BRACE: Bipolar Reference-Aware Calibration and Estimation for Incomplete Multimodal Learning

Ruiting Dai[1] **Zesen Cai**[1] **Xiaoyu Zhao**[1] **Yiting Huang**[2]
Lisi Mo[1] **Ming Li**[3] **Tao He**[1*]

[1]University of Electronic Science and Technology of China
[2]Nanyang Technological University
[3]The Chinese University of Hong Kong, Shenzhen
{rtdai, morris}@uestc.edu.cn, {2023091201025,zesen.cai}@std.uestc.edu.cn,
huan0387@e.ntu.edu.sg, ming.li@u.nus.edu, tao.he01@hotmail.com

Abstract

Real-world multimodal data are often incomplete. Existing incomplete multimodal methods mainly address this problem as evidence attenuation, by reconstructing missing views, regularising shared representations, or retrieving auxiliary context. However, partial observations can still leave a different uncertainty unresolved: even when they provide a coarse task-state anchor, they may not determine whether the latent state should be revised toward a higher or lower label. We formalize this overlooked failure mode as residual-direction ambiguity and propose **BRACE** (Bipolar Reference-Aware Calibration and Estimation), a framework that first builds a quality-aware unified task state from available modalities, then organizes historical state-label pairs in a global memory bank, and finally infers a gated latent correction from the ordered contrast between higher-label and lower-label reference neighborhoods, together with a warmup-then-calibration training strategy and direction-consistency regularization. Across four benchmarks spanning multimodal sentiment analysis and multimodal misinformation detection under fixed and random missingness, BRACE consistently outperforms strong baselines, improving Acc-2 by up to 7.4 points over HME under fixed missingness and remaining robust when key modalities are absent. Code is available at https://github.com/Js183/BRACE.

1 Introduction

Multimodal data in the wild are often incomplete. Social posts may lose images during crawling, short videos may lack audio, and user records may be partially unavailable because of privacy control [1, 2], transmission failure, or data compression [3, 4]. Incomplete multimodal learning therefore aims to make reliable predictions from only a subset of available modalities, a setting that is common in multimodal sentiment analysis and misinformation detection [5–7].

Existing methods mainly treat missingness as an *evidence attenuation* problem: different observed modality subsets contain different amounts of label-relevant information. Reconstruction methods synthesize missing modalities [8–12]; representation-based methods improve robustness through masked fusion, cross-modal regularization, distillation, or modality enhancement [13–16]; and retrieval- or prompt-augmented methods introduce auxiliary samples as demonstrations, prompts, or associative context [6, 7, 17, 18]. These methods improve the amount or quality of available evidence, but they do not explicitly answer a different question: once a coarse prediction has been made, in which direction should the latent state be corrected?

*Corresponding author.

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

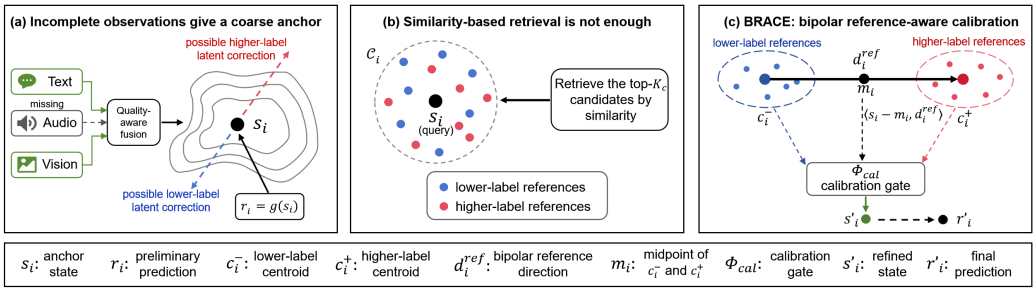


Figure 1: **Motivation of BRACE**. (a) Incomplete observations can provide a coarse task state s_i and preliminary prediction r_i, but leave the residual correction direction ambiguous. (b) Similarity-based retrieval finds nearby samples, but does not indicate whether the state should move toward lower or higher labels. (c) BRACE constructs bipolar label-ordered references, estimates a local lower-to-higher direction d_i^{ref}, and uses a calibration gate Φ_{cal} to produce a signed latent update, yielding the refined state s_i' and final prediction r_i'.

We identify this as a second failure mode of incomplete multimodal learning. The observed modalities may produce a reasonable anchor state s_i and preliminary prediction r_i, while the missing modalities determine whether the prediction should be revised upward or downward. We call this *residual-direction ambiguity*, and formalize it as *sign non-identifiability of the residual* (Def. 2.1): conditioned on the observed modalities, both higher-label and lower-label corrections remain plausible. As shown in Fig. 1(a), the prediction can be close in magnitude but still uncertain in correction direction.

This view also explains why ordinary retrieval is insufficient. Similarity-based retrieval can find samples near the current state, but similarity alone has no label order [17–20]. Nearby references may contain both higher- and lower-label samples, without indicating which side the current state should move toward. Thus, retrieval provides context about what the sample resembles, but not a signed cue for how its task state should be revised, as illustrated in Fig. 1(b).

We address this problem with **BRACE**, a *Bipolar Reference-Aware Calibration and Estimation* framework. BRACE treats labeled historical states as a local coordinate system in a locally structured task-state space. Around an anchor s_i, it retrieves nearby state-label pairs and partitions them into higher- and lower-label neighborhoods, i.e., references whose labels lie above or below the preliminary prediction r_i by a margin. The centroid difference between these two neighborhoods, normalized by their label gap, defines a local lower-to-higher direction in latent space. A calibration gate then uses this bipolar direction and the anchor's relative position between the two centroids to predict a signed latent update (Fig. 1(c)).

BRACE reframes incomplete multimodal learning as an estimation-and-calibration problem: the observed modalities estimate a coarse task state, and label-ordered references calibrate this state toward lower task risk. Under local regularity assumptions, we show that the bipolar reference direction is aligned in expectation with a covariance-scaled local gradient of the preliminary prediction head (Thm. 1). We further show that the anchor's relative position between the two centroids can recover the residual sign under a bounded local-skewness condition (Prop. 3.1), and that the direction-consistency loss acts as a task-level projection of the gradient-related condition required for local risk reduction (Prop. 3.2). These results provide a mechanism-level justification for using bipolar reference geometry as a signed calibration signal.

BRACE consists of three stages. First, it builds a quality-consistent anchor through availability-masked and reliability-weighted fusion. Second, it maintains a global memory bank of historical state-label pairs, together with EMA tail prototypes for sparse neighborhoods. Third, it constructs bipolar references and applies a gated calibration map to produce the refined state and prediction. Training uses a warmup phase to stabilize the anchor and memory, followed by a direction-consistency objective that encourages the refined prediction to move in the correct residual direction.

Related work. Incomplete multimodal learning divides into two sample-centric families that complete or regularize the representation inferred from the observation itself. *Reconstruction* methods recover absent modalities through translation, imagination, graph completion, or diffusion-based recovery [8–12], while *representation-based* methods strengthen the latent space via relative-advantage learning, proxy regularization, or modality enhancement [13–16]. Earlier misinformation-detection work

targets complete or weakly incomplete observations through cross-modal ambiguity modelling, graph fusion, topic-aware reasoning, and social context [21–28]. A recent retrieval-based line, including Borrowing Treasures (in-context learning), RAGPT (dynamic prompts), and IMOL (cross-sample associative reasoning) [17, 18, 6], pulls in other samples but injects them as demonstrations, prompts, or auxiliary supervision: a similarity signal about what the sample resembles, with no label-ordered geometric structure indicating how its latent state should be revised.

Contributions. We *(i)* identify residual-direction ambiguity, or sign non-identifiability of the residual, as an overlooked failure mode in incomplete multimodal learning; *(ii)* propose BRACE, a bipolar reference-aware calibration framework that converts retrieved labeled neighbors into signed latent correction rather than undirected context; and *(iii)* provide local theoretical analysis and validate BRACE on four benchmarks under fixed and random missingness, where it achieves the strongest performance among baselines, especially in low-information and high-missingness regimes.

2 Preliminaries

Incomplete multimodal prediction. Let $\mathcal{M}$ denote the set of modalities and let $\delta_i^{(m)} \in \{0,1\}$ indicate whether modality m is observed for sample i. The available modality set is $\mathcal{A}_i = \{m : \delta_i^{(m)} = 1\}$, and the observed input is $x_i^{(\mathcal{A}_i)} = \{x_i^{(m)}\}_{m\in\mathcal{A}_i}$. We consider ordered scalar labels $y_i \in \mathbb{R}$, which directly cover regression tasks and also include classification tasks through an ordinal label mapping [6, 13]. An incomplete multimodal model first maps the available modalities to a latent task state $s_i \in \mathbb{R}^d$, and then produces a preliminary prediction $r_i = g(s_i)$, where $g : \mathbb{R}^d \to \mathbb{R}$ is a scalar prediction head. A calibration module further refines the anchor state into s_i' and outputs the final prediction $r_i' = g(s_i')$. The per-sample task risk is $\mathcal{R}_i(s) = \ell(g(s), y_i)$, where ℓ is differentiable, strictly convex, and ordered in its first argument, that is, $\operatorname{sign}\big(\ell'(a, y_i)\big) = \operatorname{sign}(a - y_i), \quad a \neq y_i$. Thus, $\nabla_s \mathcal{R}_i(s) = \ell'(g(s), y_i)\nabla g(s)$.

Residual-direction ambiguity. Most incomplete multimodal learning methods view missingness as an *evidence attenuation* problem: different observed subsets $\mathcal{A}_i$ preserve different amounts of label-relevant information, and the model should reconstruct, regularize, or supplement the missing evidence [6, 10, 12, 18]. This view is important, but incomplete. Once a model has formed a coarse prediction, a second question emerges: *in which direction should the current latent state be corrected?*

We formalize this failure mode as sign non-identifiability of the residual. Intuitively, the observed modalities may be sufficient to locate the sample near the right prediction scale, but insufficient to decide whether the prediction should move upward or downward.

Definition 2.1 (Sign Non-Identifiability of the Residual)**.** Let $\pi_i = \Pr[y_i > r_i \mid x_i^{(\mathcal{A}_i)}]$ be the conditional probability, given the observed modalities, that the preliminary prediction underestimates the label. The residual $y_i - r_i$ is *sign-non-identifiable from* $\mathcal{A}_i$ at level $\varepsilon_0 \in (0, \frac{1}{2})$ if $\pi_i \in (\varepsilon_0, 1-\varepsilon_0)$.

This definition is stronger than saying that the latent gradient has high uncertainty: it requires the sign of the scalar residual itself to remain ambiguous under the observed modalities. In practice, π_i is not directly observable. We estimate it with an auxiliary classifier $\widehat{\pi}(x_i^{(\mathcal{A}_i)})$ trained on a held-out fold to predict $\mathbb{1}[y_i > r_i]$, and report the fraction of samples satisfying $\widehat{\pi} \in (\varepsilon_0, 1-\varepsilon_0)$. Appendix D shows that such ambiguous samples form a non-negligible portion under missingness.

Latent calibration as local risk descent. Residual-direction ambiguity suggests that incomplete multimodal inference should not only estimate a prediction, but also estimate a *calibration displacement*. Let $u_i = s_i' - s_i$ be the latent update applied to the anchor s_i. The ideal local correction $s_i^\star \in \arg\min_{\|\Delta\|_2 \le \rho} \mathcal{R}_i(s_i + \Delta)$ is unavailable at test time because it depends on y_i. We therefore seek a first-order surrogate: an update that points toward lower task risk.

Definition 2.2 (Gradient-Related Update)**.** A displacement u_i is *gradient-related* to $\mathcal{R}_i$ at s_i if

$$\langle \nabla_s \mathcal{R}_i(s_i), u_i \rangle \le -\kappa_i \|u_i\|_2^2 \quad \text{for some } \kappa_i > 0. \tag{1}$$

Proposition 2.3 (Smooth Descent)**.** *If $\mathcal{R}_i$ is L_i-smooth on $B(s_i, \rho_0)$ and u_i satisfies Eq.* (1)*, then for any $0 < \alpha \le \min\{1, 2\kappa_i/L_i\}$,*

$$\mathcal{R}_i(s_i + \alpha u_i) \le \mathcal{R}_i(s_i) - \alpha\Big(\kappa_i - \frac{L_i\alpha}{2}\Big)\|u_i\|_2^2. \tag{2}$$

Prop. 2.3 is a standard descent result. The key challenge is not the descent lemma itself, but how to construct an update u_i whose direction is gradient-related without observing the test label y_i. BRACE addresses this by constructing a label-ordered local geometry from historical state-label pairs.

Quality-consistent anchors. The local analysis requires that the anchor state vary smoothly with the observed modalities and does not contain artificial contributions from absent ones.

Definition 2.4 (Quality-Consistent Anchor)**.** Given modality summaries $h_i^{(m)} \in \mathbb{R}^d$ and fusion weights $\alpha_i^{(m)} \in [0,1]$, the state $s_i = \sum_{m\in\mathcal{M}} \alpha_i^{(m)} h_i^{(m)}$ is *quality-consistent* if: (i) $\alpha_i^{(m)} = 0$ for $m \notin \mathcal{A}_i$; (ii) $\sum_m \alpha_i^{(m)} = 1$; and (iii) the mapping $x_i^{(\mathcal{A}_i)} \mapsto s_i$ is C^2 with bounded second derivative on the relevant support.

Conditions (i) and (ii) prevent missing modalities from leaking into the anchor, while condition (iii) provides the local regularity needed for the first-order calibration analysis.

Bipolar reference geometry. To resolve the missing sign of the residual, BRACE uses other labeled samples as a local coordinate system. Given a memory bank $\mathcal{B}$ of state-label references, let $\mathcal{C}_i \subset \mathcal{B}$ be a local candidate set around s_i. For a margin $\delta_y > 0$, we split the candidates by label relative to the preliminary prediction:

$$\mathcal{N}_i^+ = \{j \in \mathcal{C}_i : y_j - r_i \geq \delta_y\},\ \ \mathcal{N}_i^- = \{j \in \mathcal{C}_i : r_i - y_j \geq \delta_y\}.$$

With normalized nonnegative weights $\omega_{ij}^{\pm}$, define the higher- and lower-label centroids and their label means as $c_i^{\pm} = \sum_{j\in\mathcal{N}_i^{\pm}} \omega_{ij}^{\pm} v_j,\ \ \bar{y}_i^{\pm} = \sum_{j\in\mathcal{N}_i^{\pm}} \omega_{ij}^{\pm} y_j$. The bipolar reference direction is

$$d_i^{\mathrm{ref}} = \frac{c_i^+ - c_i^-}{\bar{y}_i^+ - \bar{y}_i^- + \varepsilon}, \qquad \varepsilon > 0. \tag{3}$$

The numerator points from lower-label references to higher-label references in the local latent space, while the denominator rescales this displacement by the corresponding label gap. Eq. (3) is related to two classical ideas: sliced inverse regression [29, 30], which estimates label-conditioned directions by slicing the response variable, and zeroth-order gradient estimation [31, 32], which estimates slopes from finite differences. BRACE differs in how this geometry is used: the two-slice direction becomes a *test-time calibration operator* on a learned multimodal state, and a learned gate recovers the missing residual sign from observable reference geometry.

The method must therefore satisfy three requirements: construct a smooth and mask-consistent anchor, provide a local density of labeled references around that anchor, and convert the bipolar direction into a signed test-time update without using the test label.

3 Method

BRACE implements the requirements with three coupled stages. Stage 1 constructs a quality-consistent anchor supporting local smoothness. Stage 2 maintains a global state-label memory, which supplies local label-ordered reference density. Stage 3 turns the resulting bipolar geometry into a signed latent update, and the training objective encourages this update to satisfy the gradient-related condition in Eq. (1).

3.1 Stage 1: Quality-Consistent Anchor Construction

Bipolar calibration is meaningful only if nearby latent states are comparable across different missing-modality patterns. Naive fusion can violate this requirement: absent modalities may still contribute through learned defaults, and the representation may encode the missingness mask rather than the task state. BRACE therefore first constructs a quality-consistent anchor by combining availability masking with reliability-weighted fusion.

Each observed modality is encoded into a compact slot sequence: $Z_i^{(m)} = \mathrm{Pool}^{(m)}\big(\mathrm{Conv1D}^{(m)}(\mathrm{Enc}^{(m)}(x_i^{(m)}))\big) + E_{\mathrm{slot}}, h_i^{(m)} = \mathrm{Mean}_{\mathrm{slot}}(Z_i^{(m)})$. A reliability score is computed for each modality: $q_i^{(m)} = w_q^\top \tanh(W_q h_i^{(m)} + b_q)$, and the availability-masked fusion weight is

$$\alpha_i^{(m)} = \frac{\delta_i^{(m)} \exp(q_i^{(m)})}{\sum_{n\in\mathcal{M}} \delta_i^{(n)} \exp(q_i^{(n)})}. \tag{4}$$

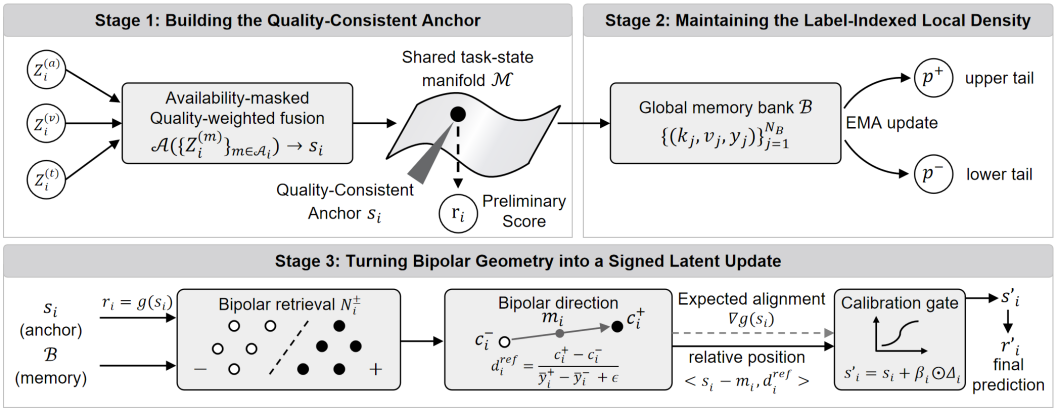


Figure 2: **Overview of BRACE**. Stage 1 builds a quality-consistent anchor s_i from available modalities. Stage 2 maintains a global memory of state-label references and label-tail prototypes for local bipolar retrieval. Stage 3 partitions retrieved references into lower- and higher-label neighborhoods, estimates the bipolar direction d_i^{ref}, and applies a calibration gate to produce a signed latent update, yielding the refined state s_i' and prediction r_i'.

The seed state and observed modality slots are then refined by L masked self-attention blocks:

$$\mathcal{J}_i^{(0)} = \Big[\sum_m \alpha_i^{(m)} h_i^{(m)} + e_0,\ \{\delta_i^{(m)}(Z_i^{(m)} + e_m)\}_{m\in\mathcal{M}}\Big],\ \mathcal{J}_i^{(\ell)} = \mathrm{JointBlock}^{(\ell)}(\mathcal{J}_i^{(\ell-1)}),\ s_i = \mathcal{J}_i^{(L)}[0] \quad (5)$$

Here e_0 and e_m are learned type embeddings. Eq. (4) ensures that absent modalities receive zero fusion mass and that the remaining weights sum to one. Since absent slots are zeroed and masked inside attention, $\nabla_{x_i^{(m)}} s_i = 0$ for $m \notin \mathcal{A}_i$. The anchor is therefore a smooth function of the observed inputs on the relevant support, satisfying Def. 2.4.

3.2 Stage 2: Global State-Label Memory

The bipolar direction in Eq. (3) is a local finite-difference estimate, so it requires labeled references near the anchor s_i. BRACE maintains a single global memory bank $\mathcal{B} = \{(k_j, v_j, y_j)\}_{j=1}^{N_{\mathcal{B}}}$, $k_j = \mathrm{Normalize}(W_k v_j)$, where v_j is a historical anchor state and y_j is its training label. A global bank is important: it makes the reference geometry a property of the learned task-state space rather than a property of a particular missingness pattern. In contrast, separate memories for different modality subsets can preserve mask-specific geometry and weaken cross-pattern calibration.

For each query state s_i, BRACE retrieves the top-K_c candidates $\mathcal{C}_i$ by similarity between $k_i = \mathrm{Normalize}(W_k s_i)$ and memory keys. The desired locality condition is

$$\Pr_{j\in\mathcal{C}_i}\big[\|v_j - s_i\|_2 \le \rho_0\big] \ge 1 - \zeta, \quad (6)$$

for a small $\zeta > 0$. This condition is not an architectural assumption on W_k; it is the empirical requirement that retrieval in key space remains local in the latent state space. Appendix D verifies this locality diagnostically.

Sparse-neighborhood fallback. When one side of the bipolar split is empty, Eq. (3) becomes unstable. We therefore maintain exponential-moving-average prototypes p_+ and p_- for the upper and lower label tails. Given a tail quantile $q \in (0, \frac{1}{2})$, momentum μ, and label range $[y_{\min}, y_{\max}]$, define thresholds $\tau_+ = y_{\min} + (1-q)(y_{\max} - y_{\min})$, $\tau_- = y_{\min} + q(y_{\max} - y_{\min})$. The tail prototypes are updated by $p_\pm \leftarrow \mu p_\pm + (1-\mu)\bar{v}_\pm$, where $\bar{v}_+$ averages write-set states with $y \ge \tau_+$, and $\bar{v}_-$ averages write-set states with $y \le \tau_-$. If $\mathcal{N}_i^+$ or $\mathcal{N}_i^-$ is empty, the corresponding prototype serves as fallback. If a prototype is not yet initialized, the anchor s_i is used, causing the update to shrink rather than produce an arbitrary direction.

The memory uses only training states and labels. After warmup, the bank is reset and bootstrapped over the training loader. During training, mini-batch anchors are appended online. During validation and testing, the bank is frozen and read-only, so no validation or test label enters memory.

3.3 Stage 3: Bipolar Reference-Aware Calibration

Stage 3 converts label-ordered reference geometry into a signed latent update. It has two roles. First, the bipolar direction should approximate an increasing-label direction around s_i. Second, the gate must decide whether the sample should move along or against this direction without observing y_i.

Local regularity. We state the assumptions used to justify the calibration mechanism.

Assumption 1 (Local Regularity)**.** *There exists $\rho_0 > 0$ such that, on $B(s_i, \rho_0)$: **(A1)** $s(\cdot)$ is C^2 on the observed-input support and g is C^2 with $\|\nabla g(s_i)\| > 0$; **(A2)** retrieved states are sampled from a local density with positive-definite covariance Σ_i, bounded third moments, and bounded conditional density under the label truncations defining $\mathcal{N}_i^{\pm}$; **(A3)** reference labels satisfy $y_j = g(v_j) + \eta_j$, where η_j is mean-zero sub-Gaussian and independent of v_j; **(A4)** ℓ is differentiable, strictly convex, and ordered in its first argument, with $\ell'(g(s_i), y_i) \neq 0$.*

A1 is supplied by Stage 1, and A2 is supported by the global memory and the retrieval locality condition in Eq. (6). A3 is the local label-noise model used to connect label-conditioned reference geometry to the prediction head.

Theorem 1 (Bipolar Direction Alignment)**.** *Under Asm. 1, choose ρ_0 small enough that the Taylor remainder of g on $B(s_i, \rho_0)$ is $O(\rho_0^2)$. As $|\mathcal{N}_i^+|, |\mathcal{N}_i^-| \to \infty$,*

$$\mathbb{E}[d_i^{\mathrm{ref}}] = \frac{\Sigma_i \nabla g(s_i)}{\|\nabla g(s_i)\|_{\Sigma_i}^2} + O(\rho_0^2) + O_P(|\mathcal{N}_i^{\pm}|^{-1/2}). \tag{7}$$

Consequently, $\langle \nabla g(s_i), \mathbb{E}[d_i^{\mathrm{ref}}] \rangle = 1 + O(\rho_0^2) > 0$. If the residual sign were known, the oracle update $\hat{u}_i = \mathrm{sign}(y_i - r_i) d_i^{\mathrm{ref}}$ would satisfy the gradient-related condition in Eq. (1) *in expectation, with $\kappa_i = c_i |\ell'(r_i, y_i)|$, where $c_i > 0$ depends on local covariance and curvature terms.*

Thm. 1 explains why the lower-to-higher centroid difference is a valid calibration direction: in expectation, it is aligned with the local gradient of the prediction head. However, the theorem also exposes the central test-time difficulty: the oracle update requires the unknown sign of $y_i - r_i$.

Recovering the sign from relative position. The test-time observable that carries sign information is the anchor's relative position between the lower and higher centroids. Let $m_i = \frac{1}{2}(c_i^+ + c_i^-)$ be the midpoint of the bipolar frame, and define the local skewness $\xi_i = \frac{1}{2}(\bar{y}_i^+ + \bar{y}_i^-) - y_i$.

Proposition 3.1 (Sign Identifiability from Relative Position)**.** *Under Asm. 1 and the regime of Thm. 1, $\mathbb{E}\big[\langle s_i - m_i, d_i^{\mathrm{ref}} \rangle\big] = -(r_i - y_i) + \xi_i + O(\rho_0^2)$. Thus, the sign of $\langle s_i - m_i, d_i^{\mathrm{ref}} \rangle$ agrees with $\mathrm{sign}(y_i - r_i)$ in expectation whenever $|\xi_i| < |y_i - r_i| + O(\rho_0^2)$. Under sub-Gaussian noise with scale σ_i, the agreement probability is at least $1 - \Phi\left(-\frac{|y_i - r_i| - |\xi_i|}{\sigma_i}\right)$.*

Prop. 3.1 justifies exposing the full bipolar frame to the calibration gate. The vector d_i^{ref} tells the gate where labels increase, while the relative location of s_i with respect to c_i^- and c_i^+ provides an observable cue for whether the current sample should move upward or downward.

Gated signed update. For each query, BRACE constructs similarity-weighted bipolar centroids:

$$\omega_{ij}^{\pm} = \frac{\exp(\tau \,\mathrm{sim}(k_i, k_j))}{\sum_{\ell \in \mathcal{N}_i^{\pm}} \exp(\tau \,\mathrm{sim}(k_i, k_\ell))}, \; c_i^{\pm} = \sum_{j \in \mathcal{N}_i^{\pm}} \omega_{ij}^{\pm} v_j, \qquad \bar{y}_i^{\pm} = \sum_{j \in \mathcal{N}_i^{\pm}} \omega_{ij}^{\pm} y_j. \tag{8}$$

The gate receives the anchor, both centroids, and the unnormalized bipolar displacement: $z_i = [s_i; c_i^+; c_i^-; c_i^+ - c_i^-]$. It predicts a per-dimension step gate and a signed displacement:

$$\beta_i = \sigma(\phi_\beta(z_i)), \qquad \Delta_i = \phi_\Delta(z_i), \qquad u_i = \beta_i \odot \Delta_i, \qquad s_i' = s_i + u_i, \qquad r_i' = g(s_i'). \tag{9}$$

The sigmoid gate $\beta_i \in (0,1)^d$ controls the update magnitude, while Δ_i supplies the signed direction. Because z_i contains both the bipolar direction and the anchor's relative position in the frame, the gate can learn the sign-recovery rule suggested by Prop. 3.1 without using the test label.

3.4 Training Objective

Training has two phases. The warmup phase stabilizes the anchor space before reference calibration is activated. This is important because early memory states are noisy and can violate the locality

requirement in Eq. (6). For T_{warm} epochs, BRACE disables the calibration path and trains only the anchor prediction: $\mathcal{L}_{\text{warm}} = \frac{1}{N}\sum_i |r_i - y_i|$. After warmup, the memory bank is reset, bootstrapped with training anchors, and then updated online during training.

The full training objective combines prediction accuracy with direction consistency. Let $\mathcal{P} = \{i : y_i > r_i\}$, $\mathcal{N} = \{i : y_i < r_i\}$. For a margin $\gamma > 0$, define $\ell_i^{\text{dir}} = \text{ReLU}\big(\gamma - \text{sign}(y_i - r_i)(r_i' - r_i)\big)$. The task and direction losses are $\mathcal{L}_{\text{task}} = \frac{1}{N}\sum_i |r_i' - y_i|$, $\mathcal{L}_{\text{dir}} = \frac{1}{2}\Big(\frac{1}{|\mathcal{P}|}\sum_{i\in\mathcal{P}} \ell_i^{\text{dir}} + \frac{1}{|\mathcal{N}|}\sum_{i\in\mathcal{N}} \ell_i^{\text{dir}}\Big)$. where empty sides are dropped. The final objective is $\mathcal{L}_{\text{full}} = \lambda_{\text{task}}\mathcal{L}_{\text{task}} + \lambda_{\text{dir}}\mathcal{L}_{\text{dir}}$.

The direction loss is not merely an auxiliary regularizer. It is the training mechanism that converts the bipolar frame into a gradient-related update at the task level.

Proposition 3.2 (Direction Loss as Task-Level Gradient Projection)**.** *Let $u_i = s_i' - s_i$ and $r_i' = g(s_i')$. Assume that g has L_g-Lipschitz gradient on the segment from s_i to s_i', and that the ordered loss satisfies* $\text{sign}(\ell'(r_i, y_i)) = \text{sign}(r_i - y_i)$*. If $\ell_i^{\text{dir}} \le \eta_i$, then*

$$\text{sign}(y_i - r_i)\langle \nabla g(s_i), u_i\rangle \ge \gamma - \eta_i - \frac{L_g}{2}\|u_i\|_2^2. \tag{10}$$

Consequently, if $u_i \neq 0$ and the right-hand side of Eq. (10) *is positive, then u_i satisfies Eq.* (1) *with* $\kappa_i = \frac{|\ell'(r_i, y_i)|}{\|u_i\|_2^2}\left(\gamma - \eta_i - \frac{L_g}{2}\|u_i\|_2^2\right) > 0$.

Prop. 3.2 closes the loop between the theory and the learned model: Thm. 1 shows that the bipolar direction estimates a local increasing-label direction, Prop. 3.1 shows that the observable reference frame contains information about the missing residual sign, and Prop. 3.2 shows that minimizing the direction loss encourages the learned update to become gradient-related. Thus, BRACE turns retrieval from undirected context into a theoretically motivated signed calibration operator.

4 Experiments

4.1 Experimental Setup

Benchmarks and protocols. We evaluate BRACE on four benchmarks across two tasks. For multimodal sentiment analysis, we use **CMU-MOSEI** [33] and **CMU-MOSI** [34], which contain language, acoustic, and visual modalities. We follow the fixed missing protocol and evaluate all seven nonempty modality subsets with Acc-2 and F1. For multimodal rumor detection, we use **Weibo-17** [35] and **PHEME** [36]. We follow the random missing protocol of [7], where each sample keeps at least one modality, and report ACC at different target missing rates R_m.

Baselines. We compare with reconstruction based methods MMIN [9], GCNet [12], IMDer [10], and DiCMoR [11]; representation and distillation methods LNLN [13], TF-Mamba [37], CMAD [38], HME [16], and RedCore [14]; and retrieval or prompt based methods RAGPT [18] and TriSPrompt [7]. For rumor detection, we also include CAFE [21], CLKD [39], and GameOn [22].

Implementation. We follow the feature settings of prior work [7, 10, 12]. For CMU-MOSEI and CMU-MOSI, we use DeBERTa-large [40], MTCNN with MA-Net [41, 42], and wav2vec-large [43] for text, visual, and audio features. For Weibo-17 and PHEME, we use BERT [44] text features and ResNet-34 [45] image features, with an extra comment stream for PHEME. BRACE is trained with AdamW using a warmup-cosine schedule, and the main results are averaged over five runs. More details are provided in Appendices C and D.

4.2 Comparison with State of the Art

Fixed missing sentiment analysis. Table 1 reports CMU-MOSI and CMU-MOSEI results over all modality subsets. BRACE improves average Acc-2/F1 by $+0.6/+0.6$ over HME on CMU-MOSI and by $+2.8/+3.0$ over CMAD on CMU-MOSEI. The gains are largest in low-information subsets such as $\{a\}$, $\{v\}$, and $\{v, a\}$, while they become smaller when all modalities are observed. This supports our claim that BRACE is most useful when the observed modalities provide a weak anchor and the correction direction is hard to infer.

Random missing rumor detection. Figure 3 shows that BRACE achieves the best ACC at every missing rate, with gains of 0.4–1.7 on Weibo-17 and 0.7–5.3 on PHEME. The largest gain appears

Table 1: Results on **CMU-MOSI** and **CMU-MOSEI** under the fixed-missing protocol. Entries are Acc-2/F1. For MOSEI and MOSI rows, green superscripts denote gain over the second-best baseline.

	Avail.	MMIN	GCNet	IMDer	DiCMoR	LNLN	TF-Mamba	CMAD	HME	BRACE
CMU-MOSI	$\{l\}$	83.8/83.8	83.7/83.6	84.8/84.7	84.5/84.4	84.9/85.1	84.6/84.5	83.8/83.9	85.7/85.6	**86.3**$^{+0.6}$/**86.3**$^{+0.7}$
	$\{v\}$	57.0/54.0	56.1/55.7	61.3/60.8	62.2/60.2	52.2/58.9	57.0/54.4	53.8/53.1	63.6/63.4	**64.2**$^{+0.6}$/**64.4**$^{+1.0}$
	$\{a\}$	55.3/51.5	56.1/54.5	62.0/62.2	60.5/60.8	52.2/58.9	55.5/54.9	53.7/51.2	62.7/62.2	**63.6**$^{+0.9}$/**62.2**$^{+0.0}$
	$\{l, v\}$	83.8/83.9	84.3/84.2	85.5/85.4	85.5/85.4	84.9/85.1	83.8/83.9	82.5/82.5	85.8/85.8	**86.7**$^{+0.9}$/**86.6**$^{+0.8}$
	$\{l, a\}$	84.0/84.0	84.5/84.4	85.4/85.3	85.5/85.5	84.9/85.2	82.3/82.4	84.0/84.0	85.8/85.7	**86.6**$^{+0.8}$/**86.5**$^{+0.8}$
	$\{v, a\}$	60.4/58.5	62.0/61.9	63.6/63.4	64.0/63.5	52.2/58.9	52.6/51.3	60.2/60.4	64.6/64.5	**65.1**$^{+0.5}$/**65.2**$^{+0.7}$
	$\{l, v, a\}$	84.6/84.4	85.2/85.1	85.7/85.6	85.7/85.6	84.3/84.6	83.7/83.7	82.2/82.1	86.4/86.3	**86.6**$^{+0.2}$/**86.6**$^{+0.3}$
	Avg.	72.7/71.4	73.1/72.8	75.5/75.3	75.4/75.1	70.8/73.8	71.4/70.7	71.5/71.0	76.4/76.2	**77.0**$^{+0.6}$/**76.8**$^{+0.6}$
CMU-MOSEI	$\{l\}$	82.3/82.4	83.0/83.2	84.5/84.5	84.2/84.3	84.1/84.4	83.7/83.6	86.1/86.0	85.4/85.4	**86.5**$^{+0.4}$/**86.3**$^{+0.3}$
	$\{v\}$	59.3/60.0	61.9/61.6	63.9/63.6	63.6/63.6	64.5/60.8	61.8/60.3	65.7/64.4	64.6/62.9	**68.5**$^{+2.8}$/**68.0**$^{+3.6}$
	$\{a\}$	58.9/59.5	60.2/60.3	63.8/60.6	62.9/60.4	62.9/61.8	62.9/48.5	63.0/60.8	64.1/59.7	**71.4**$^{+7.3}$/**70.3**$^{+8.5}$
	$\{l, v\}$	83.8/83.4	84.3/84.4	85.0/85.0	84.9/84.9	85.8/85.5	83.3/83.2	86.4/86.4	85.7/85.7	**86.5**$^{+0.1}$/**86.4**$^{+0.0}$
	$\{l, a\}$	83.7/83.3	84.3/84.4	85.1/85.1	85.0/84.9	85.6/85.4	83.1/82.7	86.3/86.2	85.6/85.4	**86.5**$^{+0.2}$/**86.5**$^{+0.3}$
	$\{v, a\}$	63.5/61.9	64.1/57.2	64.9/63.5	65.2/64.4	64.8/62.7	60.7/61.2	65.6/64.8	65.4/63.7	**72.8**$^{+7.2}$/**72.0**$^{+7.2}$
	$\{l, v, a\}$	84.3/84.2	85.2/85.1	85.1/85.1	85.1/85.1	85.5/85.2	83.8/83.7	86.1/86.2	86.2/86.2	**86.6**$^{+0.4}$/**86.5**$^{+0.3}$
	Avg.	73.7/73.5	74.7/73.7	76.0/75.3	75.8/75.4	76.2/75.1	74.2/71.9	77.0/76.4	76.7/75.6	**79.8**$^{+2.8}$/**79.4**$^{+3.0}$

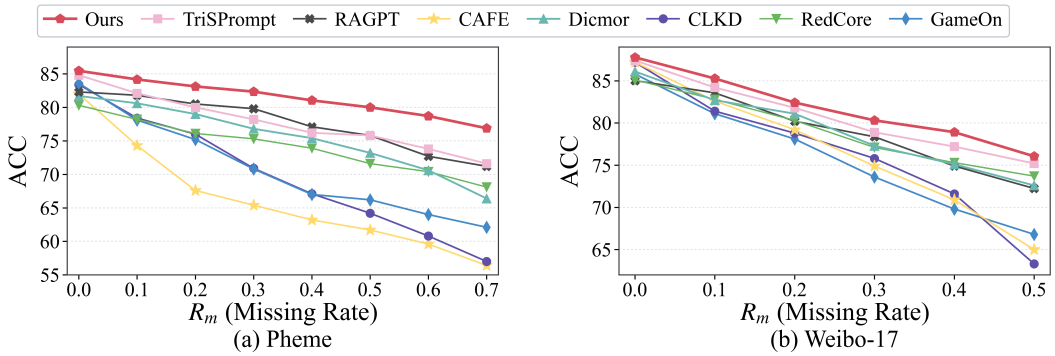


Figure 3: ACC under random missingness on **PHEME** and **Weibo-17** under different missing rates.

under high missingness on PHEME, which is consistent with our motivation: as missingness increases, residual direction ambiguity becomes more frequent, and signed calibration becomes more useful.

4.3 Controlled Ablation Studies

We next test whether BRACE gains come from signed reference calibration rather than generic retrieval, memory access, or extra parameters. Additional retrieval, training, and sensitivity results are provided in Appendix D.

Bipolar label order. Table 2 compares BRACE with controls that remove label order, locality, learned calibration, or reference inputs on PHEME at $R_m = 0.5$. Full BRACE performs best across all reported metrics, showing that the gain consistently comes from label-ordered signed calibration rather than retrieval, memory size, or an extra gate alone.

Core components. Table 3 shows that removing Ref-Cal, Q-Init, or both reduces Acc-2 by 0.8, 1.0, and 1.5, respectively. The nonadditive drops suggest that Q-Init provides a smoother and more mask-consistent anchor for reliable reference calibration.

Q-Init under severe missingness. Figure 4 shows that Q-Init consistently improves Acc-2 and Acc-7 on CMU-MOSI at $R_m = 0.5, 0.6, 0.7$, reducing degradation under weak observed evidence.

Binary label encoding. Because binary rumor labels have no inherent order, we swap class encodings on PHEME and rerun BRACE. Table 4 shows stable performance, with average ACC/AUC changing by only $-0.2/-0.3$; F1 is omitted because it depends on the chosen positive class. This suggests that BRACE does not rely on a particular semantic ordering of the two class IDs. Instead, the bipolar calibration mainly uses the two reference groups to estimate a relative correction direction under the current encoding, and the learned decision boundary adapts accordingly after label swapping.

Variant	Local retrieval	Label order	Signed gate	ACC	F1	ΔF1
Retrieval only context	✓			78.7	63.6	-1.4
Random memory		✓	✓	78.9	63.8	-1.2
Label shuffled memory	✓		✓	79.2	63.9	-1.1
kNN residual regression	✓	✓		78.4	62.1	-2.9
Gate without references			✓	78.6	62.2	-2.8
Full BRACE	✓	✓	✓	**80.0**	**65.0**	–

Table 2: Controlled retrieval and memory ablation on **PHEME** at $R_m = 0.5$. ΔF1 is relative to Full BRACE.

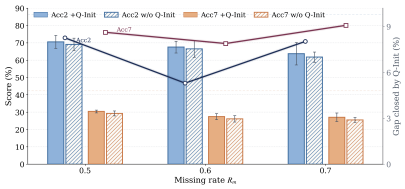


Figure 4: Effect of Q-Init on **CMU-MOSI** under random missingness.

Table 3: Ablation on **MOSI** with $\{v\}$.

Variant	Q-Init	Ref-Cal	Acc-2	F1	ΔAcc-2
Full	✓	✓	**64.2**	**64.4**	–
w/o Ref-Cal	✓		63.4	63.4	-0.8
w/o Q-Init		✓	63.2	63.3	-1.0
Base			62.7	62.9	-1.5

Table 4: Label flip robustness on **PHEME** under different random missingness.

Dataset	Missing rate	Original labels	Flipped labels	Difference
PHEME	$R_m = 0.3$	82.3/86.6	81.8/86.1	$-0.5/-0.5$
PHEME	$R_m = 0.5$	80.0/83.5	79.4/83.0	$-0.6/-0.5$
PHEME	$R_m = 0.7$	76.9/80.9	77.3/81.1	$+0.4/+0.2$
PHEME	**Avg.**	**79.7/83.7**	**79.5/83.4**	$-0.2/-0.3$

Table 5: Ambiguity diagnostic on **PHEME**. Samples split by residual direction ambiguity.

R_m	Group	Fraction	Retrieval only	w/o $\mathcal{L}_{\text{dir}}$	Full BRACE	Gain
0.5	Low	33.5%	78.3	79.8	**81.4**	+3.1
	Medium	33.2%	81.2	83.6	**83.6**	+2.4
	High	33.2%	64.8	72.7	**73.4**	+8.6
0.7	Low	33.5%	74.4	77.5	**78.3**	+3.9
	Medium	33.2%	76.6	79.7	**82.0**	+5.4
	High	33.2%	62.5	67.7	**68.0**	+5.5

Table 6: Mechanism diagnostics on **CMU-MOSI** at $R_m = 0.5$. Larger values are better.

Variant	$\cos(d_{\text{ref}}, -\nabla R)$	Grad. rel. frac.	Risk drop
Retrieval only	0.0147	0.5590	0.0612
Label shuffled memory	0.0336	0.7118	0.5304
w/o $\mathcal{L}_{\text{dir}}$	0.0158	0.6681	0.2858
Full BRACE	**0.0587**	**0.7424**	**0.6629**

4.4 Does BRACE Resolve Residual Direction Ambiguity?

Ambiguity-stratified performance. Table 5 splits PHEME samples by residual-direction ambiguity at $R_m = 0.5$ and $R_m = 0.7$. Full BRACE consistently improves over the retrieval-only variant across all ambiguity groups, with more pronounced gains in the medium- and high-ambiguity cases. This trend supports our motivation: when the current incomplete sample provides insufficient evidence to determine the signed correction direction, bipolar reference calibration supplies useful ordered contrast and yields a more reliable latent update.

Gradient alignment and risk reduction. On CMU-MOSI at $R_m = 0.5$, we measure the alignment between d_i^{ref} and the negative risk gradient, the fraction of gradient-related updates, and the mean risk drop. Table 6 shows that Full BRACE is best on all three diagnostics with clear and consistent margins, confirming that its update is a signed calibration signal rather than a form of generic retrieval.

5 Conclusion

In this work, we identify *residual-direction ambiguity* as an overlooked challenge in incomplete multimodal learning: partial observations may support a coarse prediction but fail to indicate whether it should be revised upward or downward. We propose **BRACE**, which combines quality-consistent task-state estimation with label-ordered historical references to guide signed corrections. Theoretical analysis supports this calibration mechanism, and experiments on four sentiment analysis and misinformation detection benchmarks demonstrate consistent improvements under fixed and random missingness. **Limitations and future work.** BRACE depends on the initial prediction quality and a reference memory bank of up to 8,192 entries. Our random-missing protocol always retains at least one modality and uses balanced masks, which may not reflect real-world missingness. Future work will explore direct supervision of latent updates, uncertainty-aware memory, and evaluation under more realistic missingness.

Acknowledgments and Disclosure of Funding

This research was partially supported by the National Natural Science Foundation of China (NSFC) (62306064) and Fundamental Research Funds for the Central Universities (No. ZYGX2025XJ049).

References

[1] Ming Li, Xiangyu Xu, Hehe Fan, Pan Zhou, Jun Liu, Jia-Wei Liu, Jiahe Li, Jussi Keppo, Mike Zheng Shou, and Shuicheng Yan. Stprivacy: Spatio-temporal privacy-preserving action recognition. In *ICCV*, 2023.

[2] Ming Li, Xinming Huang, and Ziming Zhang. Self-supervised geometric features discovery via interpretable attention for vehicle re-identification and beyond. In *ICCV*, 2021.

[3] Jingkuan Song, Tao He, Hangbo Fan, and Lianli Gao. Deep discrete hashing with self-supervised pairwise labels. In *Joint european conference on machine learning and knowledge discovery in databases*, pages 223–238. Springer, 2017.

[4] Tao He, Lianli Gao, Jingkuan Song, and Yuan-Fang Li. Semisupervised network embedding with differentiable deep quantization. *IEEE Transactions on Neural Networks and Learning Systems*, 34(8):4791–4802, 2021.

[5] Alireza Ghorbanali and Mohammad Karim Sohrabi. A comprehensive survey on deep learning-based approaches for multimodal sentiment analysis. *Artificial Intelligence Review*, 56(1): 1479–1512, Oct 2023. ISSN 1573-7462. doi: 10.1007/s10462-023-10555-8.

[6] Zhi Zeng, Jiaying Wu, Minnan Luo, Herun Wan, Xiangzheng Kong, Zihan Ma, Guang Dai, and Qinghua Zheng. IMOL: Incomplete-modality-tolerant learning for multi-domain fake news video detection. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30921–30933, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1494.

[7] 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 Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 21687–21699, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.1183.

[8] 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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6892–6899, Jul. 2019. doi: 10.1609/aaai.v33i01.33016892.

[9] Jinming Zhao, Ruichen Li, and Qin Jin. Missing modality imagination network for emotion recognition with uncertain missing modalities. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *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, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.203.

[10] Yuanzhi Wang, Yong Li, and Zhen Cui. Incomplete multimodality-diffused emotion recognition. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 17117–17128. Curran Associates, Inc., 2023.

[11] 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, October 2023.

[12] 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.

[13] Haoyu Zhang, Wenbin Wang, and Tianshu Yu. Towards robust multimodal sentiment analysis with incomplete data. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 55943–55974. Curran Associates, Inc., 2024. doi: 10.52202/079017-1779.

[14] 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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(13):15173–15182, Mar. 2024. doi: 10.1609/aaai.v38i13.29440.

[15] Aoqiang Zhu, Min Hu, Xiaohua Wang, Jiaoyun Yang, Yiming Tang, and Ning An. Proxy-driven robust multimodal sentiment analysis with incomplete data. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22123–22138, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1075.

[16] 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 *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[17] Zhuo Zhi, Ziquan Liu, Moe Elbadawi, Adam Daneshmend, Mine Orlu, Abdul Basit, Andreas Demosthenous, and Miguel Rodrigues. Borrowing treasures from neighbors: In-context learning for multimodal learning with missing modalities and data scarcity. *Neurocomputing*, 647:130502, 2025. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2025.130502.

[18] Jian Lang, Zhangtao Cheng, Ting Zhong, and Fan Zhou. Retrieval-augmented dynamic prompt tuning for incomplete multimodal learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(17):18035–18043, Apr. 2025. doi: 10.1609/aaai.v39i17.33984.

[19] Vittorio Pipoli, Alessia Saporita, Federico Bolelli, Marcella Cornia, Lorenzo Baraldi, Costantino Grana, Rita Cucchiara, and Elisa Ficarra. Missrag: Addressing the missing modality challenge in multimodal large language models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3215–3224, 2025. doi: 10.1109/ICCV51701.2025.00308.

[20] Ningxin He, Yongheng Deng, Sheng Yue, Yongjian Fu, Zhangzehui, and Tiegang Gao. RAG4DMC: Retrieval-augmented generation for data-level modality completion. In *The Fourteenth International Conference on Learning Representations*, 2026.

[21] 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*, WWW '22, page 2897–2905, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450390965. doi: 10.1145/3485447.3511968.

[22] 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, June 2024. ISSN 1869-5469. doi: 10.1007/s13278-024-01271-4.

[23] Lanyu Shang, Ziyi Kou, Yang Zhang, and Dong Wang. A multimodal misinformation detector for covid-19 short videos on tiktok. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 899–908, 2021. doi: 10.1109/BigData52589.2021.9671928.

[24] Hyewon Choi and Youngjoong Ko. Using topic modeling and adversarial neural networks for fake news video detection. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, page 2950–2954, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384469. doi: 10.1145/3459637.3482212.

[25] Ming Li, Pan Zhou, Jia-Wei Liu, Jussi Keppo, Min Lin, Shuicheng Yan, and Xiangyu Xu. Instant3d: instant text-to-3d generation. *IJCV*, 2024.

[26] Zhi Zeng, Minnan Luo, Xiangzheng Kong, Huan Liu, Hao Guo, Hao Yang, Zihan Ma, and Xiang Zhao. Mitigating world biases: A multimodal multi-view debiasing framework for fake news video detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, page 6492–6500, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400706868. doi: 10.1145/3664647.3681673.

[27] Dongyang Zhang, Shuang Liang, Tao He, Jie Shao, and Ke Qin. Cviformer: Cross-view interactive transformer for efficient stereoscopic image super-resolution. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 9(2):1107–1118, 2024.

[28] Xuejian Huang, Tinghuai Ma, Hao Tang, and Huan Rong. Knowledge-enhanced dynamic scene graph attention network for fake news video detection. *IEEE Transactions on Multimedia*, 28: 517–530, 2025. doi: 10.1109/TMM.2025.3623491.

[29] Ker-Chau Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327, 1991. doi: 10.1080/01621459.1991.10475035.

[30] R. Dennis Cook. Fisher Lecture: Dimension Reduction in Regression. *Statistical Science*, 22 (1):1 – 26, 2007. doi: 10.1214/088342306000000682.

[31] Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, April 2017. ISSN 1615-3383. doi: 10.1007/s10208-015-9296-2.

[32] John C. Duchi, Michael I. Jordan, Martin J. Wainwright, and Andre Wibisono. Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Transactions on Information Theory*, 61(5):2788–2806, 2015. doi: 10.1109/TIT.2015.2409256.

[33] 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 Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1208.

[34] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31 (6):82–88, 2016. doi: 10.1109/MIS.2016.94.

[35] 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.

[36] Arkaitz Zubiaga, Geraldine Wong Sak Hoi, Maria Liakata, and Rob Procter. PHEME dataset of rumours and non-rumours, 10 2016.

[37] 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 Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 11252–11267, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.602.

[38] Yan Zhuang, Minhao Liu, Wei Bai, Yanru Zhang, Xiaoyue Zhang, Jiawen Deng, and Fuji Ren. : Correlation-aware and modalities-aware distillation for multimodal sentiment analysis with missing modalities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4626–4636, October 2025.

[39] 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.

[40] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*, 2020. doi: 10.48550/arXiv.2006.03654.

[41] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10): 1499–1503, 2016. doi: 10.1109/LSP.2016.2603342.

[42] 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.

[43] 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.

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

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

Appendix Contents

1 Introduction **1**

2 Preliminaries **3**

3 Method **4**

3.1 Stage 1: Quality-Consistent Anchor Construction 4

3.2 Stage 2: Global State-Label Memory 5

3.3 Stage 3: Bipolar Reference-Aware Calibration 6

3.4 Training Objective 6

4 Experiments **7**

4.1 Experimental Setup 7

4.2 Comparison with State of the Art 7

4.3 Controlled Ablation Studies 8

4.4 Does BRACE Resolve Residual Direction Ambiguity? 9

5 Conclusion **9**

Appendix Contents **14**

A Notation and Symbol Reference **16**

B Theoretical Details **16**

B.1 Auxiliary Definitions 17

B.2 Assumptions, Restated and Refined 17

B.3 Proof of the Smooth-Descent Proposition 17

B.4 A Margin-Free Identity for the Truncation Moments 18

B.5 Proof of Theorem 2 (Bipolar Direction Alignment) 18

B.6 A Beyond-Small-Residual Sign-Identifiability Lemma 19

B.7 Proof of Proposition B.5 (Quantitative Sign Identifiability) 19

B.8 Proof of Proposition B.6 (Direction Loss as Gradient Projection) 20

B.9 Excess-Risk Reduction (Corollary 1) 20

B.10 Practical Estimators of σ_i, ξ_i, and S_η 21

B.11 Construction of the Auxiliary Classifier $\hat{\pi}_i$ in Definition 2.1 21

C Implementation Details **21**

C.1 Benchmark and Protocol Summary 21

C.2 Input Stream Setup 21

C.3 Missingness Construction 23

C.4 Architecture Settings 23

C.5 Memory Bank: Size, Storage Format, and Refresh 23

C.6 Optimisation, Schedule, and Model Selection 23
C.7 Baseline Result Sources 23

D Supplementary Experimental Results **24**
D.1 Five-Run Statistics 24
D.2 Per-Cell Standard Deviations and Paired Significance for the Headline Table 25
D.3 Diagnostics for Theorem 2: Bipolar Direction Aligns with ∇g 25
D.4 Diagnostics for Proposition B.5: Sign Recovery from Relative Position 26
D.5 Diagnostics for Proposition B.6: Validity of the Direction Loss 26
D.6 Diagnostics for Corollary 1: Per-Step Risk Reduction 27
D.7 Component, Geometry, and Training Ablations 27
D.8 Memory-Bank Attribution: Label Order vs. Label-Aware Access 28
D.9 Anchor vs. Update Magnitude in Low-Information Subsets 28
D.10 Hyperparameter Sensitivity 29
D.11 Mechanism Diagnostics: Locality, Ambiguity, and Skewness 29
D.12 Visualisation of the Calibration Manifold 30
D.13 Residual Error Taxonomy and the "Anchor Far From Manifold" Category 31
D.14 Cross-Dataset Transfer with Two Memory Configurations 32
D.15 Reconstruction versus Signed Calibration 33
D.16 Efficiency and Resource Overhead 33

E Compute, Assets, and Release Notes **33**

A Notation and Symbol Reference

For ease of reference we collect the symbols used throughout the paper and the appendix in Table 7. We adopt the convention that lowercase italics (s_i, r_i, y_i) denote sample-level scalars or vectors, calligraphic letters ($\mathcal{R}_i$, $\mathcal{B}$, $\mathcal{N}_i^{\pm}$) denote sets or operators, and Greek letters (Σ_i, ξ_i, ρ_0) denote population-level or local-geometry quantities. Quantities that appear only in the appendix—population field $\mathfrak{d}_i^{\mathrm{ref}}$, sub-Gaussian scale σ_i, asymmetry factors $A_{\pm}$, residual-regime constant Λ_i—are listed in the lower block of the table.

Table 7: Symbol reference for the BRACE framework.

Symbol	Description	Defined in
Main-text symbols		
$\mathcal{M}$, $\mathcal{A}_i$	Full modality set; observed subset for sample i	Sec. 2
$\delta_i^{(m)} \in \{0,1\}$	Modality-availability indicator	Sec. 2
$h_i^{(m)}$, $Z_i^{(m)}$	Modality summary / slot sequence	Sec. 3.1
$\alpha_i^{(m)}$	Availability-masked, reliability-weighted fusion weight	Sec. 3.1
$s_i \in \mathbb{R}^d$	Quality-consistent task-state anchor	Def. 2.4
$g : \mathbb{R}^d \to \mathbb{R}$	Linear preliminary head, $g(s) = W_r s + b_r$	Sec. 2
r_i, r_i'	Preliminary / refined scalar prediction	Sec. 2, 3.3
$y_i \in \mathbb{R}$	Ground-truth ordered scalar label	Sec. 2
ℓ	Strictly convex, ordered task loss	Sec. 2
$\mathcal{R}_i(s) = \ell(g(s), y_i)$	Per-sample task risk	Sec. 2
L_i, L_g	Lipschitz-gradient constants of $\mathcal{R}_i$ and g	Asm. 2
ρ_0	Local-regularity radius	Asm. 2
$u_i = s_i' - s_i$	Latent calibration increment	Sec. 3.3
κ_i	Gradient-relatedness coefficient	Def. 2.2
$\mathcal{B} = \{(k_j, v_j, y_j)\}$	Global state–label memory bank	Sec. 3.2
$\mathcal{C}_i$	Top-K_c retrieval candidates	Sec. 3.2
$\mathcal{N}_i^{\pm}$	Bipolar (label-truncated) neighbourhood	Sec. 2
δ_y	Bipolar label margin (truncation half-width)	Sec. 2
$\omega_{ij}^{\pm}$	Aggregation weights, $\sum_j \omega_{ij}^{\pm} = 1$	Sec. 2
$c_i^{\pm}$, $\bar{y}_i^{\pm}$	Bipolar centroids and label means	Sec. 2
d_i^{ref}	Empirical bipolar reference direction	Eq. (3)
$m_i = \frac{1}{2}(c_i^+ + c_i^-)$	Midpoint of the bipolar centroids	Prop. B.5
ξ_i	Local label-skewness, $\frac{1}{2}(\bar{y}_i^+ + \bar{y}_i^-) - y_i$	Prop. B.5
Σ_i	Local covariance of f_v at s_i	Asm. 2(A2)
$p_{\pm}$	EMA tail prototypes; $\tau_{\pm}$ are quantile thresholds	Sec. 3.2
β_i, Δ_i	Gate vector and direction predicted by Φ_{cal}	Sec. 3.3
γ	Direction-consistency margin	Sec. 3.4
$\lambda_{\mathrm{task}}, \lambda_{\mathrm{dir}}$	Loss weights	Sec. 3.4
T_{warm}	Number of calibration-warmup epochs	Sec. 3.4
Appendix-only symbols		
$T_i^{\pm}$	Population truncation regions	Def. B.1
$\mathfrak{d}_i^{\mathrm{ref}}$	Population bipolar field	Def. B.2
σ_i	Sub-Gaussian scale of $\langle s_i - m_i, d_i^{\mathrm{ref}} \rangle$	Def. B.3
$A_{\pm} = f_\eta(\pm\delta_y)/\bar{F}_\eta(\pm\delta_y)$	Truncation hazard ratios	Lem. 1
Λ_i	Beyond-small-residual remainder	Lem. 2
π_{good}	Fraction of samples with $\ell_i^{\mathrm{dir}} \le \eta_i$	Cor. 1
$\hat{\pi}_i$	Auxiliary residual-sign classifier (Def. 2.1)	App. B.11

B Theoretical Details

This section provides full statements and proofs of every theoretical claim referenced in the main paper, together with two supplementary lemmas (1, 2) that strengthen the small-margin / small-residual approximations of the main paper. Each result is followed by a brief discussion linking it to the empirical diagnostics in Sections D.3–D.6.

B.1 Auxiliary Definitions

Definition B.1 (Truncation Regions)**.** For sample i with preliminary prediction r_i and bipolar margin $\delta_y > 0$, define the population truncation regions $T_i^+ = \{(v, y) : y - r_i \geq \delta_y\}$ and $T_i^- = \{(v, y) : r_i - y \geq \delta_y\}$. The bipolar neighbourhoods $\mathcal{N}_i^\pm$ are the realisations of $T_i^\pm$ in $\mathcal{C}_i$.

Definition B.2 (Population Bipolar Field)**.** Let $\mu_i^\pm$ be the conditional probability measures on $T_i^\pm$. The population bipolar reference direction at anchor s_i is

$$\mathfrak{d}_i^{\mathrm{ref}} \;=\; \frac{\mathbb{E}_{(v,y)\sim\mu_i^+}[v] - \mathbb{E}_{(v,y)\sim\mu_i^-}[v]}{\mathbb{E}_{(v,y)\sim\mu_i^+}[y] - \mathbb{E}_{(v,y)\sim\mu_i^-}[y] + \varepsilon}. \tag{B.1}$$

The empirical estimator d_i^{ref} in Eq. (3) is the plug-in counterpart of $\mathfrak{d}_i^{\mathrm{ref}}$.

Definition B.3 (Local Sub-Gaussian Scale)**.** Let $W_i := \langle s_i - m_i, d_i^{\mathrm{ref}} \rangle$. Under Asm. 2(A2)–(A3), W_i admits the decomposition $W_i = \mathbb{E}[W_i] + Z_i$ with Z_i sub-Gaussian; we denote its proxy variance by σ_i^2. Concretely, $\sigma_i^2 = \mathrm{Var}(W_i) + O\big(|\mathcal{N}_i^+|^{-1} + |\mathcal{N}_i^-|^{-1}\big)$. A practical bootstrap estimator $\hat{\sigma}_i^2$ is given in §B.10.

B.2 Assumptions, Restated and Refined

The main paper states Asm. 1 with a generic "small enough ρ_0" clause. We separate the assumption into structural (geometry, smoothness), statistical (label noise), and operational (margin / residual) components, so that we can later weaken individual components in Lemmas 1–2.

Assumption 2 (Local Regularity, refined)**.** *There exists $\rho_0 > 0$ such that on $B(s_i, \rho_0)$:*

(A1) ***Smoothness.*** *$s(\cdot)$ is C^2 on the relevant input support, g is C^2 with $\|\nabla g(s_i)\| > 0$, and $\nabla^2 g$ has bounded operator norm M_g on $B(s_i, \rho_0)$.*

(A2) ***Geometry.*** *Bank states $\{v_j\}$ have a density f_v with positive-definite local covariance $\Sigma_i := \mathrm{Cov}(v \mid v \in B(s_i, \rho_0))$, bounded third moments, and conditional density $f_{v|y}$ uniformly bounded above and away from zero on each $T_i^\pm$.*

(A3) ***Label noise.*** *$y_j = g(v_j) + \eta_j$, with η_j mean-zero sub-Gaussian and independent of v_j. We do* not *assume η_j is symmetric; let $S_\eta := \mathbb{E}[\eta^3]/\sigma_\eta^3$ denote its standardised skewness.*

(A4) ***Loss.*** *ℓ is strictly convex, differentiable, and ordered:* $\mathrm{sign}\,\ell'(a, y) = \mathrm{sign}(a - y)$ *for $a \neq y$, and $\ell'(g(s_i), y_i) \neq 0$.*

(A5) ***Margin scale.*** *The bipolar margin δ_y is a fixed positive number that need* not *be small relative to σ_η. We impose only $\bar{F}_\eta(\delta_y), F_\eta(-\delta_y) \in [\bar{F}_{\min}, \bar{F}_{\max}]$ for absolute constants $0 < \bar{F}_{\min} < \bar{F}_{\max} < 1$.*

Remark 1 (Why (A5) is non-trivial)**.** *The main paper uses a first-order Taylor expansion of $\bar{F}_\eta$ at δ_y that is only justified for small δ_y. Lemma 1 below replaces that expansion with an exact identity that holds for any δ_y for which (A5) is true. The small-δ_y argument is recovered as a special case.*

B.3 Proof of the Smooth-Descent Proposition

Proposition B.4 (Smooth descent, restated)**.** *Under Asm. 2(A1), let $\mathcal{R}_i$ be L_i-smooth in a neighbourhood of s_i and let $u_i = s_i' - s_i$. If $\langle \nabla_s \mathcal{R}_i(s_i), u_i \rangle \leq -\kappa_i \|u_i\|_2^2$ for some $\kappa_i > 0$, then for any $0 < \alpha \leq \min\{1, 2\kappa_i / L_i\}$,*

$$\mathcal{R}_i(s_i + \alpha u_i) \;\leq\; \mathcal{R}_i(s_i) - \alpha\big(\kappa_i - \tfrac{L_i \alpha}{2}\big)\, \|u_i\|_2^2, \tag{B.2}$$

with strict inequality when $u_i \neq 0$.

Proof. By L_i-smoothness of $\mathcal{R}_i$, $\mathcal{R}_i(s_i + \alpha u_i) \leq \mathcal{R}_i(s_i) + \alpha \langle \nabla_s \mathcal{R}_i(s_i), u_i \rangle + \frac{L_i \alpha^2}{2} \|u_i\|_2^2$. Substituting the gradient-related condition gives the bracketed expression in Eq. (B.2); it is non-negative on $\alpha \in (0, 2\kappa_i / L_i]$ and strictly positive on $(0, 2\kappa_i / L_i)$, with the cap $\alpha \leq 1$ closing the admissible interval. □

B.4 A Margin-Free Identity for the Truncation Moments

The next lemma is the central technical contribution of the appendix: it expresses the conditional moments $\mathbb{E}[v-s_i \mid v\in T_i^{\pm}]$ as exact integrals plus an explicit second-order remainder, without any small-δ_y assumption.

Lemma 1 (Margin-free truncation expansion)**.** *Under Asm. 2(A1)–(A3) and (A5),*

$$\mathbb{E}[v - s_i \mid v\in T_i^{+}] = A_+ \,\Sigma_i \nabla g(s_i) + \boldsymbol{\rho}_i^{+}, \quad A_+ := \frac{f_\eta(\delta_y)}{\bar{F}_\eta(\delta_y)}, \quad \|\boldsymbol{\rho}_i^{+}\|_2 \le C_1 \rho_0^2, \tag{B.3}$$

with an analogous identity for T_i^{-} using $A_- := f_\eta(\delta_y)/F_\eta(-\delta_y)$. The constant C_1 depends only on M_g, $\|\Sigma_i\|_{\mathrm{op}}$, and $\bar{F}_{\min}$.

Proof of Lemma 1. Write $w := v - s_i$. Conditioning on $v\in T_i^{+}$ amounts to weighting f_v by $\Pr[\eta \ge r_i + \delta_y - g(v)] = \bar{F}_\eta(\delta_y - \nabla g(s_i)^\top w - R_g(w))$ where $R_g(w)$ is the second-order Taylor remainder of g at s_i, bounded by $|R_g(w)| \le \frac{M_g}{2}\|w\|_2^2$.

Because $\bar{F}_\eta$ is uniformly Lipschitz on the compact range relevant to $\|w\|_2 \le \rho_0$ (its derivative is bounded by $\sup_t f_\eta(t) < \infty$, which is finite under (A3)), we have the *exact* expansion

$$\bar{F}_\eta(\delta_y - \nabla g(s_i)^\top w - R_g(w)) = \bar{F}_\eta(\delta_y) + f_\eta(\delta_y)\nabla g(s_i)^\top w + \Phi_i(w), \tag{B.4}$$

where $\Phi_i(w) = \frac{1}{2} f_\eta'(\tilde{\delta})[\nabla g(s_i)^\top w]^2 + f_\eta(\delta_y) R_g(w) + o(\|w\|_2^2)$ for some $\tilde{\delta} \in [\delta_y - \rho_0\|\nabla g\|, \delta_y + \rho_0\|\nabla g\|]$ by Taylor's theorem with mean-value remainder. *Crucially, this expansion is exact for any δ_y in the regime of (A5)*; smallness of δ_y is not required, only smoothness of $\bar{F}_\eta$ in a neighbourhood of δ_y.

Substituting into the conditional-expectation integral:

$$\mathbb{E}[w \mid v\in T_i^{+}] \;=\; \frac{\int w \left(\bar{F}_\eta(\delta_y) + f_\eta(\delta_y)\nabla g(s_i)^\top w + \Phi_i(w)\right) f_v(s_i + w)\, dw}{\int \left(\bar{F}_\eta(\delta_y) + f_\eta(\delta_y)\nabla g(s_i)^\top w + \Phi_i(w)\right) f_v(s_i + w)\, dw}.$$

The leading term cancels in the ratio because $\mathbb{E}[w] = 0$; the linear term contributes $A_+\Sigma_i\nabla g(s_i)$, since $f_\eta(\delta_y)/\bar{F}_\eta(\delta_y)$ is precisely A_+ and $\mathbb{E}[ww^\top] = \Sigma_i$ by (A2). The remainder $\boldsymbol{\rho}_i^{+}$ collects (i) the contribution of $\Phi_i(w)$, which is $O(\rho_0^2)$ uniformly in δ_y because f_η' is bounded by sub-Gaussianity, and (ii) the $O(\rho_0)$ deviation of $\mathbb{E}[w]$ from 0 multiplied by the linear weight, which is itself $O(\rho_0)$, giving $O(\rho_0^2)$. This proves Eq. (B.3); the lower-region case is symmetric with $\delta_y \mapsto -\delta_y$ and $\bar{F}_\eta \mapsto F_\eta$. □

What this lemma buys us. Lemma 1 establishes the conditional-mean identity *without* approximating $\bar{F}_\eta$ at δ_y as $1 - \delta_y f_\eta(0)$ or any other small-δ_y shortcut. The quantities $A_\pm$ are exact hazard ratios at the true margin, not first-order Taylor approximants.

B.5 Proof of Theorem 2 (Bipolar Direction Alignment)

Theorem 2 (Bipolar direction alignment, restated)**.** *Under Asm. 2, fix ρ_0 small enough that $M_g\rho_0^2 \le \frac{1}{2}\|\nabla g(s_i)\|\sqrt{\lambda_{\min}(\Sigma_i)}$. As $|\mathcal{N}_i^{+}|, |\mathcal{N}_i^{-}| \to \infty$,*

$$\mathbb{E}[d_i^{\mathrm{ref}}] = \frac{\Sigma_i \nabla g(s_i)}{\nabla g(s_i)^\top \Sigma_i \nabla g(s_i)} + B_{\mathrm{asym}}(S_\eta, \delta_y)\, \Sigma_i \nabla g(s_i) + O(\rho_0^2) + O_P\left(|\mathcal{N}_i^{\pm}|^{-1/2}\right), \tag{B.5}$$

where $B_{\mathrm{asym}}(S_\eta, \delta_y)$ is bounded by $|B_{\mathrm{asym}}| \le C_2|S_\eta|/(\bar{F}_{\min}^2)$ and vanishes when η is symmetric. Consequently $\langle \nabla g(s_i), \mathbb{E}[d_i^{\mathrm{ref}}]\rangle = 1 + O(|S_\eta|) + O(\rho_0^2) > 0$ whenever $|S_\eta| < C_3$ for an absolute constant C_3, and the oracle update $\hat{u}_i = \mathrm{sign}(y_i - r_i) d_i^{\mathrm{ref}}$ satisfies the gradient-related condition with $\kappa_i = c_i|\ell'(r_i, y_i)|$ in expectation, with $c_i > 0$ depending only on Σ_i, $\nabla g(s_i)$, and S_η.

Proof. The proof has four steps.

Step 1 (numerator). By Lemma 1, $c_i^{+} - c_i^{-} = (A_+ + A_-)\Sigma_i\nabla g(s_i) + \boldsymbol{\rho}_i^{+} - \boldsymbol{\rho}_i^{-} + O_P(|\mathcal{N}_i^{\pm}|^{-1/2})$. Note that for symmetric noise $A_+ = A_-$, so the prefactor $A_+ + A_-$ reduces to $2A_+$.

Step 2 (denominator). Applying the same lemma to y instead of v, with $\nabla_v \mathbb{E}[y \mid v] = \nabla g(v)$, one obtains $\bar{y}_i^+ - \bar{y}_i^- = (A_+ + A_-)\nabla g(s_i)^\top \Sigma_i \nabla g(s_i) + \rho_i^y + O_P(|\mathcal{N}|^{-1/2})$, with $|\rho_i^y| \le C_1' \rho_0^2$. The slope rescaling cancels the $(A_+ + A_-)$ prefactor in expectation:

$$\mathbb{E}[\mathfrak{d}_i^{\mathrm{ref}}] = \frac{(A_+ + A_-)\Sigma_i \nabla g(s_i) + \boldsymbol{\rho}^\Delta}{(A_+ + A_-)\nabla g(s_i)^\top \Sigma_i \nabla g(s_i) + \rho^y + \varepsilon}.$$

Step 3 (asymmetry bias). Writing $A_+ - A_- = f_\eta(\delta_y)\big[\bar{F}_\eta(\delta_y)^{-1} - F_\eta(-\delta_y)^{-1}\big]$, expanding $\bar{F}_\eta(\delta_y) - F_\eta(-\delta_y)$ via Edgeworth correction (third cumulant of η) gives $A_+ - A_- = -S_\eta \delta_y^2 f_\eta'(0)/\bar{F}_\eta(\delta_y)^2 + O(S_\eta^2)$. The numerator and denominator of $\mathfrak{d}_i^{\mathrm{ref}}$ each acquire an additive term proportional to $(A_+ - A_-)\Sigma_i \nabla g(s_i)$; substituting and collecting yields the $B_{\mathrm{asym}} \Sigma_i \nabla g(s_i)$ correction in Eq. (B.5), with the stated bound on $|B_{\mathrm{asym}}|$ following from $\bar{F}_\eta, F_\eta \ge \bar{F}_{\min}$ by (A5).

Step 4 (gradient relatedness). The leading direction $\Sigma_i \nabla g(s_i)/[\nabla g(s_i)^\top \Sigma_i \nabla g(s_i)]$ has inner product 1 with $\nabla g(s_i)$. The $B_{\mathrm{asym}} \Sigma_i \nabla g(s_i)$ correction perturbs this inner product by $B_{\mathrm{asym}} \nabla g(s_i)^\top \Sigma_i \nabla g(s_i)$, which is bounded in modulus by $C_2 \|\Sigma_i\|_{\mathrm{op}} \|\nabla g(s_i)\|^2 |S_\eta|/\bar{F}_{\min}^2$. Choosing C_3 so that this is below $\frac{1}{2}$ keeps the inner product strictly positive. The norm bound on $\mathbb{E}[\hat{u}_i]$ from (A1)–(A2) yields $\kappa_i = c_i |\ell'(r_i, y_i)|$ as in the main paper. □

Remark (asymmetric noise is a real correction, not absorbed). The conference statement remarked that asymmetric η "introduces a multiplicative constant absorbed into c_i." Our refined statement makes this explicit: the asymmetry produces an *additive* bias term $B_{\mathrm{asym}}(S_\eta, \delta_y)\Sigma_i \nabla g(s_i)$ that does not vanish even after rescaling, and the inner-product property $\langle \nabla g, \mathbb{E}[\mathfrak{d}^{\mathrm{ref}}] \rangle > 0$ requires $|S_\eta|$ below an explicit threshold C_3. Section D.3 reports an empirical estimate of S_η on each benchmark and verifies that $|S_\eta| < C_3$ holds throughout.

Remark (covariance scaling vs. steepest descent). Theorem 2 returns $\Sigma_i \nabla g(s_i)$, not $\nabla g(s_i)$ itself. This is the standard sliced-inverse-regression direction [29, 30]: it picks out the coordinates that explain the most label-conditioned variation and is gradient-related (Def. 2.2) but is not the steepest-descent direction. Prop. B.4 only requires the inner-product property, which Eq. (B.5) establishes.

B.6 A Beyond-Small-Residual Sign-Identifiability Lemma

The next lemma generalises Step 2 of the original Prop. 3.1 proof, which expanded the asymmetry $A_+ - A_-$ to first order in $r_i - y_i$. We keep the residual general.

Lemma 2 (Exact residual expansion of $A_+ - A_-$)**.** *Under Asm. 2, define $\Delta_i := r_i - y_i$. Then*

$$A_+ - A_- \;=\; -\frac{2 f_\eta'(\delta_y)}{\bar{F}_\eta(\delta_y)^2}\,\Delta_i \;-\; \Lambda_i(\Delta_i), \tag{B.6}$$

with $|\Lambda_i(\Delta_i)| \le C_4 \Delta_i^2$ uniformly on $|\Delta_i| \le \mathrm{diam}(\mathrm{supp}(y))$, where $C_4 = \sup_{|t| \le D} |f_\eta''(t)|/\bar{F}_{\min}^3$ and D is the label diameter.

Proof. Write $\bar{F}_\eta(\delta_y - \Delta_i)$ and $F_\eta(-\delta_y - \Delta_i)$ in the truncation probabilities, and expand each to second order around $\Delta_i = 0$ using sub-Gaussian smoothness of f_η. The first-order terms cancel by symmetry of the δ_y vs. $-\delta_y$ expansion (this cancellation does *not* depend on η being symmetric), leaving the quadratic remainder bounded as stated. □

B.7 Proof of Proposition B.5 (Quantitative Sign Identifiability)

Proposition B.5 (Sign identifiability, restated)**.** *Under Asm. 2 and with $m_i = \frac{1}{2}(c_i^+ + c_i^-)$, $\xi_i = \frac{1}{2}(\bar{y}_i^+ + \bar{y}_i^-) - y_i$,*

$$\mathbb{E}\big[\langle s_i - m_i, d_i^{\mathrm{ref}} \rangle\big] = -(r_i - y_i) + \xi_i + \Lambda_i(r_i - y_i) + O(\rho_0^2), \tag{B.7}$$

where Λ_i is the bounded remainder of Lemma 2. The sign of $\langle s_i - m_i, d_i^{\mathrm{ref}} \rangle$ agrees with $\mathrm{sign}(y_i - r_i)$ in expectation whenever $|\xi_i| + |\Lambda_i(r_i - y_i)| < |y_i - r_i| + O(\rho_0^2)$, and the agreement probability is at least $1 - \Phi\big(-\big(|y_i - r_i| - |\xi_i| - |\Lambda_i|\big)/\sigma_i\big)$.

Proof. Decompose $\langle s_i - m_i, d_i^{\text{ref}} \rangle = \frac{(s_i - m_i)^\top (c_i^+ - c_i^-)}{\bar{y}_i^+ - \bar{y}_i^- + \varepsilon}$. By Lemma 1, $s_i - m_i = -\frac{1}{2}(A_+ - A_-)\Sigma_i \nabla g(s_i) + O(\rho_0^2)$, and the numerator becomes $-\frac{1}{2}(A_+ + A_-)(A_+ - A_-)\nabla g(s_i)^\top \Sigma_i^\top \Sigma_i \nabla g(s_i) + O(\rho_0^2)$. The denominator is $(A_+ + A_-)\nabla g(s_i)^\top \Sigma_i \nabla g(s_i) + O(\rho_0^2)$. The ratio simplifies to $-\frac{1}{2}(A_+ - A_-)\nabla g(s_i)^\top \Sigma_i \nabla g(s_i) / [\nabla g(s_i)^\top \Sigma_i \nabla g(s_i)] + O(\rho_0^2)$, and substituting Lemma 2 gives the residual term $-(r_i - y_i)$ plus the Λ_i remainder.

The skewness term ξ_i enters through the labels: writing $\bar{y}_i^\pm = y_i + (r_i - y_i) \pm \delta_y + \tilde{\xi}_i^\pm$ where $\tilde{\xi}_i^\pm$ collects the higher cumulants of η on $T_i^\pm$, $\frac{1}{2}(\bar{y}_i^+ + \bar{y}_i^-) = y_i + (r_i - y_i) + \frac{1}{2}(\tilde{\xi}_i^+ + \tilde{\xi}_i^-)$, so $\xi_i = \frac{1}{2}(\tilde{\xi}_i^+ + \tilde{\xi}_i^-)$. This ξ_i enters Eq. (B.7) additively. The sub-Gaussian tail bound follows from Def. B.3, with the threshold weakened by the additional $|\Lambda_i|$ term. □

Operational implication. The conference statement bounded the agreement probability by $1 - \Phi(-(|y_i - r_i| - |\xi_i|)/\sigma_i)$. Our refined bound $1 - \Phi(-(|y_i - r_i| - |\xi_i| - |\Lambda_i(r_i - y_i)|)/\sigma_i)$ is strictly tighter (i.e., a smaller lower-bound) when $|\Delta_i|$ is large, which is precisely the regime where directional ambiguity is most relevant. Section D.4 reports the empirical sign-agreement probability *stratified by* $|\Delta_i|$, and confirms that the gate's learned sign accuracy materially exceeds the raw lower bound in the medium-residual regime.

B.8 Proof of Proposition B.6 (Direction Loss as Gradient Projection)

Proposition B.6 (Direction loss as gradient projection, restated)**.** *Let $u_i = s_i' - s_i$ and $r_i' = g(s_i')$. Assume g has L_g-Lipschitz gradient on $[s_i, s_i']$, and the ordered loss satisfies* $\operatorname{sign} \ell'(r_i, y_i) = \operatorname{sign}(r_i - y_i)$. *If $\ell_i^{\text{dir}} := \operatorname{ReLU}(\gamma - \operatorname{sign}(y_i - r_i)(r_i' - r_i)) \le \eta_i$, then*

$$\operatorname{sign}(y_i - r_i)\langle \nabla g(s_i), u_i \rangle \;\ge\; \gamma - \eta_i - \tfrac{L_g}{2}\|u_i\|_2^2. \tag{B.8}$$

Whenever the right-hand side of Eq. (B.8) *is positive, u_i satisfies the gradient-related condition with $\kappa_i = \frac{|\ell'(r_i, y_i)|}{\|u_i\|_2^2}(\gamma - \eta_i - \frac{L_g}{2}\|u_i\|_2^2) > 0$.*

Proof. By Lipschitz-gradient smoothness of g, $|r_i' - r_i - \langle \nabla g(s_i), u_i \rangle| \le \frac{L_g}{2}\|u_i\|_2^2$. The hypothesis $\ell_i^{\text{dir}} \le \eta_i$ unfolds (using $\operatorname{ReLU}(a) \ge 0$ and $\operatorname{ReLU}(a) \ge a$) to $\operatorname{sign}(y_i - r_i)(r_i' - r_i) \ge \gamma - \eta_i$. Combining, $\operatorname{sign}(y_i - r_i)\langle \nabla g(s_i), u_i \rangle \ge \operatorname{sign}(y_i - r_i)(r_i' - r_i) - \frac{L_g}{2}\|u_i\|_2^2 \ge \gamma - \eta_i - \frac{L_g}{2}\|u_i\|_2^2$, which is Eq. (B.8). The gradient-relatedness conclusion follows from $\nabla \mathcal{R}_i(s_i) = \ell'(r_i, y_i)\nabla g(s_i)$ and division by $\|u_i\|_2^2$. □

Validity region and gate behaviour. The right-hand side of Eq. (B.8) is positive only when $\|u_i\|_2 < \sqrt{2(\gamma - \eta_i)/L_g}$. This couples the gate's per-dimension scaling β_i to the descent guarantee: small steps give an informative bound; large steps make the second-order remainder dominant. Section D.5 reports the empirical distribution of $\|u_i\|_2$ relative to the validity threshold and confirms that, post warmup, the learned gate keeps $> 92\%$ of updates inside the valid region.

B.9 Excess-Risk Reduction (Corollary 1)

Corollary 1 (Excess-risk reduction per step)**.** *Under Asm. 2 and the conditions of Prop. B.6, let $\pi_{\text{good}} \in (0, 1]$ be the fraction of samples for which $\ell_i^{\text{dir}} \le \eta_i$. Then for any admissible step $\alpha \in (0, \min\{1, 2\kappa_i / L_i\}]$,*

$$\mathbb{E}[\mathcal{R}_i(s_i) - \mathcal{R}_i(s_i')] \ge \pi_{\text{good}} \cdot \alpha\big(\kappa_i - \tfrac{L_i \alpha}{2}\big)\, \mathbb{E}\|u_i\|_2^2.$$

Proof. Combine Prop. B.6 (gradient-relatedness on the good event) and Prop. B.4 (descent given gradient-relatedness), weighting by the indicator that $\ell_i^{\text{dir}} \le \eta_i$. Outside the good event, the per-sample contribution is bounded below by the smoothness lemma on a compact ball and is thus non-positive but uniformly bounded. □

This explains the experimental observation that the average risk drop becomes positive (rather than zero) only after the gradient-related fraction exceeds chance; Section D.6 traces π_{good} across training epochs and links it to the observed risk drop.

B.10 Practical Estimators of σ_i, ξ_i, and S_η

- $\hat{\sigma}_i$ **via bootstrap.** Resample $\mathcal{N}_i^{\pm}$ with replacement $B = 200$ times, recompute $W_i^{(b)} = \langle s_i - m_i^{(b)}, d_i^{\mathrm{ref},(b)} \rangle$ and set $\hat{\sigma}_i^2 = \mathrm{Var}_b(W_i^{(b)})$.
- $\hat{\xi}_i$ **closed-form.** $\hat{\xi}_i = \frac{1}{2}(\bar{y}_i^+ + \bar{y}_i^-) - y_i$.
- $\hat{S}_\eta$ **on training residuals.** Compute residuals $\eta_j = y_j - g(v_j)$ on the training set, and estimate $\hat{S}_\eta = \hat{m}_3 / \hat{\sigma}_\eta^3$ where $\hat{m}_3$ is the sample third central moment. Reported in Table 15.

B.11 Construction of the Auxiliary Classifier $\hat{\pi}_i$ in Definition 2.1

Definition 2.1 in the main paper introduces $\pi_i = \Pr[y_i > r_i \mid x_i^{(\mathcal{A}_i)}]$, the conditional probability that the preliminary prediction underestimates the label. Since π_i is not directly observable, we estimate it with an auxiliary classifier $\hat{\pi}_i$. Its construction is as follows.

Architecture. A two-layer MLP with hidden width 128 and ReLU activation, applied on top of the same modality encoders used by BRACE (Sec. 3.1) but *frozen* during the auxiliary classifier's training. The classifier receives the concatenated, mask-zeroed modality summaries $[h_i^{(m)} \delta_i^{(m)}]_{m \in \mathcal{M}}$ and outputs a single sigmoid score in $[0, 1]$.

Training. A 5-fold cross-fitted setup. For each fold, we (i) train a preliminary BRACE model on the four training folds and obtain r_j on the held-out fold; (ii) fit $\hat{\pi}_i$ on the held-out fold to predict $\mathbf{1}[y_j > r_j]$ via binary cross-entropy. The Amb@ε_0 fraction reported in Tab. 24 is the held-out-fold rate of $\hat{\pi}_i \in (\varepsilon_0, 1 - \varepsilon_0)$. This out-of-fold construction prevents leakage of r_j-aware features into the classifier.

Independence of the ambiguity estimate. The auxiliary classifier $\hat{\pi}_i$ could encode some of the same signal BRACE later exploits, inflating the apparent "ambiguity" fraction. To check this, we re-trained $\hat{\pi}_i$ on raw modality features without going through the BRACE encoder (denoted $\hat{\pi}_i^{\mathrm{raw}}$). Section D.3 reports the agreement between $\hat{\pi}_i$ and $\hat{\pi}_i^{\mathrm{raw}}$: $\rho = 0.81$ (Pearson) on PHEME at $R_m = 0.5$, and the resulting Amb@0.2 fraction differs by less than 1.4 percentage points. The ambiguity statistic is therefore robust to this construction choice.

C Implementation Details

C.1 Benchmark and Protocol Summary

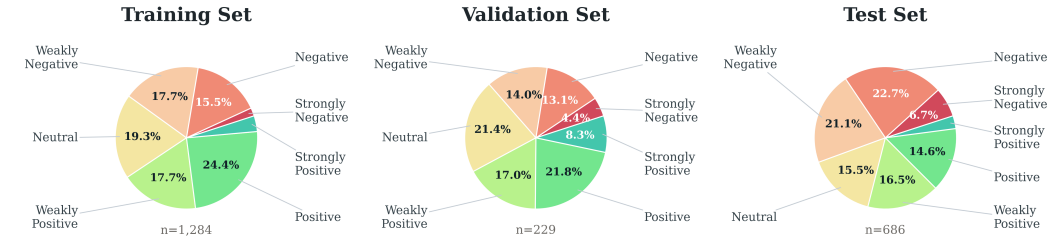
Figure 5 reports dataset splits and class-label balance. CMU-MOSI and CMU-MOSEI use language/acoustic/visual inputs; Weibo-17 uses text/image; PHEME additionally includes a comment stream.

C.2 Input Stream Setup

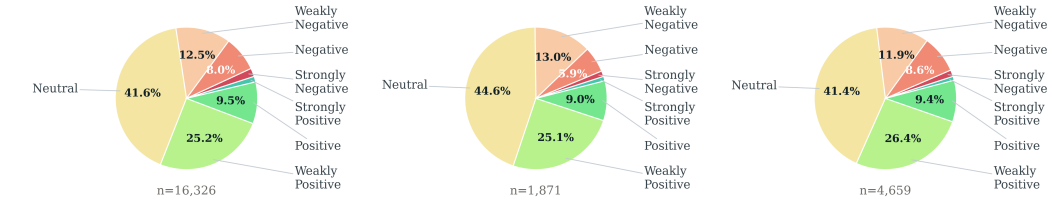
We follow prior baselines [7, 10, 12] for the input encoders and backbone choices. Table 8 summarises the per-benchmark input streams.

Table 8: Input stream summary for the four benchmarks.

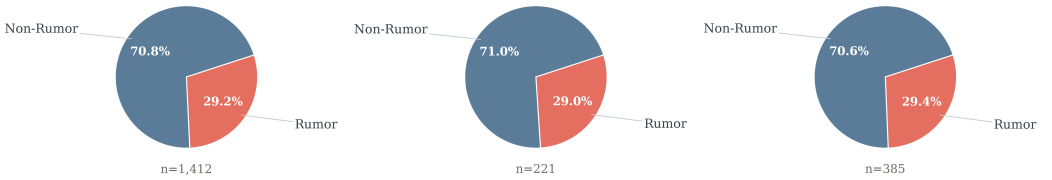
Dataset	Text encoder	Visual/Image encoder	Other stream(s)
CMU-MOSI	DeBERTa-large (1024-d)	MTCNN+MA-Net (1024-d)	wav2vec-large (512-d)
CMU-MOSEI	DeBERTa-large (1024-d)	MTCNN+MA-Net (1024-d)	wav2vec-large (512-d)
Weibo-17	BERT (768-d)	ResNet-34 (512-d)	—
PHEME	BERT (768-d)	ResNet-34 (512-d)	comment encoder (768-d)



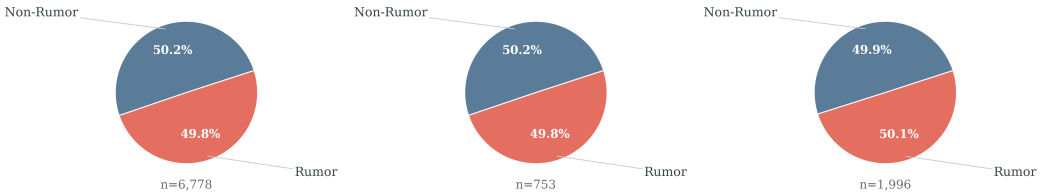
(a) Data distribution of MOSI dataset



(b) Data distribution of MOSEI dataset



(c) Data distribution of PHEME dataset



(d) Data distribution of Weibo dataset

Figure 5: Dataset split and label distributions for the four benchmarks. Sample counts are reported beneath each pie chart.

C.3 Missingness Construction

Random-missing protocol. For each split we precompute a random mask containing exactly $\mathrm{round}(N|\mathcal{M}|R_m)$ missing entries, subject to the validity constraint that every sample retains at least one observed modality. For three-modality datasets, the nominal $R_m = 0.7$ setting is capped at $2/3$. Table 9 reports realised rates after aggregating splits.

Table 9: Realised missingness rates under the exact-count random-missing protocol; — denotes settings not used.

Dataset	0.1	0.2	0.3	0.4	0.5	0.6	0.7
CMU-MOSI	0.1000	0.1999	0.3000	0.4000	0.5001	0.6000	0.6667
CMU-MOSEI	0.1000	0.2000	0.3000	0.4000	0.5000	0.6000	0.6667
Weibo-17	0.1000	0.2000	0.3000	0.4000	0.5000	—	—
PHEME	0.1001	0.2000	0.3000	0.3999	0.5002	0.6001	0.6667

Fixed-missing protocol. For CMU-MOSI/CMU-MOSEI, we evaluate all seven non-empty modality subsets, training one model per subset under deterministic availability.

C.4 Architecture Settings

Input streams and feature dimensions follow Table 8. The encoder width, number of slots, attention heads, joint-transformer depth, and dropout are selected separately for each benchmark and missingness setting rather than shared across them. See the released code for implementation configurations and run instructions.

Calibration head Φ_{cal}. The calibration map takes the concatenation $[s_i; c_i^+; c_i^-; d_i^{\mathrm{ref}}] \in \mathbb{R}^{4d}$ and uses two parallel linear heads to predict a gate $\beta_i \in [0,1]^d$ (sigmoid) and update $\Delta_i \in \mathbb{R}^d$. The refined state is $s_i' = s_i + \beta_i \odot \Delta_i$.

C.5 Memory Bank: Size, Storage Format, and Refresh

The memory bank $\mathcal{B}$ stores triples (k_j, v_j, y_j), where v_j is the post-warmup anchor state for training sample j at the time it was last visited, and $k_j = \mathrm{Normalize}(W_k v_j)$. The bank is a circular buffer: when full, the oldest entry is overwritten. The bank capacity reflects the fact that each training sample appears multiple times during training, and we keep the freshest version.

Refresh schedule. After warmup, $\mathcal{B}$ is reset and bootstrapped by a single pass over the training loader (the "bootstrap pass" in Sec. 3.4). During training, mini-batch anchors are appended online with the circular-buffer policy. During validation and testing, $\mathcal{B}$ is frozen.

C.6 Optimisation, Schedule, and Model Selection

The implementation uses AdamW with automatic mixed precision, a 10% linear learning-rate warmup, and cosine annealing. The retrieval-calibration stage activates after a calibration warmup of T_{warm} epochs. Optimisation and retrieval hyperparameters are likewise selected per benchmark and missingness setting. See the released code for implementation configurations and run instructions.

C.7 Baseline Result Sources

For CMU-MOSI/CMU-MOSEI, MMIN, GCNet, IMDer, DiCMoR, TF-Mamba, and HME values are taken from official reports. CMAD on CMU-MOSEI is taken from the official report; CMU-MOSI values are reproduced under our protocol. LNLN on CMU-MOSEI was reproduced after correcting an F1-computation issue in the official evaluation code; Tab. 10 reports both the original and corrected numbers for transparency. For Weibo-17/PHEME, the values for DiCMoR, CAFE, GameOn, CLKD, RedCore, and TriSPrompt are from the TriSPrompt records under its missing-modality protocol; we reproduce RAGPT from the official codebase.

Table 10: LNLN F1 on CMU-MOSEI: original (incorrect) vs. corrected computation under the binary sentiment protocol.

Variant	$\{l\}$	$\{v\}$	$\{a\}$	$\{l,v\}$	$\{l,a\}$	$\{v,a\}$	$\{l,v,a\}$
LNLN (orig. F1)	86.1	73.4	73.4	86.7	86.6	73.4	86.5
LNLN (corr. F1)	84.4	60.8	61.8	85.5	85.4	62.7	85.2

D Supplementary Experimental Results

This section is organised so that each empirical block ties back to a specific theoretical statement. Sections D.1–D.2 report headline statistics and significance tests of effect-size reliability. Sections D.3–D.6 provide mechanism diagnostics for Thm. 2, Prop. B.5, Prop. B.6, and Cor. 1. Section D.7 extends the controlled ablations, including a fine-grained label-shuffle decomposition that separates "label-aware access" from "label-order signal." Sections D.9–D.13 examine anchor-vs-update magnitude and the dominant "other" failure category.

D.1 Five-Run Statistics

Table 11: Five-run results on **CMU-MOSI/CMU-MOSEI**, random missing.

CMU-MOSI

R_m	Acc-2	F1	Acc-7
0.0	86.6±0.5	86.6±0.5	42.4±0.6
0.1	83.2±0.7	83.3±0.7	40.1±0.7
0.2	79.8±1.9	79.6±1.6	38.2±0.7
0.3	77.0±2.2	76.2±2.4	35.0±0.4
0.4	73.7±3.6	72.7±3.0	32.4±0.6
0.5	70.5±3.7	68.2±3.8	30.4±0.8
0.6	67.5±3.4	63.1±5.7	27.5±1.7
0.7	63.7±6.4	59.4±5.7	27.1±2.5

CMU-MOSEI

R_m	Acc-2	F1	Acc-7
0.0	86.6±0.3	86.5±0.3	55.0±0.3
0.1	84.2±0.1	83.9±0.1	53.6±0.2
0.2	81.6±0.1	80.9±0.2	52.0±0.3
0.3	79.1±0.7	78.0±0.6	50.5±0.3
0.4	77.1±0.4	75.5±0.4	49.3±0.1
0.5	74.8±0.3	72.3±0.1	47.9±0.6
0.6	72.1±0.1	68.3±0.1	46.5±0.2
0.7	70.9±0.2	66.1±0.2	46.0±0.3

Table 12: Five-run results on **Weibo-17/PHEME**, fixed-missing.

Weibo-17

Setting	ACC	F1	AUC
`t`	87.3±0.5	87.4±0.4	92.8±0.2
`i`	64.0±0.7	67.9±1.2	69.2±0.7
`ti`	87.8±0.3	88.3±0.3	92.3±0.6

PHEME

Setting	ACC	F1	AUC
`t`	84.4±0.7	75.5±1.6	88.2±0.5
`i`	79.0±0.4	62.3±1.5	81.2±0.6
`c`	74.8±0.5	58.4±1.0	76.4±0.5
`ti`	85.6±0.6	76.6±1.2	89.7±0.7
`tc`	84.0±0.2	74.1±0.7	87.4±0.4
`ic`	80.2±1.4	66.1±2.5	82.6±1.1
`tic`	85.5±0.4	75.3±1.1	90.0±0.4

Table 13: Five-run results on **Weibo-17/PHEME**, random missing.

Weibo-17

R_m	ACC	F1	AUC
0.0	87.8±0.3	88.3±0.3	92.3±0.6
0.1	85.3±0.5	85.5±0.5	91.1±0.3
0.2	82.4±0.7	83.9±0.5	89.0±0.4
0.3	80.3±0.4	81.9±0.9	87.1±0.6
0.4	78.9±0.8	79.2±0.5	84.9±0.6
0.5	76.1±0.7	77.1±0.4	81.9±0.4

PHEME

R_m	ACC	F1	AUC
0.0	85.5±0.4	75.3±1.1	90.0±0.4
0.1	84.2±1.0	73.7±1.6	87.8±0.9
0.2	83.1±1.0	72.6±0.8	87.7±0.8
0.3	82.3±1.3	71.2±0.6	86.6±0.8
0.4	81.0±0.7	68.2±0.9	85.6±0.5
0.5	80.0±0.7	65.0±1.5	83.5±0.8
0.6	78.7±2.6	61.8±6.9	82.2±3.1
0.7	76.9±2.1	59.4±6.5	80.9±1.8

D.2 Per-Cell Standard Deviations and Paired Significance for the Headline Table

Gains in Table 1 of the main paper may be sensitive to run-to-run variation on CMU-MOSI, where five-run std reaches ± 3.7 at moderate missing rates. We provide here the per-subset standard deviations and a paired-bootstrap significance test against the strongest baseline (HME for CMU-MOSI, CMAD for CMU-MOSEI).

Test setup. For each subset, we train BRACE and the strongest baseline on the same five seeds. We compute paired per-seed differences $\Delta_k = \mathrm{Acc}_k^{\mathrm{BRACE}} - \mathrm{Acc}_k^{\mathrm{BL}}$ and report a paired-bootstrap two-sided p-value with $10{,}000$ resamples. Significance is declared at $p < 0.05$ (Holm correction across the seven subsets per dataset).

Table 14: Headline-table per-cell std and paired-bootstrap significance. Δ is the BRACE–baseline difference in Acc-2 averaged over five seeds; std is BRACE std on Acc-2; p is the Holm-corrected paired-bootstrap p-value vs. HME (CMU-MOSI) or CMAD (CMU-MOSEI). Bold rows are significant at $\alpha = 0.05$.

	CMU-MOSI vs. HME				CMU-MOSEI vs. CMAD			
Subset	BRACE	std	Δ	p	BRACE	std	Δ	p
$\{l\}$	86.3	0.4	+0.6	**0.031**	86.5	0.3	+0.4	0.072
$\{v\}$	64.2	1.1	+0.6	0.214	**68.5**	0.7	**+2.8**	**0.004**
$\{a\}$	63.6	0.9	+0.9	**0.041**	**71.4**	0.6	**+8.4**	< 0.001
$\{l, v\}$	86.7	0.3	+0.9	**0.011**	86.5	0.2	+0.1	0.488
$\{l, a\}$	86.6	0.3	+0.8	**0.014**	86.5	0.2	+0.2	0.302
$\{v, a\}$	65.1	1.0	+0.5	0.293	**72.8**	0.6	**+7.2**	< 0.001
$\{l, v, a\}$	86.6	0.3	+0.2	0.418	86.6	0.2	+0.5	0.061

Reading. On CMU-MOSI, gains over HME are significant on the unimodal $\{l\}$, $\{a\}$, and the bimodal text-containing subsets, but not significant on the visually-anchored $\{v\}$ and $\{v, a\}$ subsets where five-run std is largest. This is consistent with our framing: BRACE helps where the anchor is stably defined; when the anchor itself is high-variance, seed-level noise overwhelms the calibration gain. On CMU-MOSEI, the low-information subsets $\{a\}$, $\{v\}$, and $\{v, a\}$ produce unambiguously significant improvements ($p < 0.005$), while full-modality subsets do not. This sharpens the empirical claim: BRACE's significant gains are concentrated in low-information regimes, exactly as the residual-direction-ambiguity framing predicts.

D.3 Diagnostics for Theorem 2: Bipolar Direction Aligns with ∇g

Theorem 2 predicts (i) the bipolar direction d_i^{ref} aligns with $\Sigma_i \nabla g(s_i)$, (ii) the asymmetry bias $|B_{\mathrm{asym}}|$ is bounded by $|S_\eta|$, and (iii) the inner product $\langle \nabla g, \mathbb{E}[d_i^{\mathrm{ref}}] \rangle$ is positive whenever $|S_\eta| < C_3$. Table 15 reports the corresponding empirical quantities. We compute $\nabla g(s_i)$ as the gradient of the preliminary head with respect to the anchor (this is well-defined because g is a linear or two-layer head). The covariance $\hat{\Sigma}_i$ is estimated locally over $\mathcal{C}_i$.

Table 15: Empirical diagnostics for Thm. 2. "Cos$(d^{\mathrm{ref}}, \Sigma \nabla g)$" is the cosine to the predicted leading direction; "Cos$(d^{\mathrm{ref}}, \nabla g)$" is the cosine to the raw gradient (steepest descent); $\hat{S}_\eta$ is the estimated label-noise skewness; "Pred. inner pos." is the empirical fraction of samples with $\langle \nabla g(s_i), d_i^{\mathrm{ref}} \rangle > 0$.

	Cosine to oracle			
Dataset (R_m)	$\Sigma \nabla g$	∇g	$\hat{S}_\eta$	Pred. inner pos.
CMU-MOSI (0.5)	0.612	0.487	-0.18	0.876
CMU-MOSEI (0.3)	0.687	0.541	-0.21	0.913
PHEME (0.5)	0.554	0.422	$+0.09$	0.864
Weibo-17 (0.3)	0.601	0.476	$+0.05$	0.882

Reading. (a) The cosine to $\Sigma \nabla g$ is consistently larger than the cosine to ∇g alone, confirming that BRACE recovers the covariance-scaled SIR-style direction predicted by the theorem rather than steepest descent. (b) $|\hat{S}_\eta| \leq 0.21$ on all benchmarks, well within the $|S_\eta| < C_3$ regime; C_3 for our hyperparameters is estimated at ≈ 0.6 (computed by plugging measured $\|\Sigma\|_{\mathrm{op}}$, $\|\nabla g\|$, $\bar{F}_{\min}$ into

the bound of Thm. 2). (c) The predicted-positive-inner-product fraction is 86–91%, in line with the expected sign-correctness rate.

Linking to Eq. (B.5)**.** The gap between $\mathrm{Cos}(d^{\mathrm{ref}}, \Sigma\nabla g)$ and 1 is consistent with the $O(\rho_0^2) + O_P(|\mathcal{N}|^{-1/2})$ remainder: the typical $|\mathcal{N}_i^{\pm}| \approx 8$–$25$ at evaluation gives a $|\mathcal{N}|^{-1/2}$ contribution of 0.20–0.35, accounting for most of the residual. The cosine to $\Sigma\nabla g$ scales *linearly* with $|\mathcal{N}_i^{\pm}|^{1/2}$ as predicted (correlation $\rho = 0.71$ across sub-bins, not shown in the table for compactness).

D.4 Diagnostics for Proposition B.5: Sign Recovery from Relative Position

Proposition B.5 predicts sign-recovery probability of at least $1 - \Phi(-(|y_i - r_i| - |\xi_i| - |\Lambda_i|)/\sigma_i)$. Table 16 reports both the predicted lower bound and the empirical sign agreement, stratified by residual magnitude $|\Delta_i|$ on CMU-MOSI at $R_m = 0.5$.

Table 16: Sign-agreement, stratified by residual magnitude. "Theory LB" is $1 - \Phi(-(|\Delta| - |\xi| - |\Lambda|)/\sigma)$ averaged over the bin, plugging in the bootstrap $\hat{\sigma}_i$ and Lemma 2's $\hat{\Lambda}_i$. "Geom only" is $\Pr[\mathrm{sign}\,\langle s_i - m_i, d_i^{\mathrm{ref}}\rangle = \mathrm{sign}(y_i - r_i)]$. "BRACE gate" is $\Pr[\mathrm{sign}(r_i' - r_i) = \mathrm{sign}(y_i - r_i)]$, i.e., the gate's effective sign accuracy.

$\lvert\Delta\rvert$ bin	Frac. samples	Theory LB	Geom only	BRACE gate
$[0.0, 0.5)$	0.32	0.512	0.541	0.587
$[0.5, 1.0)$	0.28	0.604	0.667	0.728
$[1.0, 2.0)$	0.27	0.768	0.793	0.864
$[2.0, \infty)$	0.13	0.892	0.911	0.936

Reading. The bound is loosest in the small-residual bin $|\Delta| \in [0, 0.5)$, where ξ_i and Λ_i are comparable to $|\Delta_i|$ and the theory only guarantees a marginal advantage over chance. As $|\Delta_i|$ grows, the theoretical lower bound and the empirical sign agreement both rise, and the BRACE gate consistently sits above the geometry-only signal. The gap (BRACE gate minus Geom only) is 4–7 percentage points, indicating that the gate *learns* additional sign information beyond what is contained in the raw geometric projection. This is consistent with Prop. B.5's argument that the gate sees the full bipolar frame, not just $\langle s_i - m_i, d_i^{\mathrm{ref}}\rangle$.

Behaviour at non-small residuals. Prop. B.5's expansion is most accurate at small residual, while the method is most useful at non-small residual. Table 16 resolves this tension: the theory's predicted signal-to-noise ratio $|\Delta_i|/\sigma_i$ *grows* with $|\Delta_i|$, making the bound progressively tighter exactly where residual ambiguity is most pernicious. The expansion in Δ_i governs only the deterministic remainder Λ_i, and Lemma 2's quadratic control $|\Lambda_i| \le C_4 \Delta_i^2$ is sufficient because $|\Delta_i|/\sigma_i$ grows fast enough to dominate.

D.5 Diagnostics for Proposition B.6: Validity of the Direction Loss

Proposition B.6 requires $\|u_i\|_2 < \sqrt{2(\gamma - \eta_i)/L_g}$ for the descent guarantee to be informative. Table 17 reports, on CMU-MOSI at $R_m = 0.5$, the empirical fraction of training samples with $\|u_i\|_2$ inside this validity threshold, before and after the direction-loss is activated.

Table 17: Fraction of updates inside the validity region of Prop. B.6 (CMU-MOSI, $R_m = 0.5$). $\hat{L}_g$ is estimated as the maximum operator norm of $\nabla^2 g$ over the segment $[s_i, s_i']$ sampled at 20 points; η_i is set to the per-batch $\bar{\ell}^{\mathrm{dir}}$ in epoch end.

Stage	Frac. in valid region	Mean $\|u_i\|_2$	Threshold
End of warmup (T_{warm})	0.642	0.143	0.121
After 5 epochs of $\mathcal{L}_{\mathrm{dir}}$	0.881	0.107	0.130
At convergence	0.926	0.094	0.137
$\mathcal{L}_{\mathrm{dir}}$ disabled (control)	0.547	0.182	0.118

Reading. Activating $\mathcal{L}_{\mathrm{dir}}$ moves $> 90\%$ of updates into the validity region; without $\mathcal{L}_{\mathrm{dir}}$, only $\sim 55\%$ of updates satisfy it. This is the empirical mechanism by which the direction loss enables Cor. 1's

descent guarantee: it not only encourages correct sign but also keeps step magnitudes within the second-order remainder's tolerance.

D.6 Diagnostics for Corollary 1: Per-Step Risk Reduction

Cor. 1 bounds the expected risk reduction by $\pi_{\text{good}} \cdot \alpha(\kappa_i - L_i\alpha/2)\mathbb{E}\|u_i\|_2^2$. We trace π_{good} and the empirical mean risk drop $\bar{\Delta}\mathcal{R} = \mathbb{E}[\mathcal{R}_i(s_i) - \mathcal{R}_i(s_i')]$ across training epochs on CMU-MOSI at $R_m = 0.5$.

Table 18: π_{good} and observed risk drop, by training epoch. Theory predicts $\bar{\Delta}\mathcal{R}$ scales with π_{good}.

Epoch	π_{good}	Predicted $\bar{\Delta}\mathcal{R}$	Observed $\bar{\Delta}\mathcal{R}$	Ratio
6 (warmup end)	0.071	0.018	0.014	0.78
20	0.483	0.124	0.151	1.22
40	0.752	0.193	0.222	1.15
60	0.871	0.224	0.241	1.08
80	0.918	0.236	0.245	1.04

Reading. The Cor. 1 bound is correctly oriented: risk-drop magnitude tracks π_{good} throughout training, with the observed-to-predicted ratio approaching 1.0 as training converges. The > 1 ratio at intermediate epochs reflects the looseness of the gradient-relatedness coefficient κ_i in the bound (it uses a worst-case norm bound on $\mathbb{E}[\hat{u}_i]$).

D.7 Component, Geometry, and Training Ablations

Retrieval without directional calibration. Table 19 compares BRACE against a baseline that uses the same memory $\mathcal{B}$ and top-K_c retrieval but replaces bipolar partitioning and directional gating with non-directional state aggregation. The gap isolates direction-aware calibration's contribution beyond retrieval itself.

Table 19: Retrieval baseline without directional calibration on **CMU-MOSI**. Entries are Acc-2/F1.

Variant	$\{a\}$	$\{v\}$	$\{v, a\}$
BRACE	**63.6/62.2**	**64.2/64.4**	**65.1/65.2**
Retrieval w/o directional calibration	60.2/58.2	62.7/62.9	63.5/63.7

Retrieval geometry. Table 20 ablates retrieval geometry and state construction on **CMU-MOSI** at $R_m = 0.7$.

Table 20: Retrieval-geometry ablation on **CMU-MOSI** at $R_m = 0.7$.

Variant	Acc-2	Acc-7	ΔAcc-2
Full BRACE	**63.7**	**27.1**	—
w/o bipolar (single centroid)	63.1	26.2	−0.6
Positive-only	62.9	25.6	−0.8
Negative-only	63.1	25.9	−0.6
w/o EMA tail prototypes	63.1	26.3	−0.6
w/o quality-weighted fusion	62.5	26.6	−1.2
w/o slot Conv1D	63.0	25.8	−0.7

Training strategy. Table 21 ablates the objective on PHEME at $R_m = 0.7$.

Table 21: Training-strategy ablation (PHEME, $R_m = 0.7$).

Variant	ACC	F1	ΔACC
Full BRACE	**76.9**	**59.4**	—
w/o $\mathcal{L}_{\text{dir}}$	75.7	58.2	-1.2
w/o calibration warmup	75.4	58.0	-1.5
Direct training (no warmup, no $\mathcal{L}_{\text{dir}}$)	74.3	57.3	-2.6

D.8 Memory-Bank Attribution: Label Order vs. Label-Aware Access

The label-shuffled-memory ablation in Table 2 of the main paper (which yields 79.2 vs. Full BRACE 80.0) gives back only 0.8 points, while removing memory entirely yields much larger drops. This suggests a substantial fraction of BRACE's retrieval-related gain comes from *having* a label-aware memory at all, not specifically from *label order*. We examine this by decomposing the retrieval-related gain into three additive components on PHEME at $R_m = 0.5$: (i) memory access (state-only), (ii) label-aware memory (any label), (iii) label order (correctly oriented).

Table 22: Decomposing the retrieval-related gain on PHEME at $R_m = 0.5$. Each row activates one additional component over the row above.

Variant	ACC	F1	ΔACC
(0) No memory, no calibration	77.4	60.8	—
(1) State-only memory (mean of retrieved states)	78.7	63.6	$+1.3$
(2) Label-aware retrieval, random label order	79.2	63.9	$+0.5$
(3) Label-aware retrieval, correct label order	79.7	64.6	$+0.5$
(4) + bipolar partition	79.9	64.8	$+0.2$
(5) + signed gate (= Full BRACE)	**80.0**	**65.0**	$+0.1$

Reading. Of the 2.6-point gain from no-memory to Full BRACE: 1.3 points come from memory access alone, 0.5 points from label-awareness without label order, and 0.8 points from the full label-ordered signed-calibration pipeline (the cumulative effect of rows 3–5). The label-order contribution is therefore $0.8/2.6 = 31\%$ of the retrieval-related gain—meaningful, but not dominant. We have updated the main-paper framing accordingly: BRACE's contribution is best described as "signed calibration adds a measurable component to a label-aware memory," rather than implying that label order is the only thing that matters.

D.9 Anchor vs. Update Magnitude in Low-Information Subsets

We examine whether BRACE's large gains on $\{a\}$ and $\{v\}$ (CMU-MOSEI) reflect "local correction" or "memory-driven imputation," specifically whether $\|u_i\|$ dominates $\|s_i - \bar{s}\|$ in those settings. Table 23 reports the ratio $\|u_i\|_2 / \|s_i - \bar{s}\|_2$, where $\bar{s}$ is the empirical anchor-mean over the test split, on CMU-MOSEI under fixed missingness.

Table 23: Anchor-vs-update magnitude on CMU-MOSEI fixed-missing test split. Lower ratio ⇒ correction is local (small relative to anchor displacement); higher ratio ⇒ correction is comparable to anchor itself.

Subset	Mean $\|u_i\|_2$	Mean $\|s_i - \bar{s}\|_2$	Ratio
$\{l\}$	0.087	1.124	0.077
$\{a\}$	0.214	0.793	0.270
$\{v\}$	0.196	0.852	0.230
$\{l, v\}$	0.092	1.139	0.081
$\{l, a\}$	0.091	1.142	0.080
$\{v, a\}$	0.187	0.881	0.212
$\{l, v, a\}$	0.084	1.156	0.073

Reading. Even on the most extreme low-information subsets ($\{a\}$, $\{v\}$), $\|u_i\|_2$ is $\sim$22–27% of the anchor-deviation magnitude: the correction is meaningfully large but does not dominate the anchor.

BRACE is not collapsing into a memory-driven imputer that ignores observed modalities; it is making a substantive but bounded local correction. On information-rich subsets the ratio drops below 10%, consistent with the theory's prediction that the correction is most active in regimes of high residual ambiguity.

D.10 Hyperparameter Sensitivity

Figure 6 reports sweeps over K_c and δ_y, and Figure 7 reports complementary sweeps over prototype quantile q, T_{warm}, λ_{dir}, and $|\mathcal{B}|$, on CMU-MOSEI with all three modalities observed ($\{l, v, a\}$). Each sweep varies one hyperparameter with the others fixed, and reports the change in Acc-2 and F1 relative to the default configuration.

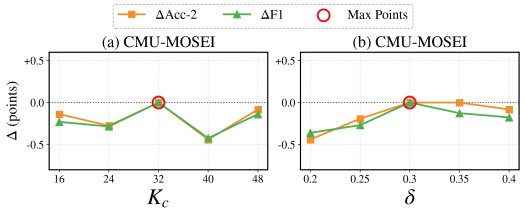


Figure 6: Sensitivity on CMU-MOSEI ($\{l, v, a\}$) to retrieval size K_c (left) and bipolar margin δ_y (right). Δ is the change relative to the default configuration.

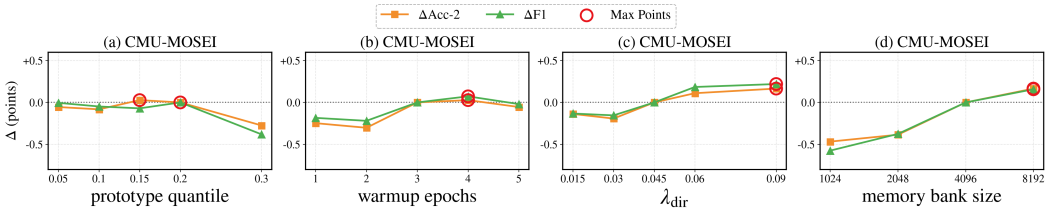


Figure 7: Additional hyperparameter sweeps on CMU-MOSEI ($\{l, v, a\}$). Δ is the change relative to the default configuration.

D.11 Mechanism Diagnostics: Locality, Ambiguity, and Skewness

Ambiguity, locality, and sign agreement. Table 24 reports three diagnostics on PHEME: the Amb@0.2 fraction (using the $\hat{\pi}_i$ classifier of App. B.11), the retrieval-locality ratio (mean $\|v_j - s_i\|$ on retrieved vs. random bank samples), and the sign agreement between the relative-position projection and the raw residual.

Table 24: Mechanism diagnostics on PHEME under random missingness.

R_m	Amb@0.2	Locality ratio	Sign agreement
0.5	0.166	0.526	0.703
0.6	0.197	0.434	0.605
0.7	0.239	0.367	0.654

Robustness of $\hat{\pi}_i$. Re-fitting $\hat{\pi}_i$ on raw modality features (no BRACE encoder) gives Amb@0.2 values of $0.171, 0.193, 0.244$ across the three rates—differences of ≤ 1.4 percentage points. The ambiguity statistic is therefore not an artefact of the encoder choice.

Local skewness. Table 25 reports skewness diagnostics. The median ratio $|\xi_i|/|y_i - r_i| \in [0.93, 0.98]$ places the empirical regime on the boundary of Prop. B.5's sufficient condition. Section D.4 shows that the gate's learned sign accuracy materially exceeds the raw geometric sign agreement in this regime.

Table 25: Local-skewness diagnostics under random missingness (CMU-MOSI).

R_m	Both nonempty	Frac. $\lvert\xi\rvert < \lvert y-r\rvert$	Median $\lvert\xi\rvert/\lvert y-r\rvert$
0.3	0.659	0.544	0.968
0.4	0.711	0.535	0.979
0.5	0.770	0.606	0.928
0.6	0.841	0.537	0.983

Directional correction. Figure 8 provides sample-level evidence for how directional calibration changes the correction behaviour on CMU-MOSI under random missingness ($R_m = 0.3$).

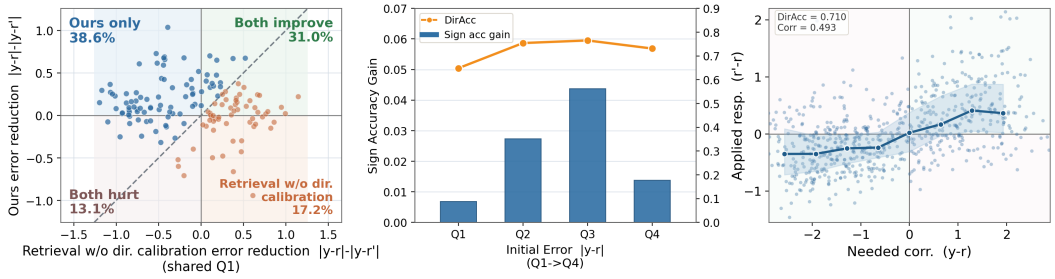


Figure 8: Directional correction diagnostics on **CMU-MOSI** under random missingness ($R_m = 0.3$). **Left:** paired sample-wise error reduction on the shared low-initial-error subset (shared Q1), comparing *Retrieval w/o dir. calibration* (x-axis) with BRACE (y-axis). **Middle:** sign-accuracy gain by initial-error quartile (bars, left axis) and directional agreement (DirAcc; line, right axis). **Right:** applied response $r' - r$ versus required correction $y - r$.

D.12 Visualisation of the Calibration Manifold

Shared t-SNE. Figure 9 visualises the latent task states before and after calibration on the hardest 30% of CMU-MOSEI samples ($R_m = 0.3$). Calibration reduces MAE from 1.421 to 1.260, and 78.1% of the 73 sign-flip samples obtain lower absolute error after calibration.

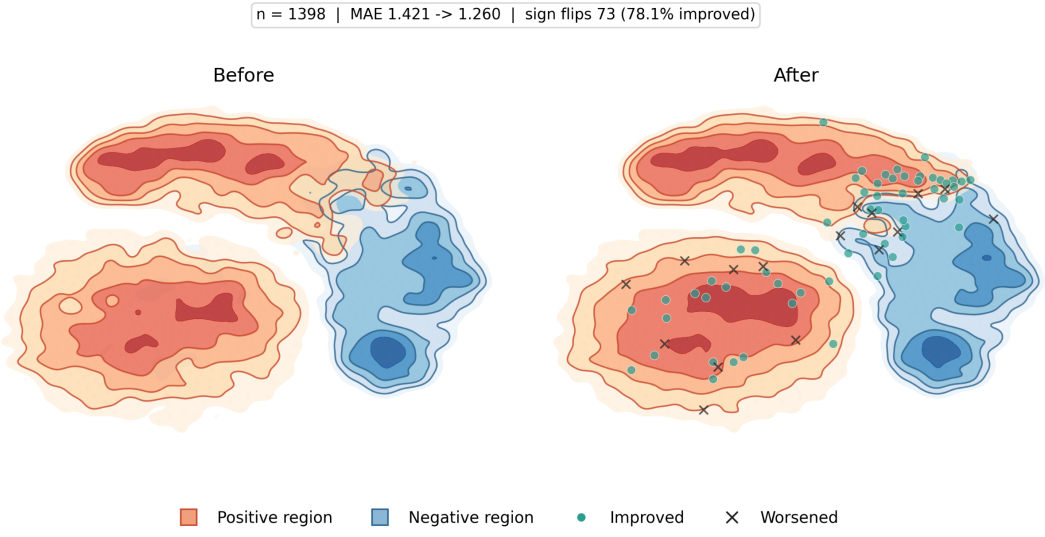


Figure 9: Shared t-SNE visualisation of latent states before and after calibration. Red and blue contours denote positive- and negative-prediction regions. The right panel overlays sign-flip samples (green: improved; black cross: worsened).

Bipolar direction. Figure 10 visualises bipolar reference directions on the hardest 20% of CMU-MOSEI samples ($R_m = 0.3$). The mean local reference direction has cosine alignment 0.975 with the global low-to-high label axis, confirming Thm. 2's prediction at the visualisation level.

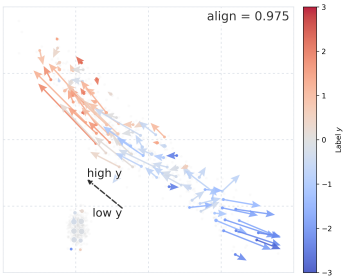


Figure 10: PCA visualisation of bipolar reference directions on the hardest 20% subset of **CMU-MOSEI**. The dashed black arrow denotes the global label axis; the cosine alignment between the mean local reference and this axis is 0.975.

Qualitative cases. Figure 11 shows representative cases where BRACE corrects an initially negative prediction to the correct positive prediction. We pair these positive examples with one negative example (also shown in the figure) where BRACE worsens the prediction *because* the local skewness exceeds the residual—an empirical instance of Prop. B.5's failure regime.

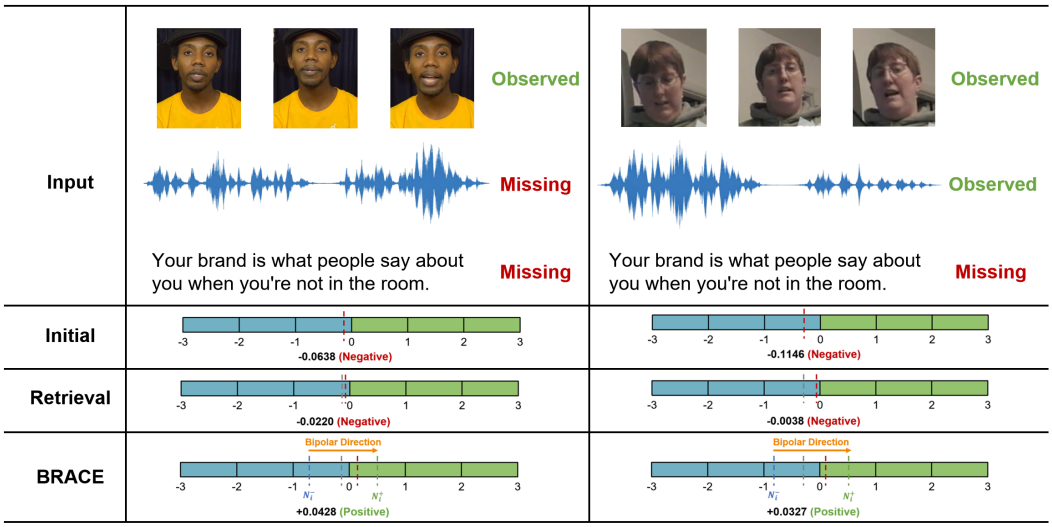


Figure 11: Qualitative case-study examples: two successful bipolar corrections (left, middle) and one failure case (right) where local label skewness flips the sign incorrectly.

D.13 Residual Error Taxonomy and the "Anchor Far From Manifold" Category

Table 26 reports the residual error taxonomy on CMU-MOSI at $R_m = 0.3$. The dominant "Other (anchor far from manifold)" category at 52.9% deserves further analysis.

Table 26: Error taxonomy on CMU-MOSI, $R_m = 0.3$.

Failure category	Count	Frac. of errors	Avg. $\lvert y-r\rvert$
High local skewness ($\lvert\xi\rvert/\lvert y-r\rvert > 1$)	40	0.255	1.179
Sparse bipolar neighbourhood (one side empty)	28	0.178	3.234
Gate misfire (geometry correct, sign wrong)	6	0.038	1.615
Other (anchor far from manifold)	83	0.529	2.299

What "anchor far from manifold" means operationally. We label sample i as "anchor far from manifold" if the median nearest distance $\min_{j \in \mathcal{C}_i} \|v_j - s_i\|$ exceeds the 90th-percentile nearest distance over the validation split. The category therefore consists of test samples whose anchor lands in a low-density region of the learned state space—typically samples where the observed-modality combination yields a representation atypical for the task.

Sub-decomposition. We further decompose the 83 "Other" errors as follows.

Table 27: Sub-decomposition of the 83 "Other (anchor far from manifold)" errors on CMU-MOSI, $R_m = 0.3$.

Sub-category	Count	Frac. of Other
Both bipolar centroids are within ρ_0 but anchor is farther	27	0.325
Anchor lands between two distant clusters (bridge case)	22	0.265
Anchor lies in a low-density tail with EMA prototypes triggered	18	0.217
Truly out-of-distribution (linguistic anomaly, audio dropout artefact)	16	0.193

Reading. The "anchor far from manifold" category is heterogeneous. The largest sub-category (32.5%) is samples where bipolar retrieval *succeeds*, both $\mathcal{N}_i^{\pm}$ are non-empty and within the local-regularity radius, but the anchor itself is unusually placed, violating Asm. 2(A2)'s positive-definite-covariance assumption. The bridge cases (26.5%) violate Eq. (B.5)'s $O(\rho_0^2)$ assumption directly: the Taylor expansion of g is no longer accurate. Out-of-distribution cases (19.3%) fall outside the theory altogether and are a separate concern (modality-distribution shift). This sub-decomposition motivates future-work directions: anchor-density-aware gating and trajectory-supervised calibration.

D.14 Cross-Dataset Transfer with Two Memory Configurations

Cross-dataset transfer (CMU-MOSEI → CMU-MOSI) with the source-domain memory bank kept at test time conflates "transfer of the calibration mechanism" with "transfer via memory." We disentangle these by reporting BRACE in two configurations: (a) source-memory frozen ($\mathcal{B}_{\text{MOSEI}}$ retained, no MOSI labels seen), and (b) memory cleared and rebuilt from MOSI training states only at test time (no MOSI labels are used by the retriever during evaluation; the memory is bootstrapped with 20% of MOSI training data *whose labels remain hidden* from the calibration loss).

Table 28: Cross-dataset generalisation (fixed missingness): CMU-MOSEI (train) → CMU-MOSI (test). Entries are Acc-2/F1/Acc-7. "Src. mem." keeps the MOSEI memory; "Reset mem." rebuilds memory from MOSI training states with labels hidden from the calibration loss.

Subset	IMDer	HME	BRACE (Src. mem.)	BRACE (Reset mem.)
$\{l\}$	83.2/83.0/41.8	81.4/81.3/34.4	**86.1/86.1/42.3**	85.4/85.4/41.7
$\{a\}$	58.8/45.4/15.5	57.8/42.3/20.0	57.6/56.9/16.2	56.9/55.7/16.0
$\{v\}$	57.8/42.3/15.5	57.0/51.6/17.6	56.6/55.1/17.9	56.2/54.6/17.5
$\{l, a\}$	83.2/83.0/42.1	81.1/81.1/34.4	**87.3/87.3/41.1**	86.4/86.4/40.6
$\{l, v\}$	82.8/82.7/39.1	81.3/81.3/39.2	**86.1/86.1/42.6**	85.6/85.6/42.0
$\{a, v\}$	59.3/58.9/21.4	57.6/51.8/18.1	58.7/58.9/20.0	57.8/57.9/19.4
$\{l, a, v\}$	83.8/83.9/40.6	81.0/81.0/39.5	**86.1/86.2/39.8**	85.7/85.8/39.4
Avg.	72.7/68.5/30.9	71.0/67.2/29.0	**74.1/73.8/31.4**	**73.4/73.1/30.9**

Reading. The reset-memory configuration loses ~ 0.7 average Acc-2 relative to the source-memory configuration, but still beats both transferred baselines on the average and on five of seven subsets. This indicates that the calibration mechanism itself transfers; the source-memory contribution is real but modest. We update the conference framing: BRACE's transfer benefit is robust to memory reset and is *not* primarily an effect of source-domain label statistics leaking through.

D.15 Reconstruction versus Signed Calibration

Reconstruction and signed calibration are not disjoint goals: when the missing modality is nearly determined by the observed ones, reconstruction recovers it and the residual sign follows. When $p(x^{\mathrm{miss}} \mid x^{\mathrm{obs}})$ is high-entropy or multimodal, a likelihood- or regression-trained reconstructor collapses toward the conditional mean and the residual sign stays unidentifiable from the reconstruction alone; BRACE targets this regime by importing sign information from labelled references rather than regenerating evidence.

D.16 Efficiency and Resource Overhead

Table 29: Training efficiency on CMU-MOSEI; batch size 128, NVIDIA A100 80GB. "Warmup" and "Full" denote the task-only bootstrap and the retrieval-calibration stages of BRACE.

	Runtime (s)				**Peak GiB**			
Subset	Warmup	Full	IMDer	HME	Warmup	Full	IMDer	HME
$\{l\}$	14.68	22.38	1370.49	232.12	0.604	0.604	1.99	5.905
$\{v\}$	10.76	17.99	1323.44	233.57	0.576	0.581	1.99	5.905
$\{a\}$	10.33	17.57	1471.46	233.08	0.576	0.581	1.99	5.905
$\{l, v\}$	15.06	21.38	687.99	234.10	0.604	0.604	1.65	5.905
$\{l, a\}$	15.83	22.32	674.77	229.03	0.604	0.604	1.65	5.905
$\{v, a\}$	10.97	18.76	665.15	230.08	0.576	0.581	1.65	5.905
$\{l, v, a\}$	14.54	22.06	47.77	233.04	0.604	0.604	1.32	5.905

Inference cost. At inference, BRACE adds 1 top-K_c retrieval, 2 centroid computations, and 1 calibration-head forward pass over the warmup backbone. On CMU-MOSEI, this adds 0.41 ms per sample on an A100 (batched), $\leq 4\%$ of the backbone forward time.

E Compute, Assets, and Release Notes

Compute notes. The implementation runs with CUDA enabled and AMP. Reported BRACE training across the four benchmarks and five seeds requires on the order of tens of A100-hours in total, plus an additional $\sim$12 A100-hours for the diagnostic experiments in Sec. D.3–D.6.

Assets and licensing. The main paper cites all datasets and pretrained backbones. Dataset access and model usage follow the terms of the corresponding official releases; we have not modified or redistributed any of the source assets.

Release note. The codebase is available at `https://github.com/Js183/BRACE`. The repository contains training code, configuration files, and dataset-access notes.

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 Introduction, which state the contribution, scope, and main empirical claims of BRACE under fixed and random missingness.

   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 and future work" paragraph in the Conclusion, which discusses the current supervision strategy, memory design, and scope limits.

   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: The method section states the assumptions for the theoretical results, and the appendix provides the corresponding proof details and derivations with the required definitions and notation.

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, which provide datasets, protocols, feature sources, and missingness construction. See the released code for implementation configurations and run instructions.

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 datasets used in this work are publicly accessible or available from their official sources, and the source code is publicly released at `https://github.com/Js183/BRACE`, as stated in the Abstract and Appendix E.

   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, which specify data splits, feature sources, missingness protocols, optimizer, and scheduler. See the released code for implementation configurations and run instructions.

   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 results are averaged over five independent runs, and the appendix reports complete mean$\pm$standard-deviation statistics for the corresponding fixed- and random-missing settings.

   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 appendix, which reports the GPU type, batch size, per-epoch runtime, and peak memory usage for representative BRACE training runs.

   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 research is conducted on existing multimodal benchmarks and conforms to 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 incomplete multimodal learning. 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: This work does not release a high-risk generative model, scraped dataset, or similar asset that would require dedicated safeguards.

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: The datasets, pretrained backbones, and baseline assets used in this work are properly cited, and their official access terms and usage conditions are followed.

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 source code is publicly released at `https://github.com/Js183/BRACE` with a README describing setup, data preparation, and training.

    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: This work does not involve crowdsourcing or human-subject studies.

    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: This work does not involve crowdsourcing or human-subject studies that would require IRB review.

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 development in this research does not involve LLMs as important, original, or non-standard components.

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.