r|} \sum_{i \in \mathcal{T}_r} \mathbf{1}[\langle \mathbf{g}_i, \mathbf{r}_{S_i} \rangle < 0], \qquad \mathcal{T}_r = \left\{ i : \|\mathbf{r}_{S_i}\|_2 > 10^{-8} \right\}.$$

Table 8: **Per-setting empirical verification of direction alignment.** We report representative settings: IEMOCAP-4 single-modality subsets $\{l\}$, $\{v\}$, and $\{a\}$, Weibo-17 at $R_m = 0.5$, and PHEME at $R_m = 0.7$. The non-vanishing event is operationalized as $\|\mathbf{q}_S\|_2 > 10^{-3}$.

Setting	$\mathbb{P}(\mathcal{E}_S)$ on test (%)	Debt-weighted sign alignment (%)	Frac_q (%)
IEMOCAP-4 $\{l\}$	100.00	99.20	96.11
IEMOCAP-4 $\{v\}$	100.00	94.45	90.44
IEMOCAP-4 $\{a\}$	100.00	94.30	87.97
Weibo-17 ($R_m = 0.5$)	91.03	99.34	79.39
PHEME ($R_m = 0.7$)	100.00	85.60	72.17

However, the correction head is conservative, so in some settings the realized residual is numerically negligible even when the retrieved debt direction is informative. For that reason, the per-setting table reports the pre-gated direction $\mathbf{q}_{S_i}$ instead:

$$\mathrm{Frac}_q = \frac{1}{|\mathcal{T}_q|} \sum_{i \in \mathcal{T}_q} \mathbf{1}[\langle \mathbf{g}_i, \mathbf{q}_{S_i} \rangle < 0], \qquad \mathcal{T}_q = \left\{ i : \|\mathbf{q}_{S_i}\|_2 > 10^{-3} \right\},$$

together with the debt-weighted support-coordinate sign-agreement statistic used in Table 8. Concretely, for each training sample we mark competitor coordinates with positive debt mass in the row indexed by the true class, and we compute the fraction of that debt support on which $q_{S_i,k}$ has the opposite sign to $(\mathbf{g}_i)_k$.

We instantiate this protocol on five representative settings: IEMOCAP-4 single-modality subsets $\{l\}$, $\{v\}$, and $\{a\}$, Weibo-17 at $R_m = 0.5$, and PHEME at $R_m = 0.7$. For each setting, we compute gradient-alignment statistics on the training split and the non-vanishing event on the test split.

The detailed rows confirm that the retrieved debt direction remains active and usually points toward descent across all five settings, while also showing the cross-setting spread behind the pooled average.

C Implementation Details and Data Splits

C.1 Training Details

We follow the encoders used by the corresponding protocols: wav2vec-large [29], DeBERTa-large [30], and MA-Net [31] for CMU-MOSEI/IEMOCAP; BERT [32] and ResNet-34 [33] for Weibo-17/PHEME, with PHEME additionally using comments; and CLIP ViT-B/16 [34] for Hateful Memes. All experiments were run on a single NVIDIA A100 80GB GPU. Unless otherwise stated, we use AdamW [35] with gradient clipping. For CMU-MOSEI, Hateful Memes, PHEME, and Weibo-17, we use a cosine learning-rate schedule [41], bfloat16 mixed precision, and TF32 matrix multiplication on CUDA. The dataset-specific optimization settings, SCBC parameters, and loss weights are specified in the corresponding code configurations.

For IEMOCAP-4, we use a dialogue-oriented pipeline in which each optimizer step processes one dialogue.

C.2 Dataset Splits

For IEMOCAP-4, we use a fixed session-based split: Sessions 1–4 provide 3,600 training utterances, and Session 5 provides 943 test utterances. CMU-MOSEI contains 16,326 training, 1,871 validation, and 4,659 test samples.

For Weibo-17, the training, validation, and test splits contain 6,778, 753, and 1,996 samples, respectively. PHEME contains 1,412 training, 221 validation, and 385 test samples. Hateful Memes contains 8,500 training, 540 validation, and 2,000 test samples.

Table 9 summarizes the feature preprocessing pipeline used for each active benchmark.

Table 9: Feature preprocessing summary for the evaluated benchmarks.

Dataset	Text feature	Visual feature	Other stream(s)
IEMOCAP-4	DeBERTa-large (1024-d)	MTCNN + MA-Net (1024-d)	wav2vec-large audio (512-d)
CMU-MOSEI	DeBERTa-large (1024-d)	MTCNN + MA-Net (1024-d)	wav2vec-large audio (512-d)
Weibo-17	BERT (768-d)	ResNet-34 image (512-d)	none
PHEME	BERT (768-d)	ResNet-34 image (512-d)	comment feature stream (768-d)
Hateful Memes	CLIP text token features (77×512)	CLIP ViT-B/16 image embedding (768-d)	none

C.3 Baseline Result Sources

For transparency, we state the provenance of the non-SCBC entries used in the main comparisons. For IEMOCAP-4, all baseline values in Table 1 are taken from the comparison records reported by HME [9] under the same fixed- and random-missing settings. For CMU-MOSEI, the recent missing-modality baselines GCNet, IMDer, TF-Mamba, and HME in Table 2 follow the official reported results of the corresponding methods [2, 3, 9, 16]. The exception among the recent CMU-MOSEI baselines is LNLN [12]: its released evaluation code computes F1 differently from the binary sentiment protocol used here, so we corrected the F1 calculation and reproduced LNLN on the CMU-MOSEI split and feature setting used in our experiments.

For Weibo-17 and PHEME, the reported DiCMoR, CAFE, GameOn, CLKD, RedCore, and TriSPrompt baselines follow the records reported by TriSPrompt [15] under its missing-modality evaluation protocol; we additionally reproduced HME [9] under the same data split, feature setting, and missingness protocol. For Hateful Memes, the prior baseline and SyP entries in Table 3 are taken from SyP [11], and HME is reproduced under the same SyP evaluation conditions.

D Additional Experimental Results

D.1 CMU-MOSEI Missingness Tables

Tables 10 and 11 report mean±std SCBC results on CMU-MOSEI under fixed and random missingness, respectively.

Table 10: **SCBC results on CMU-MOSEI under fixed missingness.**

Setting	Acc_2	F1	Acc_7
a	70.5±0.4	68.8±0.2	42.2±0.1
v	67.3±0.3	66.3±0.3	42.6±0.2
l	86.7±0.0	86.6±0.0	53.5±0.2
av	69.8±0.1	68.7±0.2	43.5±0.4
la	86.9±0.0	86.8±0.0	54.5±0.2
lv	86.8±0.1	86.7±0.1	54.6±0.2
lva	86.8±0.1	86.7±0.0	55.0±0.2

Table 11: **SCBC results on CMU-MOSEI under random missingness.**

R_m	Acc_2	F1	Acc_7
0.0	86.8±0.1	86.7±0.0	55.0±0.2
0.1	85.3±0.2	85.2±0.2	54.1±0.0
0.2	83.4±0.1	83.2±0.0	52.8±0.1
0.3	81.8±0.1	81.4±0.2	51.0±0.1
0.4	80.2±0.2	79.7±0.2	50.2±0.1
0.5	78.2±0.1	77.6±0.1	48.8±0.1
0.6	76.2±0.2	75.4±0.1	47.1±0.3
0.7	74.8±0.2	73.8±0.3	46.1±0.1

D.2 Cross-Dataset Transfer Summary

To probe whether the learned subset-conditioned compensation transfers across dataset distributions, we train on CMU-MOSEI and evaluate on CMU-MOSI under the fixed-missing protocol without any fine-tuning. Table 12 reports the resulting cross-dataset generalization performance against transferred baselines.

D.3 PHEME Missingness Tables

Tables 13 and 14 report mean±std SCBC results on PHEME under fixed and random missingness, respectively.

Table 12: **Cross-dataset generalization under fixed missingness.** Models are trained on CMU-MOSEI and tested on CMU-MOSI. Entries are Acc_2/F1/Acc_7.

Subset	IMDer	HME	SCBC
{l}	83.2/83.0/41.8	81.4/81.3/34.4	86.3/86.3/44.8
{a}	58.8/45.4/15.5	57.8/42.3/20.0	57.2/57.3/18.1
{v}	57.8/42.3/15.5	57.0/51.6/17.6	56.2/56.5/19.2
{l,a}	83.2/83.0/42.1	81.1/81.1/34.4	85.7/85.7/45.6
{l,v}	82.8/82.7/39.1	81.3/81.3/39.2	86.1/86.1/45.0
{a,v}	59.3/58.9/21.4	57.6/51.8/18.1	57.5/57.7/20.4
{l,a,v}	83.8/83.9/40.6	81.0/81.0/39.5	86.0/86.0/43.7
Avg.	72.7/68.5/30.9	71.0/67.2/29.0	**73.6/73.7/33.8**

Table 13: **SCBC results on PHEME under fixed missingness.**

Setting	Acc	F1	AUC
text	85.3±0.3	85.4±0.4	90.5±0.2
image	79.9±0.1	79.4±0.3	83.4±0.4
comment	74.5±0.1	74.6±0.1	77.3±0.3
text+image	84.1±0.4	83.9±0.2	89.7±0.5
text+comment	85.5±0.4	85.6±0.5	89.7±0.1
image+comment	79.8±0.7	79.5±0.6	84.2±0.3
text+image+comment	84.9±0.3	83.8±0.4	89.6±0.3

Table 14: **SCBC results on PHEME under random missingness.**

R_m	Acc	F1	AUC
0.0	84.9±0.3	83.8±0.4	89.6±0.3
0.1	83.9±0.5	83.9±0.4	88.8±0.7
0.2	83.4±0.7	83.4±0.7	87.9±0.5
0.3	83.4±0.4	83.0±0.3	88.0±0.6
0.4	82.6±0.5	83.0±0.6	87.7±0.8
0.5	82.5±0.4	82.5±0.5	87.5±0.5
0.6	82.2±0.7	81.9±0.6	86.6±1.4
0.7	79.7±0.7	81.2±0.6	85.7±0.8

D.4 Weibo-17 Missingness Tables

Tables 15 and 16 report mean±std SCBC results on Weibo-17 under single-modality settings and the random-missing protocol, respectively.

Table 15: **SCBC results on Weibo-17 under fixed subset settings.**

Setting	Acc	F1	AUC
text	87.9±0.4	87.9±0.4	94.4±0.2
image	64.5±0.5	64.5±0.5	69.4±0.5
text+image (full)	87.4±0.4	87.4±0.4	94.0±0.2

Table 16: **SCBC results on Weibo-17 under random missingness.**

R_m	Acc	F1	AUC
0.0	87.4±0.4	87.4±0.4	94.0±0.2
0.1	85.6±0.6	85.6±0.6	93.2±0.2
0.2	83.3±0.1	83.3±0.1	91.3±0.4
0.3	80.9±0.1	80.9±0.1	89.5±0.1
0.4	78.5±0.3	78.5±0.3	86.9±0.6
0.5	76.1±0.1	76.1±0.1	84.6±0.5

D.5 Hateful Memes Missingness Tables

Tables 17 and 18 report mean±std SCBC results on Hateful Memes under fixed and random missingness, respectively.

Table 17: **SCBC results on Hateful Memes under fixed missingness.**

Setting	Acc	F1	AUROC
text	62.5±0.2	61.1±0.3	62.7±0.3
image	73.0±0.1	72.2±0.1	77.3±0.2
text+image	74.3±0.2	73.5±0.3	79.0±0.3

Table 18: **SCBC results on Hateful Memes under random missingness.**

R_m	Missing text			Missing image			Both missing		
	Acc	F1	AUROC	Acc	F1	AUROC	Acc	F1	AUROC
0.1	73.8±0.1	73.3±0.3	78.5±0.2	73.0±0.4	72.4±0.5	77.7±0.3	73.7±0.4	72.9±0.4	77.9±0.3
0.2	73.5±0.1	72.9±0.1	78.1±0.0	72.1±0.5	71.5±0.3	76.3±0.3	72.5±0.5	71.8±0.6	76.9±0.4
0.3	73.2±0.3	72.6±0.5	78.0±0.3	70.4±0.1	69.4±0.2	74.4±0.1	72.2±0.3	71.3±0.3	76.2±0.4
0.4	73.3±0.2	72.5±0.3	77.7±0.1	69.4±0.3	68.7±0.1	73.1±0.4	71.5±0.6	70.6±0.5	75.6±0.4
0.5	73.0±0.4	72.4±0.5	77.4±0.2	68.8±0.2	68.0±0.2	72.0±0.2	70.9±0.3	69.9±0.5	74.8±0.6
0.6	73.1±0.6	72.6±0.6	77.6±0.2	66.8±0.2	66.1±0.2	69.9±0.3	70.2±0.2	69.5±0.1	74.3±0.3
0.7	72.9±0.3	72.0±0.4	77.3±0.1	65.7±0.1	64.8±0.1	67.7±0.3	69.9±0.1	69.0±0.2	73.5±0.4
0.8	72.8±0.4	72.1±0.4	77.4±0.3	64.2±0.1	63.1±0.2	65.6±0.1	69.7±0.6	68.7±0.6	72.8±0.7
0.9	72.7±0.2	72.4±0.1	77.5±0.3	63.2±0.2	62.1±0.3	64.5±0.1	68.4±0.3	67.7±0.1	71.9±0.1

D.6 Hyperparameter Sensitivity

Figure 5 reports sensitivity curves for the evaluated hyperparameter grids on PHEME and Weibo-17.

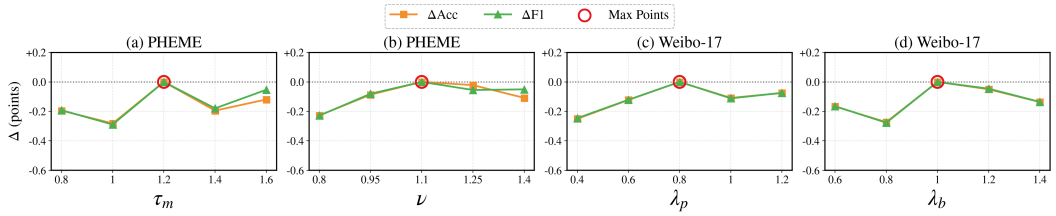

Figure 5: **Hyperparameter sensitivity.** Δ is the change in mean Acc/F1 relative to the default configuration, averaged over the full range of random missing rates on PHEME ($R_m \in \{0.0, ..., 0.7\}$) and Weibo-17 ($R_m \in \{0.0, ..., 0.5\}$). The circled points mark the maxima.

D.7 Subset-Conditioned Confusion Visualization

Figure 6 visualizes direct off-diagonal confusion rates on IEMOCAP-4 across all seven observed subsets $\{l\}$, $\{v\}$, $\{a\}$, $\{lv\}$, $\{la\}$, $\{av\}$, and $\{lva\}$. In each heatmap, rows denote true labels and columns denote predicted labels; diagonal entries are set to zero so the color scale focuses on misclassification flow. The highlighted high-rate cells vary strongly by subset, confirming that subset-conditioned errors are structured rather than reducible to global class difficulty.

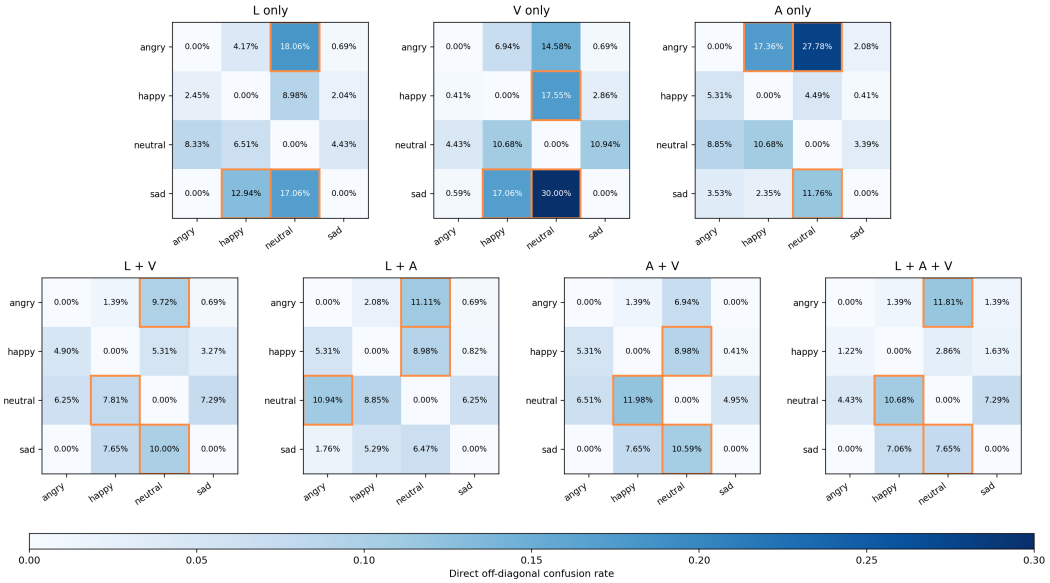

Figure 6: **Direct off-diagonal confusion rates across observed subsets on IEMOCAP-4.** Rows are true labels and columns are predicted labels. Each off-diagonal cell reports the percentage of samples from a true class that are predicted as another class under the indicated observed subset, with diagonal entries set to zero. Orange outlines mark high-confusion pairs; their subset-dependent locations support that the memory targets structured subset-conditioned confusion rather than a global class-difficulty prior.

D.8 Additional Case Studies

Figure 7 provides two representative masked-view examples that illustrate how the learned correction behaves under severe missingness. The top row is a CMU-MOSEI audio-only example and shows a near-tie base prediction: the signed change signal shifts probability mass away from the incorrect class $+1$ and toward the true class 0, yielding a correct final decision. The bottom row is an IEMOCAP-4 text-only example and starts from a much stronger wrong preference for *angry*; here the retrieved correction suppresses the incorrect *angry* prediction and amplifies *neutral*, again recovering the true label. Across both cases, the adjustment remains selective rather than global, changing only a few logits while preserving the overall score structure.

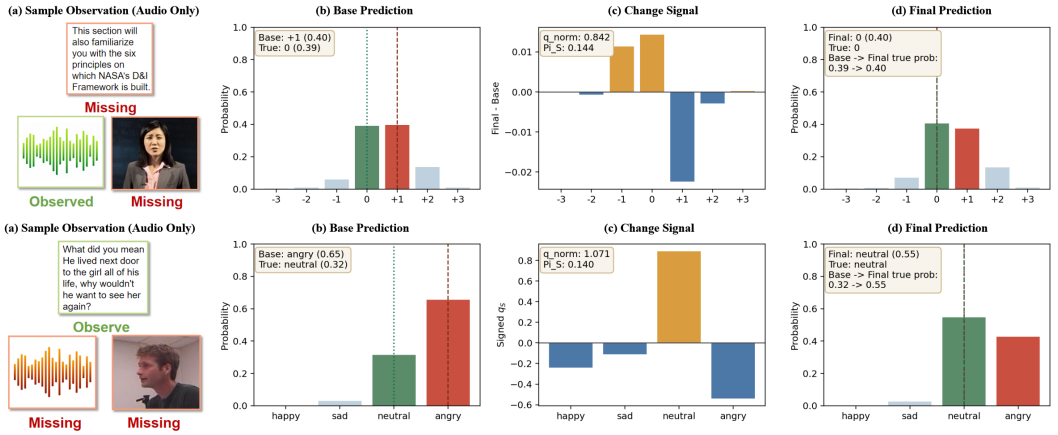

Figure 7: **Appendix case studies on CMU-MOSEI and IEMOCAP-4 under severe missingness.** Each row shows one masked-view example with the observed input, the base prediction, the signed change signal, and the final corrected prediction. The top row is a CMU-MOSEI audio-only example, where the correction moves probability mass from class $+1$ to the true class 0. The bottom row is an IEMOCAP-4 text-only example, where it suppresses the incorrect *angry* prediction and raises the true *neutral* class. These examples illustrate that the historical debt direction can produce targeted label repair without accessing the full reference view at inference.

Figure 8 further visualizes a local CMU-MOSEI audio-only repair pattern for the $+1/+2$ class pair. The selected subset contains the hardest 10% samples within this pair, defined by the smallest absolute base-stage pairwise margin, so it focuses on examples closest to the local pairwise transition.

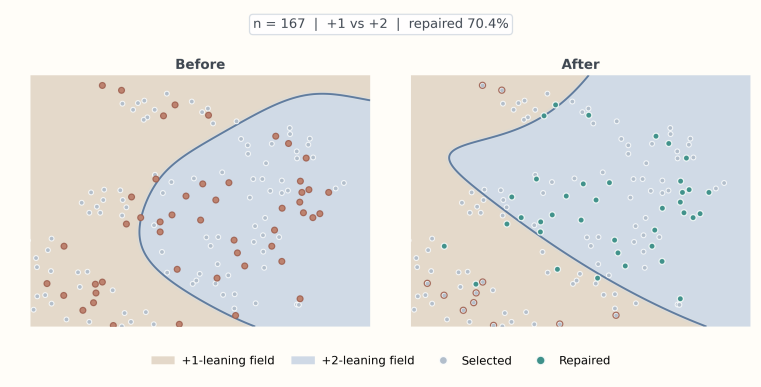

Figure 8: **Pairwise confidence-field visualization on CMU-MOSEI.** We visualize the $+1/+2$ class pair under the audio-only setting using t-SNE[42]. Background colors denote a smoothed pairwise confidence field leaning toward $+1$ or $+2$ in the visualization coordinates; the dark blue curve marks the equal-confidence contour $p(+1) = p(+2) = 0.5$. The subset contains the hardest 10% samples within this class pair, selected by the smallest absolute base-stage pairwise margin. The left panel marks initially misclassified samples, while the right panel highlights samples repaired after SCBC correction; residual orange circles indicate errors that remain unrepaired. Among 54 initial errors, 38 are repaired (70.4%).

D.9 Efficiency and Resource Overhead

To quantify the practical overhead of our method, we compare SCBC with IMDer and HME on CMU-MOSEI (7-class) under the fixed seven-subset protocol $\{\{l\},\{v\},\{a\},\{l,v\},\{l,a\},\{a,v\},\{l,v,a\}\}$. All runs use batch size 128, one NVIDIA A100 80GB GPU, and one training epoch. Runtime is reported as the full-epoch wall-clock time for each subset, and memory is the maximum per-process CUDA allocation during the timed training run. The same epoch-level accounting is used for all compared methods.

Figure 9 shows that SCBC stays in a 25.91–35.92 second range across all seven subsets, whereas IMDer ranges from 47.77 seconds on $\{l, v, a\}$ to 1471.46 seconds on the single-modality settings and HME stays around 229.03–234.10 seconds. This yields a $1.5\times$ speedup over IMDer for the complete-input setting and a $19.6\times$–$56.2\times$ speedup when one or two modalities are missing. Compared with HME, SCBC is $6.5\times$–$8.9\times$ faster across subsets. SCBC is also more memory-efficient, using only 0.62–0.64 GiB of maximum allocated CUDA memory versus 1.32–1.99 GiB for IMDer and 5.905 GiB for HME.

We additionally time the inference-only correction overhead over the corresponding base forward pass. The added latency is 0.000109, 0.003442, 0.002785, and 0.001475 ms per sample on CMU-MOSEI, Weibo-17, PHEME, and Hateful Memes, respectively.

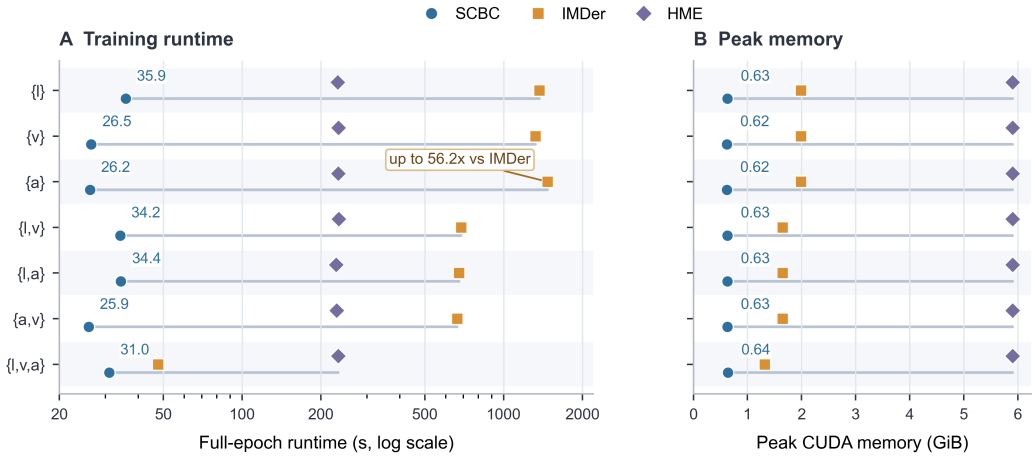

Figure 9: **Efficiency and resource overhead on CMU-MOSEI (7-class).** Runtime reports full-epoch wall-clock time, and memory reports peak per-process CUDA allocation; lower is better in both panels. All numbers are measured during training with batch size 128 on one NVIDIA A100 80GB GPU.

D.10 Theoretical Verification

Near-violation condition across benchmarks. Assumption 1(A3) requires that violation margins concentrate near the target threshold τ_m, so that the surrogate loss activation slope is approximately constant on the active set. This condition underpins the linearization in Theorem 1 and is empirically supported by Figure 1(c) for the IEMOCAP-4 $\{v\}$ setting. Table 19 extends this verification to all five benchmarks, reporting the fraction of misclassified samples whose negative logit gaps fall in $[-3, 0)$ and in the stricter interval $[-2, 0)$. The condition holds consistently across datasets and missingness protocols, with a mean of 82.6% under $[-3, 0)$ and 69.3% under $[-2, 0)$, confirming that the near-violation regime is a structural property of the failure mode rather than an artifact of any single setting.

Violation memory convergence. To verify that the modality-factorized memory converges to a stable estimate of $\mathbf{\Delta}_S$, we track the relative drift of the running factorized-memory estimate on IEMOCAP-4 fixed-missing settings. The main text reports the corresponding convergence plot, which shows that the aggregate drift decreases rapidly during training and remains in a low-drift regime in later epochs, consistent with the steady-state result derived in Appendix B.4. The write-after-read protocol ensures that each forward pass consumes historical rather than self-referential statistics, preventing the drift that would otherwise arise from online updates.

Per-setting direction-alignment breakdown. Table 20 reports the per-setting statistics behind the pooled averages quoted in the main text, including the non-vanishing event rate $\mathbb{P}(\mathcal{E}_S)$, the debt-weighted sign alignment, and the active-sample fraction Frac_q. The retrieved debt direction

Table 19: Empirical verification of the near-violation condition (Assumption 1(A3)) across all five benchmarks. We report the fraction of misclassified samples with negative logit gaps in $[-3, 0)$ and $[-2, 0)$, supporting the local-linearity approximation underlying Theorem 1. The IEMOCAP-4 $\{v\}$ entry matches Figure 1(c).

Setting	$[-3, 0)$ **(%)**	$[-2, 0)$ **(%)**
IEMOCAP-4 $\{v\}$ fixed	87.1	70.6
IEMOCAP-4 $\{a\}$ fixed	73.6	57.3
CMU-MOSEI $\{v\}$ fixed	93.1	81.4
CMU-MOSEI $\{a\}$ fixed	96.8	93.4
Weibo-17 $R_m = 0.5$	67.7	57.5
PHEME $R_m = 0.7$	91.1	73.3
Hateful Memes 50% image missing	68.6	51.5
Mean	**82.6**	**69.3**

remains active and consistently points toward descent across all five settings, while the cross-setting spread confirms that the pooled average is representative rather than dominated by any single setting.

Table 20: Per-setting empirical verification of direction alignment. We report representative settings: IEMOCAP-4 single-modality subsets $\{l\}$, $\{v\}$, and $\{a\}$, Weibo-17 at $R_m = 0.5$, and PHEME at $R_m = 0.7$. The non-vanishing event is operationalized as $\|\mathbf{q}_S\|_2 > 10^{-3}$.

Setting	$\mathbb{P}(\mathcal{E}_S)$ **on test (%)**	**Sign alignment (%)**	**Frac**$_q$ **(%)**
IEMOCAP-4 $\{l\}$	100.00	99.20	96.11
IEMOCAP-4 $\{v\}$	100.00	94.45	90.44
IEMOCAP-4 $\{a\}$	100.00	94.30	87.97
Weibo-17 ($R_m = 0.5$)	91.03	99.34	79.39
PHEME ($R_m = 0.7$)	100.00	85.60	72.17
Mean	**98.21**	**94.58**	**85.22**