CONSENSUS–COMPLEMENT SUBSPACE FUSION FOR MULTIMODAL SENTIMENT ANALYSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal sentiment analysis (MSA) seeks to infer the sentiment of an utterance by integrating its language, acoustic and visual streams, whose importance and agreement vary across utterances. Traditional fusion methods either align the modalities to exploit their agreement or decouple them into invariant and specific parts, with heuristic losses whose balance is tuned by hand. Recent advances have placed alignment on a principled footing by driving the representation matrix of an instance towards rank one. However, full alignment discards modality-specific information, and no rule states how much of it should be retained. To address the challenges, in this paper, we propose PRISM, a novel framework that fuses modalities through an exact orthogonal decomposition of the instance matrix. Specifically, we are grounded in a theoretical insight. The rank-one truncation of the instance matrix spans a consensus subspace, and its Eckart–Young residual carries modality-specific innovation, termed the complement. We prove that full alignment incurs an excess risk equal to the label variance explained by modality-specific latents. Under a linear-Gaussian latent model, the Bayes-optimal predictor instead adds a consensus score to individually shrunk complement scores. PRISM realises this predictor. It weights the consensus by the leading right singular vector, a training-free modality router certified by the consensus ratio. It shrinks each complement by a certificate-driven gate, the minimum-mean-square-error coefficient of the specific score. Besides, a Davis–Kahan analysis yields a spectral-gap regulariser that stabilises the decomposition. Extensive experiments on CMU-MOSI, CMU-MOSEI and CH-SIMS benchmarks demonstrate that PRISM achieves state-of-the-art performance.

1 INTRODUCTION

Humans judge the sentiment of a speaker from what is said, how it is said and the accompanying facial expression. Multimodal sentiment analysis (MSA) (Zadeh et al., 2016; 2018b; Yu et al., 2020) seeks to simulate this capability by integrating the language, acoustic, and visual streams of an utterance into a single prediction of sentiment intensity. The modalities are *heterogeneous*. Language usually dominates, whereas acoustic and visual features are noisy and only intermittently informative, so their importance changes from sample to sample (Peng et al., 2022; Fang et al., 2025). They are also only *partially redundant*. The three streams share a common affective content, while the unshared parts, e.g., a sarcastic tone contradicting positive words, often decide the label (Hazarika et al., 2020; Li et al., 2023b). Figure 1(c) shows such an utterance.

Fusion methods respond to these observations in two ways. Alignment-based methods, from tensor and attention fusion (Zadeh et al., 2017; Tsai et al., 2019a; Rahman et al., 2020) to recent volume-based objectives (Cicchetti et al., 2025), pull the modalities together so that their agreement can be exploited. At the extreme, driving the representation matrix of an instance towards rank one collapses all modalities onto a single self-discovered direction. It removes modality imbalance by construction, as shown in Figure 1(a). Decoupling-based methods (Hazarika et al., 2020; Yu et al., 2021; Li et al., 2023b) instead split every modality into an invariant and a specific part. Dynamic-expert methods (Fang et al., 2025) route the modalities with a learned gate. However, the two families sit at opposite ends of a continuum without a rule for where to stop. Full alignment discards the modality-specific part that MSA relies on. Decoupling and routing retain it through auxiliary losses

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

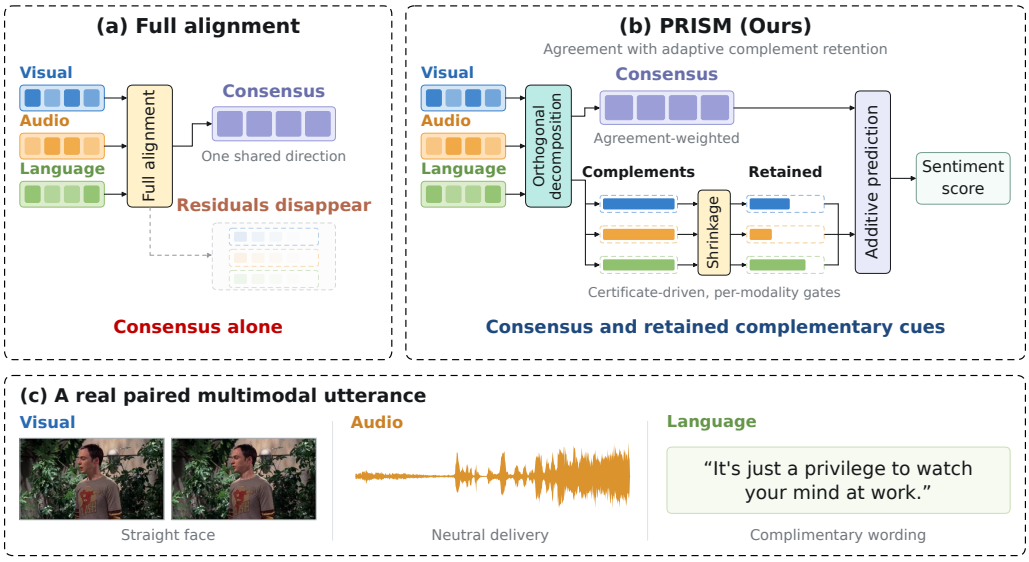


Figure 1: **Full alignment versus PRISM.** (a) Full alignment collapses the modalities onto one shared direction and discards the residuals that carry modality-specific cues. (b) PRISM decomposes the representations into an agreement-weighted consensus and per-modality complements, which certificate-driven gates shrink before an additive head. (c) An utterance whose complimentary wording is contradicted by a straight face and a neutral delivery: the contradiction lives in the complements.

whose hand-tuned weights set the balance between agreement and distinctness. These limitations underscore the need for a fusion method with sound principles.

In this work, we initiate our research from the geometry of a single utterance. We stack its normalised representations into an instance matrix $\boldsymbol{Z} = [\boldsymbol{z}^1, \ldots, \boldsymbol{z}^k]$. The leading singular direction $\boldsymbol{u}_1$ spans a one-dimensional *consensus* subspace. Its orthogonal complement collects the *innovation* of every modality relative to the consensus (Kailath et al., 2000; Xi et al., 2026). This complement is exactly the Eckart–Young residual (Eckart & Young, 1936) that full alignment drives to zero. This leads to a critical insight: the residual cannot be discarded. Under a natural latent-variable model, any fully aligned representation is blind to the modality-specific latents, with an excess risk equal to the label variance they explain (Theorem 4). Full alignment fixes the balance problem at the price of the specialisation problem (Fang et al., 2025). We then quantify what each part buys (Theorem 5). The consensus averages the shared factor over k noisy views, a *balance gain*. The complements add the specific factors, a *specialisation gain*. Under a linear-Gaussian instance of this model, the Bayes-optimal use of a complement is to shrink its specific score by $\alpha_m = 1/(1+\kappa_m)$, where κ_m is its inverse signal-to-noise ratio. Finally, a Davis–Kahan argument (Yu et al., 2015) bounds the sensitivity of the decomposition by the inverse spectral gap (Theorem 7), which motivates regularising the *gap* rather than the residual.

Motivated and implemented by these principles, we term our method PRISM (**PR**incipled consensus–complement **I**nstance-wise **S**ubspace fusion for **M**ultimodal sentiment), depicted in Figure 1(b). Specifically, the consensus vector $\boldsymbol{Z}\boldsymbol{v}_1 = \sigma_1 \boldsymbol{u}_1$ is a weighted combination of the modalities whose weights, the leading right singular vector $\boldsymbol{v}_1$, are proportional to each modality's cosine agreement with the consensus and are certifiably balanced once the consensus ratio exceeds $1 - 1/k$, i.e., a per-instance modality router without a router network. Besides, the complements are shrunk by certificate gates $\alpha_m = 1/(1 + \kappa_m)$ predicted from the spectral statistics of the instance, and the prediction is additive over the two parts, mirroring the Bayes predictor of Theorem 5. We verify PRISM with extensive experiments on three MSA benchmarks and demonstrate its superiority over strong baselines. Our contributions are summarised as follows:

- We prove that full multimodal alignment is predictively incomplete: the excess risk of any fully aligned representation is lower bounded by the label variance explained by modality-specific latents (Theorem 4).
- To operationalise this insight, we introduce PRISM, which fuses modalities through an exact orthogonal consensus–complement decomposition of the instance matrix. Its agreement weights are closed-form, with a sign pattern certified by the consensus ratio (Lemma 2, Proposition 3). Its complement shrinkage is the Bayes-optimal coefficient of the modality-specific score under a linear-Gaussian latent model and equals a minimum-norm subspace edit (Theorem 5). Its gap regulariser is justified by a Davis–Kahan bound (Theorem 7).
- Extensive experiments on CMU-MOSI, CMU-MOSEI and CH-SIMS demonstrate that PRISM achieves state-of-the-art performance under an identical protocol and matches a learned router at no parameter cost.

2 Related Work

Multimodal sentiment analysis. MSA regresses the sentiment of an utterance from its language, acoustic and visual streams (Zadeh et al., 2016; 2018b; Yu et al., 2020; Mao et al., 2022; Liang et al., 2024). Fusion methods model cross-modal interactions explicitly (Zadeh et al., 2017; Liu et al., 2018; Zadeh et al., 2018a; Pham et al., 2019; Tsai et al., 2019a) or, with pre-trained language models (Devlin et al., 2019), inject the non-verbal cues into the text stream (Rahman et al., 2020; Han et al., 2021a; Wang et al., 2023b;a; Zhang et al., 2023a; Sun et al., 2022; Hu et al., 2022). Decoupling methods split every modality into a shared and a specific part (Tsai et al., 2019b; Sun et al., 2020; Hazarika et al., 2020; Yang et al., 2022; 2023; Li et al., 2023b; Yu et al., 2021; Han et al., 2021b; Sun et al., 2024), kept apart by auxiliary losses whose weights are tuned by hand.

Modality imbalance and dynamic fusion. A dominant modality suppresses the others in joint training (Wang et al., 2020; Huang et al., 2022; Wu et al., 2022). Remedies modulate gradients (Peng et al., 2022; Li et al., 2023a; Fan et al., 2023; Wei & Hu, 2024), alternate unimodal learning (Du et al., 2023; Zhang et al., 2024; Hua et al., 2024), fuse dynamically per sample (Xue & Marculescu, 2023; Zhang et al., 2023b; Feng et al., 2024) or route a mixture of modality experts (Jacobs et al., 1991; Shazeer et al., 2017; Fang et al., 2025). PRISM instead obtains the sample-level weights and the shared/specific split from one SVD of the instance.

Principled learning with alignment and SVD. Alignment guides representation learning from contrastive pairs (Radford et al., 2021) and anchor binding (Girdhar et al., 2023; Zhu et al., 2024) to the Gram volume (Cicchetti et al., 2025). Its cost is known information-theoretically: shared information suffices under the multi-view assumption (Sridharan & Kakade, 2008; Federici et al., 2020; Tsai et al., 2021) but not when task information is unique to one modality (Liang et al., 2023b;a; Dufumier et al., 2025; Huang et al., 2021; Lu, 2023), which Theorem 4 makes explicit for the rank-one limit. PRISM instead derives the fusion rule from the spectrum of the instance: the leading singular pair (Eckart & Young, 1936; Hotelling, 1936; Andrew et al., 2013) defines a consensus that rewards agreement, its residual isolates each modality's innovation, minimum-mean-square-error shrinkage (Kailath et al., 2000; James & Stein, 1961) sets how much of it the Bayes predictor keeps, and a Davis–Kahan bound (Davis & Kahan, 1970; Yu et al., 2015) turns stability into a regulariser, so that the subspace decompositions recently used as interventions in vision–language models (Lin et al., 2026; Xi et al., 2026) become the fusion operator itself.

3 Preliminary

Multimodal sentiment analysis. An utterance consists of three sequences $(\boldsymbol{x}^\ell, \boldsymbol{x}^a, \boldsymbol{x}^v)$ for the language, acoustic and visual streams, and is annotated with a sentiment score $y \in \mathbb{R}$. MSA learns a predictor of y from the three streams. Modality encoders f^m map each stream to a normalised representation $\boldsymbol{z}^m = f^m(\boldsymbol{x}^m)/\|f^m(\boldsymbol{x}^m)\| \in \mathbb{S}^{d-1}$, $m \in \mathcal{M}$, $|\mathcal{M}| = k$ ($k = 3$ here). Superscripts index modalities and the instance index is omitted. We stack the representations of one instance into the *instance matrix* $\boldsymbol{Z} = [\boldsymbol{z}^1, \ldots, \boldsymbol{z}^k] \in \mathbb{R}^{d\times k}$. Its Gram matrix $\boldsymbol{G} = \boldsymbol{Z}^\top \boldsymbol{Z}$ satisfies $\boldsymbol{G}_{mm} = 1$ and $\operatorname{tr} \boldsymbol{G} = k$.

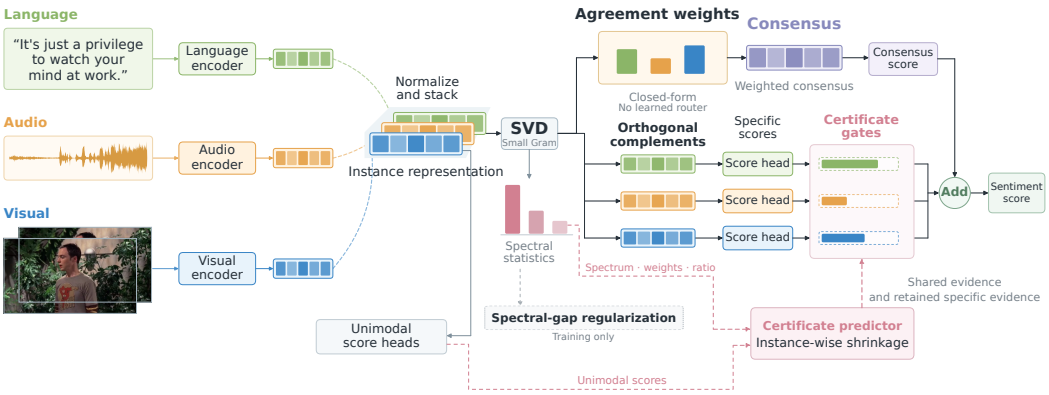


Figure 2: **The overall framework of PRISM.** The encoded and normalised modalities form the instance matrix $\boldsymbol{Z}$, whose 3×3 Gram matrix yields the spectral statistics, the agreement weights $\boldsymbol{v}_1$, the consensus $\boldsymbol{c} = \boldsymbol{Z}\boldsymbol{v}_1$ and the complements $\boldsymbol{q}^m$. A consensus head and modality-specific heads score the two parts, the certificate predictor turns the spectral statistics and unimodal scores into gates $\alpha_m = 1/(1+\kappa_m)$, and the prediction adds the consensus score to the gated specific scores (Eq. (5)), trained with auxiliary supervision and the spectral regularisers (Eq. (6)).

SVD and eigenvalues. Let $\boldsymbol{Z} = \sum_{j=1}^{k} \sigma_j \boldsymbol{u}_j \boldsymbol{v}_j^\top$ be the SVD of the instance matrix with $\sigma_1 \geq \cdots \geq \sigma_k \geq 0$. The eigenvalues of $\boldsymbol{G}$ are $\lambda_j = \sigma_j^2$ and satisfy $\sum_j \lambda_j = k$. By the Eckart–Young theorem (Eckart & Young, 1936), $\tilde{\boldsymbol{Z}} = \sigma_1 \boldsymbol{u}_1 \boldsymbol{v}_1^\top$ is the best rank-one approximation of $\boldsymbol{Z}$, and the residual energy is $\|\boldsymbol{Z} - \tilde{\boldsymbol{Z}}\|_F^2 = \sum_{j \geq 2} \lambda_j$. As is standard for alignment methods, we assume that representations of the same instance do not form obtuse angles. Proposition 3 later certifies this assumption up to a sign flip per modality by a verifiable spectral condition, and the hinge of Eq. (6) removes the flips during training.

Assumption 1 (Non-obtuseness)**.** *For every instance, $\langle \boldsymbol{z}^m, \boldsymbol{z}^{m'} \rangle \geq 0$ for all $m, m' \in \mathcal{M}$, i.e., $\boldsymbol{G}$ is entrywise non-negative.*

Full alignment. Alignment-based methods maximise the pairwise similarities $\boldsymbol{G}_{mm'}$ of an instance. The ideal case is $\boldsymbol{z}^1 = \cdots = \boldsymbol{z}^k$, termed *full alignment*, where every entry of $\boldsymbol{G}$ equals one. Under Assumption 1, full alignment holds iff $\operatorname{rank}(\boldsymbol{G}) = 1$ iff $\sigma_1 = \sqrt{k}$ (Lemma 1). A full-alignment objective therefore drives σ_1 to $\sqrt{k}$, e.g., through a softmax over singular values, $\mathcal{L}_{\text{align}} = -\log \frac{\exp(\sigma_1/\tau)}{\sum_j \exp(\sigma_j/\tau)}$, which makes the Eckart–Young residual vanish. Minimising the Gram volume $\det \boldsymbol{G} = \prod_j \lambda_j$ (Cicchetti et al., 2025) only requires $\operatorname{rank}(\boldsymbol{G}) \leq k-1$, i.e., $\sigma_k \to 0$, but in practice collapses the whole spectrum (Section 5.3).

4 Principled Consensus–Complement Fusion

4.1 Principled Learning

Definition 1 (Consensus subspace, consensus ratio, spectral gap)**.** *The* consensus subspace *of an instance is $\mathcal{C}(\boldsymbol{Z}) = \operatorname{span}(\boldsymbol{u}_1)$ with orthogonal projector $\Pi_{\mathcal{C}} = \boldsymbol{u}_1 \boldsymbol{u}_1^\top$ and complement projector $\Pi_\perp = \boldsymbol{I} - \Pi_{\mathcal{C}}$. Its* consensus ratio *is $\rho(\boldsymbol{Z}) = \lambda_1/k \in [1/k, 1]$ and its* spectral gap *is $\Delta(\boldsymbol{Z}) = (\lambda_1 - \lambda_2)/k \in [0, 1]$. For each modality, $\boldsymbol{q}^m = \Pi_\perp \boldsymbol{z}^m$ is the* complement *(innovation) of modality m and $r_m = \|\boldsymbol{q}^m\|^2$ its innovation energy; $\mathbf{Q} = \Pi_\perp \boldsymbol{Z} = [\boldsymbol{q}^1, \ldots, \boldsymbol{q}^k]$ collects the complements.*

We first show that the consensus is the optimal common direction with closed-form, spectrally certified weights (Lemmas 1 and 2, Proposition 3), then that discarding the complement is lossy (Theorem 4), how much of it to keep (Theorem 5) and that stability is governed by the spectral gap (Theorem 7). Proofs are in Appendix A.

① **Consensus and its weights.** We begin with the leading singular direction and the weights that produce it.

Lemma 1 (Consensus is the optimal common direction)**.** *For every unit vector $\boldsymbol{u}$, $\sum_m \langle \boldsymbol{u}, \boldsymbol{z}^m \rangle^2 \leq \lambda_1$, with equality iff $\boldsymbol{u} = \pm \boldsymbol{u}_1$ when $\lambda_1 > \lambda_2$; hence $\boldsymbol{u}_1$ minimises the total complement energy $\sum_m \|(\boldsymbol{I} - \boldsymbol{u}\boldsymbol{u}^\top)\boldsymbol{z}^m\|^2$, whose minimum is $k - \lambda_1$, and $1 - \rho(\boldsymbol{Z}) = \frac{1}{k}\sum_m \|\boldsymbol{q}^m\|^2$ is the normalised disagreement among modalities. Under Assumption 1, $\rho(\boldsymbol{Z}) = 1$ iff $\boldsymbol{z}^1 = \cdots = \boldsymbol{z}^k$ (full alignment), and $\rho(\boldsymbol{Z}) = 1/k$ iff the modalities are pairwise orthogonal.*

Lemma 2 (Closed-form consensus weights)**.** *Let $\boldsymbol{v}_1$ be the leading right singular vector of $\boldsymbol{Z}$. Then (i) $\boldsymbol{Z}\boldsymbol{v}_1 = \sigma_1 \boldsymbol{u}_1$, i.e., the consensus vector is the weighted combination $\boldsymbol{c} = \sum_m v_{1,m} \boldsymbol{z}^m$; (ii) $\boldsymbol{v}_1 = \arg\max_{\|\boldsymbol{w}\|=1} \|\boldsymbol{Z}\boldsymbol{w}\|$, unique up to sign when $\lambda_1 > \lambda_2$ (*maximal agreement*); (iii) $v_{1,m} = \langle \boldsymbol{u}_1, \boldsymbol{z}^m \rangle / \sigma_1$ is the cosine agreement $\langle \boldsymbol{u}_1, \boldsymbol{z}^m \rangle$ of modality m with the consensus, normalised by σ_1; (iv) if $\boldsymbol{G}$ is* balanced*, i.e., $\mathbf{D}\boldsymbol{G}\mathbf{D}$ is entrywise non-negative for some sign matrix $\mathbf{D} = \mathrm{diag}(\pm 1)$ (Assumption 1 is the case $\mathbf{D} = \boldsymbol{I}$), then $\mathbf{D}\boldsymbol{v}_1$ can be chosen entrywise non-negative, and is unique and strictly positive when $\mathbf{D}\boldsymbol{G}\mathbf{D}$ is entrywise positive (Perron–Frobenius); (v) $\sum_m v_{1,m} \boldsymbol{q}^m = \mathbf{0}$.*

Remark. Lemma 2 provides a per-instance, agreement-rewarding weighting without the router network of EMOE (Fang et al., 2025), and item (v) shows that consensus fusion cancels the complements exactly. Flipping the sign of a modality's representation leaves every spectral quantity invariant, so what the spectrum can certify is *balance*, i.e., non-negativity up to such a flip.

Proposition 3 (Spectral certificate of balance)**.** *Let $\mathbf{D} = \mathrm{diag}(\mathrm{sign}(\boldsymbol{v}_1))$. If $\mathbf{D}\boldsymbol{G}\mathbf{D}$ has a negative entry $g < 0$, then $\lambda_1 \leq k - 1 - g^2$. Hence $\rho(\boldsymbol{Z}) > 1 - 1/k$ implies that $\boldsymbol{G}$ is balanced, i.e., $\mathbf{D}\boldsymbol{G}\mathbf{D} \geq \mathbf{0}$ entrywise, and then $\mathbf{D}\boldsymbol{v}_1 \geq \mathbf{0}$ by Lemma 2(iv); if in addition Assumption 1 holds, $\mathbf{D} = \boldsymbol{I}$ is admissible and $\boldsymbol{v}_1 \geq \mathbf{0}$. For $k = 3$, $\boldsymbol{G}$ is balanced iff $\boldsymbol{G}_{12}\boldsymbol{G}_{13}\boldsymbol{G}_{23} \geq 0$, and a frustrated triangle forces $\lambda_1 < 2$.*

Remark. The spectral condition constrains the *sign pattern* of $\boldsymbol{G}$, not the raw weights, since $\boldsymbol{Z} = [\boldsymbol{u}, \boldsymbol{u}, -\boldsymbol{u}]$ has $\rho = 1$ and $\boldsymbol{v}_1 = (1, 1, -1)^\top/\sqrt{3}$. The per-pair hinge of Eq. (6) removes such flips, so non-negativity of $\boldsymbol{v}_1$ is a verified property of the trained encoders. At convergence, $\boldsymbol{G} \geq \mathbf{0}$ holds for $100\%/100\%/99.3\%$ and $\rho > 2/3$ for $100\%/99.95\%/95.7\%$ of MOSI/MOSEI/SIMS test utterances (Appendix C.4).

② **Full alignment is predictively incomplete.** We now formalise the intuition that collapsing the modalities onto $\boldsymbol{u}_1$ discards label-relevant information. The argument rests on a latent-variable model.

Assumption 2 (Shared and specific latents)**.** *There exist mutually independent random variables s (shared) and $e^1, \ldots, e^k$ (modality-specific) with $\mathbb{E}[e^m] = 0$, $\mathrm{Var}(e^m) < \infty$, such that $\boldsymbol{x}^m = h^m(s, e^m)$ for injective measurable h^m, and $y = g(s) + \sum_m \beta_m e^m + \eta$ with η zero-mean noise independent of $(s, e^{1:k})$, $\mathbb{E}[g(s)^2] < \infty$.*

Theorem 4 (Predictive incompleteness of full alignment)**.** *Under Assumption 2, let the encoders be fully aligned, i.e., $\boldsymbol{z}^1 = \cdots = \boldsymbol{z}^k =: \bar{\boldsymbol{z}}$ a.s. (Lemma 1), and let $\hat{y} = \varphi(\bar{\boldsymbol{z}})$ for any measurable φ. Then*

$$\mathbb{E}\big[(y - \hat{y})^2\big] \;\geq\; \underbrace{\mathrm{Var}(\eta)}_{\textit{Bayes risk}} \;+\; \textstyle\sum_m \beta_m^2 \mathrm{Var}(e^m), \tag{1}$$

whereas the Bayes-optimal multimodal predictor $\mathbb{E}[y \mid \boldsymbol{x}^1, \ldots, \boldsymbol{x}^k]$ achieves $\mathrm{Var}(\eta)$.

Remark. A fully aligned representation is measurable with respect to $\bigcap_m \sigma(\boldsymbol{x}^m)$, the common information of the modalities. Under Assumption 2, this common information coincides with $\sigma(s)$, so every modality-specific latent is invisible to it, however full alignment is reached. The theorem does *not* say that every consensus-only predictor is lossy: the consensus $\boldsymbol{c} = \boldsymbol{Z}\boldsymbol{v}_1$ is computed after seeing all modalities, so its direction and magnitude can carry specific information, which Eq. (2) quantifies under the oracle direction.

③ **What the consensus and the complement each contribute.** We next quantify *how much* complement to keep and what the consensus buys. To this end, we instantiate Assumption 2 as a linear-Gaussian model in representation space.

Assumption 3 (Linear-Gaussian latent model)**.** *$\boldsymbol{z}^m = s\boldsymbol{a} + e^m \boldsymbol{b}^m + \boldsymbol{\epsilon}^m$ with orthonormal $\boldsymbol{a}, \boldsymbol{b}^1, \ldots, \boldsymbol{b}^k \in \mathbb{R}^d$, independent $s \sim \mathcal{N}(0, v_s)$, $e^m \sim \mathcal{N}(0, v_m)$, $\boldsymbol{\epsilon}^m \sim \mathcal{N}(\mathbf{0}, \varsigma^2 \boldsymbol{I}_d)$, and $y = \beta_0 s + \sum_m \beta_m e^m + \eta$, $\eta \sim \mathcal{N}(0, v_\eta)$; $\gamma_s = v_s/\varsigma^2$ and $\gamma_m = v_m/\varsigma^2$ are the shared and specific signal-to-noise ratios (SNR).*

Theorem 5 (Risk decomposition and optimal complement shrinkage)**.** *Under Assumption 3, with the oracle projector $\Pi_{\mathcal{C}} = \boldsymbol{a}\boldsymbol{a}^\top$ and uniform consensus weights, let $R_{\mathcal{C}}$, R_m and R_{PRISM} be the minimum mean-squared errors of predictors using the consensus $\Pi_{\mathcal{C}}\boldsymbol{Z}$ only, a single modality $\boldsymbol{z}^m$ only, and $(\boldsymbol{c}, \boldsymbol{q}^1, \dots, \boldsymbol{q}^k)$. Then, with $\bar\gamma = \gamma/(1+\gamma)$ and $\bar\gamma_{s,k} = k\gamma_s/(1+k\gamma_s)$, the Bayes predictor of y from the full representation $\boldsymbol{Z}$ is additive over the two parts and free of interaction terms, $\mathbb{E}[y \mid \boldsymbol{Z}] = \beta_0\bar\gamma_{s,k}\,\boldsymbol{a}^\top\boldsymbol{c}/\sqrt{k} + \sum_m \beta_m\bar\gamma_m\,\boldsymbol{b}^{m\top}\boldsymbol{q}^m$, so that $R_{PRISM} = v_\eta + \beta_0^2\frac{v_s}{1+k\gamma_s} + \sum_m \beta_m^2\frac{v_m}{1+\gamma_m}$ is the Bayes risk of $\boldsymbol{Z}$, and*

$$\underbrace{R_{\mathcal{C}} - R_{PRISM} = \textstyle\sum_m \beta_m^2 v_m\bar\gamma_m}_{\textit{specialisation gain}}, \qquad R_m - R_{PRISM} = \underbrace{\beta_0^2 v_s\tfrac{(k-1)\gamma_s}{(1+\gamma_s)(1+k\gamma_s)}}_{\textit{balance gain}} + \underbrace{\textstyle\sum_{m'\neq m}\beta_{m'}^2 v_{m'}\bar\gamma_{m'}}_{\textit{missing-modality deficit}}. \tag{2}$$

Moreover, the Bayes-optimal predictor uses each complement only through the shrunk score along $\boldsymbol{b}^m$,

$$\mathbb{E}[e^m \mid \boldsymbol{q}^m] = \alpha_m^\star\boldsymbol{b}^{m\top}\boldsymbol{q}^m, \quad \mathbb{E}[e^m\boldsymbol{b}^m \mid \boldsymbol{q}^m] = \alpha_m^\star\boldsymbol{b}^m\boldsymbol{b}^{m\top}\boldsymbol{q}^m, \quad \alpha_m^\star = \tfrac{\gamma_m}{1+\gamma_m} = \tfrac{1}{1+\kappa_m}, \tag{3}$$

and, with $\kappa_m = \gamma_m^{-1}$, for every $\kappa_m \geq 0$ the whole-complement *edit $\alpha_m^\star\boldsymbol{q}^m$ is the unique solution $\boldsymbol{q}^m + \boldsymbol{\delta}^\star$ of the minimum-norm editing problem $\min_{\boldsymbol{\delta}\in\mathrm{range}(\Pi_\perp)} \frac{1}{2}\|\boldsymbol{\delta}\|^2 + \frac{\kappa_m}{2}\|\boldsymbol{q}^m + \boldsymbol{\delta}\|^2$.*

Remark. That $(\boldsymbol{c}, \boldsymbol{q}^{1:k})$ lose no information is immediate, since they reconstruct $\boldsymbol{Z}$. The content of Theorem 5 is the *form* of the Bayes predictor: a consensus score plus k individually shrunk complement scores. Three qualifications matter. (i) What is Bayes-optimal is the shrinkage of the *scalar* score $\boldsymbol{b}^{m\top}\boldsymbol{q}^m$. (ii) PRISM does not know $\boldsymbol{b}^m$ and gates the whole complement, the minimum-norm edit, which reproduces the Bayes term whenever the head is linear along $\boldsymbol{b}^m$, so the projection is delegated to the head; for the non-linear heads of Section 4.2 the gate multiplies the score, $g_m(\alpha\boldsymbol{q}) \neq \alpha g_m(\boldsymbol{q})$, and the properties of the edit below describe this linear-head regime. (iii) A *constant* positive gate is absorbed by a freely learned head, so only the instance-dependent part of α_m can change the risk, which Proposition 6 quantifies. The *balance gain* stems from the consensus raising the SNR of the shared factor from γ_s to $k\gamma_s$. As κ_m is unknown and instance dependent, PRISM estimates it with a small *certificate network* (Section 4.2). Any non-negative estimate keeps the edit $T_{\boldsymbol{\alpha}}(\boldsymbol{Z}) = \Pi_{\mathcal{C}}\boldsymbol{Z} + \Pi_\perp\boldsymbol{Z}\,\mathrm{diag}(\boldsymbol{\alpha})$ consensus-preserving, non-expansive and consensus-monotone (Proposition 9, Appendix A.8).

Proposition 6 (Value of instance-dependent gates)**.** *If the specific SNR varies across instances, $\gamma_{m,i} = v_{m,i}/\varsigma^2$ with Bayes gate $\alpha_{m,i} = \gamma_{m,i}/(1+\gamma_{m,i})$, any constant gate $\bar\alpha_m$ incurs the additional risk $\beta_m^2\mathbb{E}_i[(\alpha_{m,i} - \bar\alpha_m)^2(v_{m,i} + \varsigma^2)]$.*

Constant gates are optimal iff the specific SNR is homogeneous; the converged gates spread by 0.02–0.06 around means of 0.15–0.31 (Appendix C.4), so the predicted gain of instance-dependent gates is small, as the ablation confirms.

④ **Stability and the spectral gap.** The decomposition is only useful if $\boldsymbol{u}_1$ is a stable function of the representations. The next result bounds its sensitivity to perturbations of any modality, e.g., feature noise or a corrupted stream.

Theorem 7 (Davis–Kahan stability of the consensus subspace)**.** *Let $\tilde{\boldsymbol{Z}} = \boldsymbol{Z} + \boldsymbol{E}$ and let $\tilde{\boldsymbol{u}}_1$ be its leading left singular vector. If $\lambda_1 > \lambda_2$, then $\sin\angle(\boldsymbol{u}_1, \tilde{\boldsymbol{u}}_1) = \|\tilde\Pi_{\mathcal{C}} - \Pi_{\mathcal{C}}\|_2 \leq \frac{2(2\sigma_1\|\boldsymbol{E}\|_2 + \|\boldsymbol{E}\|_2^2)}{k\,\Delta(\boldsymbol{Z})}$, and, for fixed gates $\boldsymbol{\alpha} \in [0,1]^k$, $\|T_{\boldsymbol{\alpha}}(\tilde{\boldsymbol{Z}}) - T_{\boldsymbol{\alpha}}(\boldsymbol{Z})\|_F \leq \|\boldsymbol{E}\|_F + \sqrt{k}\,\sin\angle(\boldsymbol{u}_1, \tilde{\boldsymbol{u}}_1)$.*

Corollary 8 (Plug-in consensus direction and data-dependent weights)**.** *Under Assumption 3, applying the additive Bayes predictor of Theorem 5 to the decomposition computed from the empirical $\boldsymbol{u}_1$ of $\boldsymbol{Z}$ instead of the oracle $\boldsymbol{a}$ costs at most $2\big(\beta_0^2/k + \sum_m \beta_m^2\big)\,\mathbb{E}\big[\|\boldsymbol{Z}\|_F^2\sin^2\angle(\boldsymbol{u}_1, \boldsymbol{a})\big]$ over R_{PRISM}, and replacing the uniform consensus weights of Theorem 5 by the data-dependent $\boldsymbol{v}_1$ costs at most $\frac{\beta_0^2}{k}\,\mathbb{E}\big[\lambda_1\|\boldsymbol{v}_1 - \mathbf{1}/\sqrt{k}\|^2\big]$, where $\|\boldsymbol{v}_1 - \mathbf{1}/\sqrt{k}\|^2 \leq \min\{2,\, 2(\lambda_1 - \mathbf{1}^\top\boldsymbol{G}\mathbf{1}/k)/(\lambda_1 - \lambda_2)\}$; the two substitutions together cost at most twice the sum. Theorem 7 applied to $\boldsymbol{Z} = s\boldsymbol{a}\mathbf{1}^\top + \boldsymbol{E}$ gives $\sin\angle(\boldsymbol{u}_1, \boldsymbol{a}) \leq \min\{1,\, 2(2|s|\sqrt{k}\|\boldsymbol{E}\|_2 + \|\boldsymbol{E}\|_2^2)/(ks^2)\}$, where the truncation keeps the expectation finite under the Gaussian prior on s.*

Theorem 7 says that the quantity to control is the *gap*, not the leading singular value. We therefore regularise it directly:

$$\mathcal{L}_{\mathrm{gap}} = 1 - \Delta(\boldsymbol{Z}) = 1 - \tfrac{1}{k}(\lambda_1 - \lambda_2) = \tfrac{1}{k}\big(\|\mathbf{Q}\|_F^2 + \lambda_2\big) \in [0,1], \qquad \|\mathbf{Q}\|_F^2 = \textstyle\sum_m r_m = k - \lambda_1. \tag{4}$$

The second form shows that $\mathcal{L}_{\text{gap}}$ does penalise complement energy and in isolation is minimised only at $\lambda_1 = k$, like any full-alignment objective. What differs is the gradient, $-\frac{2}{k}\boldsymbol{Z}(\boldsymbol{v}_1\boldsymbol{v}_1^\top - \boldsymbol{v}_2\boldsymbol{v}_2^\top)$, which acts only in the plane of the top two right singular vectors, whereas $\mathcal{L}_{\text{align}}$ pushes on every σ_j and the log-volume is dominated by the smallest singular value, as Appendix A.10 details. The complement is kept open by the task loss and the auxiliary heads, so the equilibrium is interior and dataset dependent, at $\rho = 0.97$, 0.95 and 0.85. Its position is set by w_g, which is validated like any loss weight: what PRISM removes is the router network and the per-modality gate parameters, not this macro-level trade-off.

4.2 Consensus–Complement Fusion with Certificate Gates

Building upon the theoretical insights, we propose PRISM, which fuses the modalities through the consensus–complement decomposition of every instance. Figure 2 depicts the overall framework.

Encoders and decomposition. Language tokens are encoded with fine-tuned BERT (Devlin et al., 2019), and acoustic and visual sequences with light Transformer encoders. Attention pooling, a two-layer MLP and ℓ_2-normalisation give $\boldsymbol{z}^m$. We eigendecompose $\boldsymbol{G} = \boldsymbol{Z}^\top\boldsymbol{Z} \in \mathbb{R}^{3\times3}$ without gradient and re-express $\lambda_j = \boldsymbol{v}_j^\top \boldsymbol{G}\boldsymbol{v}_j$ with the eigenvectors held fixed, which by the Hellmann–Feynman identity gives the exact gradient of λ_j without the ill-conditioned terms of eigenvector back-propagation. The sign of $\boldsymbol{v}_1$ is fixed by $\sum_m v_{1,m} > 0$, and $\boldsymbol{c} = \boldsymbol{Z}\boldsymbol{v}_1$, $\boldsymbol{u}_1 = \boldsymbol{c}/\|\boldsymbol{c}\|$, $\boldsymbol{q}^m = \boldsymbol{z}^m - \langle \boldsymbol{u}_1, \boldsymbol{z}^m\rangle \boldsymbol{u}_1$ and $r_m = \|\boldsymbol{q}^m\|^2$ follow.

Certificate gates and additive prediction. Unimodal heads predict $\hat{y}_m$ from $\boldsymbol{z}^m$, and a consensus head predicts $\hat{y}_c$ from $\boldsymbol{c}$. The certificate network ψ maps the statistics $(\lambda_{1:k}, \boldsymbol{v}_1, \boldsymbol{r}, \hat{y}_{1:k}, \hat{y}_c)$, treated as constants, to $\boldsymbol{\kappa} = \text{softplus}(\psi(\cdot)+b) \in \mathbb{R}^k_+$, where the bias b sets the prior amount of complement. The gates are $\alpha_m = 1/(1+\kappa_m)$, cf. Eq. (3). Following the Bayes predictor of Theorem 5, the prediction is additive over the two parts, with two-layer heads g_c and g_m. The gate multiplies the score rather than the complement, so that $g_m(\boldsymbol{q}^m)$ is shared with the ungated complement-only prediction $\hat{y}_\perp$ used for auxiliary supervision:

$$\hat{y} = g_c(\boldsymbol{c}) + \textstyle\sum_m \alpha_m\, g_m(\boldsymbol{q}^m), \qquad \hat{y}_\perp = \textstyle\sum_m g_m(\boldsymbol{q}^m). \tag{5}$$

Objective. With the ℓ_1 loss ℓ used by MSA benchmarks, the overall objective combines the task loss with auxiliary supervision and the spectral regularisers:

$$\mathcal{L} = \ell(\hat{y}, y) + w_u \tfrac{1}{k}\textstyle\sum_m \ell(\hat{y}_m, y) + w_\perp \ell(\hat{y}_\perp, y) + w_c \ell(\hat{y}_c, y) + w_g \mathcal{L}_{\text{gap}} + w_{\text{pos}} \textstyle\sum_{m<m'} [-\boldsymbol{G}_{mm'}]_+. \tag{6}$$

The unimodal terms keep every modality predictive, which counters the specialisation problem of Fang et al. (2025), the complement term keeps the innovation label-relevant, $\mathcal{L}_{\text{gap}}$ implements Theorem 7, and the hinge enforces Assumption 1 and breaks the sign symmetry (Appendix B.5); Appendix B.6 gives the algorithm and code.

4.3 Further Analysis

Connecting to full-alignment objectives, decoupling and dynamic experts. The singular-value objective $\mathcal{L}_{\text{align}}$ maximises σ_1 relative to $\sum_j \sigma_j$ and the GRAM objective minimises the log-volume $\log\det \boldsymbol{G}$ (Cicchetti et al., 2025), which we instantiate as an unconstrained per-instance penalty without its contrastive normalisation, whereas $\mathcal{L}_{\text{gap}}$ maximises $\lambda_1 - \lambda_2$. All three share the same isolated optimum and differ only in the gradient they apply away from it, hence in the equilibrium they reach against the task loss (Section 5.3). Decoupling methods approximate orthogonality with penalties (Hazarika et al., 2020; Li et al., 2023b), whereas PRISM obtains an *exact* split from a 3×3 eigenproblem per instance, at about 5% of a training step, and its weights are the router that dynamic experts (Fang et al., 2025) learn (Lemma 2).

5 Experiments

5.1 Experimental Setups

Datasets. CMU-MOSI (Zadeh et al., 2016) has 2,199 English utterances, CMU-MOSEI (Zadeh et al., 2018b) 22,856 and CH-SIMS (Yu et al., 2020) 2,281 Chinese utterances. We use the MMSA

Table 1: MOSI and MOSEI use word-aligned features and BERT with Acc-2 on negative versus positive. SIMS uses unaligned features and Chinese BERT. Unmarked rows are published results (Zhang et al., 2023a; Mao et al., 2022). Rows marked † are re-run under our protocol as mean±std over five seeds and ten for PRISM on MOSI and SIMS, and rows marked u use unaligned MOSI features as released, see Appendices C.2 and C.9, which also report F1. **Bold** marks the best mean under the identical protocol and Δ the gain over the EMOE.

Method	CMU-MOSI			CMU-MOSEI			CH-SIMS				
	Acc-7↑	Acc-2↑	MAE↓	Acc-7↑	Acc-2↑	MAE↓	Acc-2↑	Acc-3↑	Acc-5↑	MAE↓	Corr↑
MISA	42.3	83.4	0.783	52.2	85.5	0.555	76.5	-	-	0.447	0.563
MAG-BERT	43.6	84.4	0.727	52.7	84.8	0.543	74.4	-	-	0.492	0.399
Self-MM	45.8	84.8	0.712	53.5	85.0	0.529	80.0	65.5	41.5	0.425	0.595
MMIM	46.7	86.1	0.700	54.2	86.0	0.526	-	-	-	-	-
DMD	46.2	83.2	0.721	52.4	84.8	0.546	-	-	-	-	-
ALMT	49.4	86.4	0.683	54.3	86.8	0.526	81.2	68.9	45.7	0.404	0.619
EMOE	47.7	85.4	0.710	54.1	85.3	0.536	-	-	-	-	-
Self-MM$^{\dagger u}$	46.7±0.9	84.4±0.4	0.711±0.011	53.8±0.4	84.0±1.1	0.535±0.004	78.0±1.0	64.8±1.0	42.1±1.8	0.421±0.013	0.587±0.017
ALMT$^{\dagger u}$	46.2±1.1	83.7±0.5	0.733±0.006	53.0±0.6	85.1±0.4	0.541±0.007	78.1±0.7	64.4±0.6	40.6±1.0	0.422±0.007	0.584±0.008
DMD†	46.3±0.4	83.9±0.5	0.717±0.008	52.6±0.2	85.0±0.3	0.540±0.002	77.3±1.5	64.2±0.7	42.2±2.3	0.424±0.011	0.575±0.011
EMOE†	44.4±1.5	84.7±0.3	0.735±0.020	53.4±0.4	85.0±0.2	0.541±0.007	77.2±0.9	63.7±1.6	42.2±2.3	0.425±0.010	0.570±0.015
PRISM†	**47.8**±0.7	**85.5**±0.3	**0.701**±0.008	**54.2**±0.3	**85.2**±0.2	**0.531**±0.003	**80.4**±1.4	**67.9**±1.6	**44.2**±2.2	**0.401**±0.008	**0.627**±0.011
Δ vs. EMOE†	(+3.4)	(+0.8)	(−0.034)	(+0.8)	(+0.2)	(−0.010)	(+3.2)	(+4.2)	(+2.0)	(−0.024)	(+0.057)

features (Mao et al., 2022), word-aligned for MOSI and MOSEI and unaligned for SIMS, following EMOE and ALMT (Appendix B.1).

Baselines and evaluation metrics. We compare with TFN (Zadeh et al., 2017), MulT (Tsai et al., 2019a), MISA (Hazarika et al., 2020), MAG-BERT (Rahman et al., 2020), Self-MM (Yu et al., 2021), MMIM (Han et al., 2021b), DMD (Li et al., 2023b), ALMT (Zhang et al., 2023a) and EMOE (Fang et al., 2025), and with MFN (Zadeh et al., 2018a), Graph-MFN (Zadeh et al., 2018b), CENet (Wang et al., 2023b) and TETFN (Wang et al., 2023a) in Appendix C.2. Since published MSA numbers are typically single-run peaks, we re-run the strongest baselines with their official code under our features, seeds and model selection. We report Acc-7, Acc-2 and MAE on MOSI and MOSEI and Acc-2, Acc-3, Acc-5, MAE and correlation on SIMS, with F1 and the metric definitions in Appendices C.2 and B.3. PRISM numbers are means over ten seeds on MOSI and SIMS and five on MOSEI, selected on validation MAE.

Model architecture and hyperparameters. The language encoder is `bert-base-uncased` or `bert-base-chinese`. The acoustic and visual streams are standardised, projected to $d = 128$ and encoded by two-layer Transformers with attention pooling, the heads are two-layer MLPs and the certificate network has 32 hidden units. We train with AdamW (Loshchilov & Hutter, 2019), the ℓ_1 loss and early stopping on validation MAE, set $w_u = w_c = 1$, $w_\perp = 0.5$, $w_{\text{pos}} = 1$ and $b = 1$, and choose w_g on validation MAE from $\{0, 0.05, 0.1, 0.3, 0.5, 1, 2\}$, which gives 0.3 on all three datasets.

5.2 Main Results

Table 1 compares PRISM with the published results and with the strongest baselines re-run under our features, seeds and model selection, and Appendix C.3 tests every difference at the seed level. On CH-SIMS, PRISM outperforms the re-run EMOE by 3.2 Acc-2, 4.2 Acc-3, 0.024 MAE and 5.7 correlation points, with gains over all re-run baselines significant at $p < 0.05$ or better on Acc-2, Acc-3, MAE and correlation. On CMU-MOSEI, it achieves the best Acc-7, Acc-2 and MAE under the identical protocol, with the MAE gain significant against three of the four re-run baselines. On CMU-MOSI, it improves on the re-run EMOE by 3.4 Acc-7 and 0.034 MAE and on the strongest re-run baseline, Self-MM, by 1.1 Acc-7. PRISM thus achieves state-of-the-art performance on the three benchmarks under the identical protocol, with the largest gains on CH-SIMS, whose annotations are the least language-dominated, as Theorem 4 predicts.

5.3 Further Empirical Analysis

Ablation study. Figure 3 cancels the core designs of PRISM in turn and Table 3 reports all variants and metrics. **Firstly**, the gap regulariser is essential, as removing it costs 2.1 Acc-7 on MOSI and 1.5 Acc-3 on SIMS, since the consensus direction is ill defined without it, and the singular-

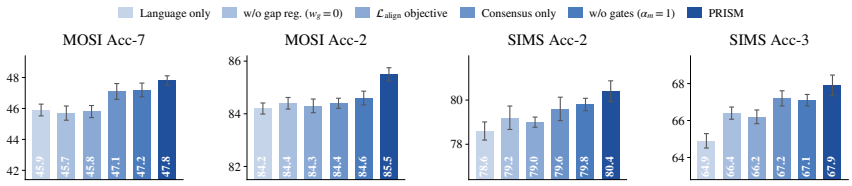
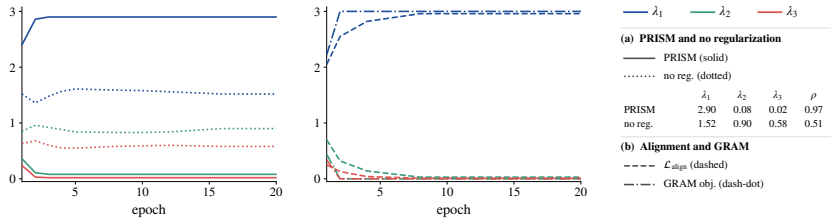


Figure 3: **Ablation on CMU-MOSI and CH-SIMS.** The gap regulariser, the complement and the certificate gates are cancelled in turn, $\mathcal{L}_{\text{gap}}$ is replaced by $\mathcal{L}_{\text{align}}$, and all variants are compared with the language-only head. Bars show the mean and standard error over seeds, and Table 3 reports all metrics.



	λ_1	λ_2	λ_3	ρ
PRISM	2.90	0.08	0.02	0.97
no reg.	1.52	0.90	0.58	0.51

Figure 4: **Eigenvalue analysis.** Eigenvalues $\lambda_{1:3}$ of $\boldsymbol{G}$ during training on CMU-MOSI under $\mathcal{L}_{\text{gap}}$ and without a regulariser in (a), and under the singular-value objective $\mathcal{L}_{\text{align}}$ and the GRAM objective in (b). Curves show the mean over seeds.

value objective is no better, at 2.0 Acc-7 and 1.7 Acc-3. **Secondly**, the complement matters, as the consensus alone loses 0.7 Acc-7 and 1.1 Acc-2 on MOSI and 0.8 Acc-2 and 0.7 Acc-3 on SIMS, the empirical face of the specialisation gain in Eq. (2). **Thirdly**, the certificate gates matter, as the ungated complement loses 0.6 Acc-7 and 0.9 Acc-2 on MOSI and 0.6 Acc-2 and 0.8 Acc-3 on SIMS, in line with Proposition 6. **Fourthly**, fusion is necessary, as the language-only head trails PRISM by 1.9 Acc-7 on MOSI and by 1.8 Acc-2 and 3.0 Acc-3 on SIMS. **Eigenvalue analysis.** Figure 4 tracks the spectrum of $\boldsymbol{G}$ during training on MOSI. Without a spectral regulariser the eigenvalues stay near 1.5, 0.9 and 0.6, the singular-value objective drives λ_1 to 2.96, GRAM reaches 3, 0 and 0 within two epochs, whereas under $\mathcal{L}_{\text{gap}}$ λ_1 rises to 2.9 and $\lambda_2 + \lambda_3$ remains a small but non-zero residual set by the task loss. The equilibrium differs per dataset, with $\bar{\rho} = 0.97$, 0.95 and 0.85 on MOSI, MOSEI and SIMS, consistent with the largest fusion gains on SIMS, and Appendix C.4 reports the per-modality loadings, innovation energies and gates.

Alignment–distinctness continuum. The gap weight moves the representations between independent modalities at $\rho = 1/k$ and full alignment at $\rho = 1$, and Table 9 in Appendix C.6 reports accuracy against the consensus ratio at convergence. On MOSI, Acc-7 rises from 45.7 at $w_g = 0$ to 47.8 at $\rho = 0.97$ and then declines slightly, to 47.2 when over-aligned, so alignment acts as consensus denoising, the balance gain of Theorem 5. On SIMS, Acc-2 peaks at $\rho \approx 0.85$ with 80.4 and falls by 1.8 points and 0.025 MAE when over-aligned, the empirical face of Theorem 4. Neither dataset wants full alignment, and no single hand-tuned balance serves both.

6 Conclusion

In this paper, we propose to fuse the modalities of an utterance through an exact orthogonal decomposition of its instance matrix into a consensus subspace and its complement, and to retain the complement through certificate-driven shrinkage. The proposed method is grounded on the theoretical insight that full alignment is predictively incomplete, with an excess risk of at least the label variance explained by modality-specific latents, whereas under a linear-Gaussian latent model the Bayes-optimal predictor adds a consensus score to individually shrunk complement scores. Closed-form agreement weights, a minimum-mean-square-error gate and a Davis–Kahan gap regulariser are then derived from this insight. A series of empirical results on CMU-MOSI, CMU-MOSEI and CH-SIMS demonstrates the effectiveness of our PRISM framework and further showcases its rational design. In the future, we plan to extend the consensus to rank-r subspaces, to infer with missing modalities, and to bring the consensus–complement view to retrieval-oriented learning.

AI USE STATEMENT

We have not used generative AI tools for generating research ideas, theoretical results or experimental results. All AI-assisted text was reviewed and edited by the authors, all proofs were checked by hand, and all AI-assisted code was verified and tested for correctness by the authors. We take responsibility for the final content of this work.

ETHICS STATEMENT

This work uses publicly available, widely adopted MSA benchmarks (CMU-MOSI, CMU-MOSEI, CH-SIMS) released for research purposes by their authors; we use the pre-extracted features distributed with the MMSA benchmark and do not process raw videos or identify individuals. Sentiment models may reflect biases of the collected data and should not be deployed for surveillance or profiling. We do not foresee further ethical concerns.

REPRODUCIBILITY STATEMENT

All theoretical results are stated with their assumptions in Section 4 and proved in Appendix A. Datasets, features, metrics, architectures and every hyper-parameter are listed in Appendix B, the algorithm flow and pseudo code in Appendix B.6, and the re-run protocol of the baselines in Appendix C.2. Source code, configuration files and the scripts that generate every table and figure of this paper will be released, as detailed in Appendix D; limitations are discussed in Appendix E.

REFERENCES

Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *ICML*, 2013.

Giordano Cicchetti, Eleonora Grassucci, Luigi Sigillo, and Danilo Comminiello. Gramian multimodal representation learning and alignment. In *ICLR*, 2025.

Chandler Davis and William M. Kahan. The rotation of eigenvectors by a perturbation. III. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pp. 4171–4186, 2019.

Chenzhuang Du, Jiaye Teng, Tingle Li, Yichen Liu, Tianyuan Yuan, Yue Wang, Yang Yuan, and Hang Zhao. On uni-modal feature learning in supervised multi-modal learning. In *ICML*, 2023.

Benoit Dufumier, Javiera Castillo-Navarro, Devis Tuia, and Jean-Philippe Thiran. What to align in multimodal contrastive learning? In *ICLR*, 2025.

Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

Yunfeng Fan, Wenchao Xu, Haozhao Wang, Junxiao Wang, and Song Guo. PMR: Prototypical modal rebalance for multimodal learning. In *CVPR*, 2023.

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

Marco Federici, Anjan Dutta, Patrick Forré, Nate Kushman, and Zeynep Akata. Learning robust representations via multi-view information bottleneck. In *ICLR*, 2020.

Xinyu Feng, Yuming Lin, Lihua He, You Li, Liang Chang, and Ya Zhou. Knowledge-guided dynamic modality attention fusion framework for multimodal sentiment analysis. In *Findings of EMNLP*, 2024.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. ImageBind: One embedding space to bind them all. In *CVPR*, 2023.

Wei Han, Hui Chen, Alexander Gelbukh, Amir Zadeh, Louis-Philippe Morency, and Soujanya Poria. Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis. In *ICMI*, 2021a.

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

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

Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.

Guimin Hu, Ting-En Lin, Yi Zhao, Guangming Lu, Yuchuan Wu, and Yongbin Li. UniMSE: Towards unified multimodal sentiment analysis and emotion recognition. In *EMNLP*, 2022.

Cong Hua, Qianqian Xu, Shilong Bao, Zhiyong Yang, and Qingming Huang. ReconBoost: Boosting can achieve modality reconcilement. In *ICML*, 2024.

Yu Huang, Chenzhuang Du, Zihui Xue, Xuanyao Chen, Hang Zhao, and Longbo Huang. What makes multi-modal learning better than single (provably). In *NeurIPS*, 2021.

Yu Huang, Junyang Lin, Chang Zhou, Hongxia Yang, and Longbo Huang. Modality competition: What makes joint training of multi-modal network fail in deep learning? (Provably). In *ICML*, 2022.

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

Willard James and Charles Stein. Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pp. 361–379, 1961.

Thomas Kailath, Ali H. Sayed, and Babak Hassibi. *Linear Estimation*. Prentice Hall, 2000.

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

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

Paul Pu Liang, Yun Cheng, Xiang Fan, Chun Kai Ling, Suzanne Nie, Richard Chen, Zihao Deng, Nicholas Allen, Randy Auerbach, Faisal Mahmood, Ruslan Salakhutdinov, and Louis-Philippe Morency. Quantifying & modeling multimodal interactions: An information decomposition framework. In *NeurIPS*, 2023a.

Paul Pu Liang, Zihao Deng, Martin Q. Ma, James Zou, Louis-Philippe Morency, and Ruslan Salakhutdinov. Factorized contrastive learning: Going beyond multi-view redundancy. In *NeurIPS*, 2023b.

Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Foundations & trends in multimodal machine learning: Principles, challenges, and open questions. *ACM Computing Surveys*, 56(10): 1–42, 2024.

Yangguang Lin, Quan Fang, Yufei Li, Jiachen Sun, Junyu Gao, and Jitao Sang. HulluEdit: Single-pass evidence-consistent subspace editing for mitigating hallucinations in large vision-language models. In *CVPR*, 2026.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Efficient low-rank multimodal fusion with modality-specific factors. In *ACL*, 2018.

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

Zhou Lu. A theory of multimodal learning. In *NeurIPS*, 2023.

Huisheng Mao, Ziqi Yuan, Hua Xu, Wenmeng Yu, Yihe Liu, and Kai Gao. M-sena: An integrated platform for multimodal sentiment analysis. In *ACL (Demo)*, 2022.

Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *CVPR*, pp. 8238–8247, 2022.

Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabás Póczos. Found in translation: Learning robust joint representations by cyclic translations between modalities. In *AAAI*, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, Amir Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. Integrating multimodal information in large pretrained transformers. In *ACL*, 2020.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *ICLR*, 2017.

Karthik Sridharan and Sham M. Kakade. An information theoretic framework for multi-view learning. In *COLT*, 2008.

Hao Sun, Hongyi Wang, Jiaqing Liu, Yen-Wei Chen, and Lanfen Lin. CubeMLP: An MLP-based model for multimodal sentiment analysis and depression estimation. In *ACM Multimedia*, 2022.

Licai Sun, Zheng Lian, Bin Liu, and Jianhua Tao. Efficient multimodal transformer with dual-level feature restoration for robust multimodal sentiment analysis. *IEEE Transactions on Affective Computing*, 15(1):309–325, 2024.

Zhongkai Sun, Prathusha Sarma, William Sethares, and Yingyu Liang. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. In *AAAI*, 2020.

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

Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. Learning factorized multimodal representations. In *ICLR*, 2019b.

Yao-Hung Hubert Tsai, Yue Wu, Ruslan Salakhutdinov, and Louis-Philippe Morency. Self-supervised learning from a multi-view perspective. In *ICLR*, 2021.

Di Wang, Xutong Guo, Yumin Tian, Jinhui Liu, LiHuo He, and Xuemei Luo. TETFN: A text enhanced transformer fusion network for multimodal sentiment analysis. *Pattern Recognition*, 136:109259, 2023a.

Di Wang, Shuai Liu, Quan Wang, Yumin Tian, Lihuo He, and Xinbo Gao. Cross-modal enhancement network for multimodal sentiment analysis. *IEEE Transactions on Multimedia*, 25:4909–4921, 2023b.

Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *CVPR*, 2020.

Yake Wei and Di Hu. MMPareto: Boosting multimodal learning with innocent unimodal assistance. In *ICML*, 2024.

Nan Wu, Stanisław Jastrzębski, Kyunghyun Cho, and Krzysztof J. Geras. Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks. In *ICML*, 2022.

Gongli Xi, Ye Tian, Mengyu Yang, Huahui Yi, Liang Lin, Xiaoshuai Hao, Kun Wang, and Wendong Wang. Large vision–language models get lost in attention. In *ICML*, 2026.

Zihui Xue and Radu Marculescu. Dynamic multimodal fusion. In *CVPR Workshops*, 2023.

Dingkang Yang, Shuai Huang, Haopeng Kuang, Yangtao Du, and Lihua Zhang. Disentangled representation learning for multimodal emotion recognition. In *ACM Multimedia*, 2022.

Jiuding Yang, Yakun Yu, Di Niu, Weidong Guo, and Yu Xu. ConFEDE: Contrastive feature decomposition for multimodal sentiment analysis. In *ACL*, 2023.

Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *ACL*, pp. 3718–3727, 2020.

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

Yi Yu, Tengyao Wang, and Richard J. Samworth. A useful variant of the davis–kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015.

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.

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

Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Memory fusion network for multi-view sequential learning. In *AAAI*, 2018a.

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 *ACL*, pp. 2236–2246, 2018b.

Haoyu Zhang, Yu Wang, Guanghao Yin, Kejun Liu, Yuanyuan Liu, and Tianshu Yu. Learning language-guided adaptive hyper-modality representation for multimodal sentiment analysis. In *EMNLP*, 2023a.

Qingyang Zhang, Haitao Wu, Changqing Zhang, Qinghua Hu, Huazhu Fu, Joey Tianyi Zhou, and Xi Peng. Provable dynamic fusion for low-quality multimodal data. In *ICML*, 2023b.

Xiaohui Zhang, Jaehong Yoon, Mohit Bansal, and Huaxiu Yao. Multimodal representation learning by alternating unimodal adaptation. In *CVPR*, 2024.

Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, HongFa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, Wancai Zhang, Zhifeng Li, Wei Liu, and Li Yuan. LanguageBind: Extending video-language pretraining to N-modality by language-based semantic alignment. In *ICLR*, 2024.

Appendix

A Theoretical Analysis 14
A.1 Proof of Lemma 1 14
A.2 Proof of Lemma 2 14
A.3 Proof of Proposition 3 (spectral certificate) 15
A.4 Proof of Proposition 6 (instance-dependent gates) 15
A.5 Proof of Corollary 8 (plug-in direction) 15
A.6 Proof of Theorem 4 16
A.7 Proof of Theorem 5 16
A.8 Properties of the Consensus-preserving Edit 17
A.9 Proof of Theorem 7 17
A.10 Gradient Analysis of the Gap Regulariser 18
B Implementation Details 18
B.1 Datasets and Features 18
B.2 Baselines 18
B.3 Evaluation Metrics 19
B.4 Hyperparameter Settings 19
B.5 Model Architecture and Sign Symmetry 19
B.6 Algorithm and Pseudo Code 20
C Additional Results 21
C.1 Full Ablation Study 21
C.2 Baselines Re-run under the Identical Protocol 21
C.3 Statistical Significance of the Gains 22
C.4 Spectral Statistics 23
C.5 Robustness 24
C.6 Sensitivity Analysis 24
C.7 Interpretation of the Spectral Statistics 25
C.8 Per-utterance Visualisation 26
C.9 Unaligned Features 27
D Reproducibility 27
E Limitations 27

A Theoretical Analysis

A.1 Proof of Lemma 1

Recall. For every unit vector $\boldsymbol{u}$, $\sum_m \langle \boldsymbol{u}, \boldsymbol{z}^m \rangle^2 \le \lambda_1$ with equality iff $\boldsymbol{u} = \pm \boldsymbol{u}_1$ when $\lambda_1 > \lambda_2$; $\boldsymbol{u}_1$ minimises the total complement energy, whose minimum is $k - \lambda_1$; under Assumption 1, $\rho = 1$ iff the columns of $\boldsymbol{Z}$ coincide and $\rho = 1/k$ iff they are pairwise orthogonal.

Proof. For a unit vector $\boldsymbol{u}$, $\sum_m \langle \boldsymbol{u}, \boldsymbol{z}^m \rangle^2 = \boldsymbol{u}^\top \boldsymbol{Z}\boldsymbol{Z}^\top \boldsymbol{u}$ is the Rayleigh quotient of the positive semidefinite matrix $\boldsymbol{Z}\boldsymbol{Z}^\top$, whose largest eigenvalue is λ_1 with eigenvector $\boldsymbol{u}_1$; the quotient is maximised exactly at the top eigenvectors, i.e., at $\pm \boldsymbol{u}_1$ when $\lambda_1 > \lambda_2$. By the Pythagorean identity $\|(\boldsymbol{I} - \boldsymbol{u}\boldsymbol{u}^\top)\boldsymbol{z}^m\|^2 = \|\boldsymbol{z}^m\|^2 - \langle \boldsymbol{u}, \boldsymbol{z}^m \rangle^2 = 1 - \langle \boldsymbol{u}, \boldsymbol{z}^m \rangle^2$, so the total complement energy is $k - \sum_m \langle \boldsymbol{u}, \boldsymbol{z}^m \rangle^2$, minimised at $\boldsymbol{u}_1$ with value $k - \lambda_1 = \sum_{j \ge 2} \lambda_j$ because $\sum_j \lambda_j = \operatorname{tr} \boldsymbol{G} = k$. Dividing by k gives $1 - \rho = \frac{1}{k} \sum_m \|\boldsymbol{q}^m\|^2$. If $\rho = 1$ then $\lambda_j = 0$ for $j \ge 2$, so $\operatorname{rank} \boldsymbol{G} = 1$ and $\boldsymbol{G} = \boldsymbol{v}\boldsymbol{v}^\top$ with $v_m^2 = \boldsymbol{G}_{mm} = 1$, i.e., $\boldsymbol{G}_{mm'} = \pm 1$; under Assumption 1 all entries equal 1, and $\langle \boldsymbol{z}^m, \boldsymbol{z}^{m'} \rangle = 1$ for unit vectors forces $\boldsymbol{z}^m = \boldsymbol{z}^{m'}$. Conversely equal columns give $\boldsymbol{G} = \boldsymbol{1}\boldsymbol{1}^\top$ with $\lambda_1 = k$. Finally $\rho = 1/k$ iff $\lambda_1 = 1$; since λ_1 is the largest of k non-negative numbers summing to k, this forces $\lambda_j = 1$ for all j, i.e., $\boldsymbol{G} = \boldsymbol{I}$, i.e., pairwise orthogonality. □

A.2 Proof of Lemma 2

Proof. (i) From the SVD, $\boldsymbol{Z}\boldsymbol{v}_1 = \sum_j \sigma_j \boldsymbol{u}_j \boldsymbol{v}_j^\top \boldsymbol{v}_1 = \sigma_1 \boldsymbol{u}_1$. (ii) $\|\boldsymbol{Z}\boldsymbol{w}\|^2 = \boldsymbol{w}^\top \boldsymbol{G}\boldsymbol{w}$ is maximised over unit $\boldsymbol{w}$ at the top eigenvector of $\boldsymbol{G}$, which is $\boldsymbol{v}_1$. (iii) From $\boldsymbol{Z}^\top \boldsymbol{u}_1 = \sigma_1 \boldsymbol{v}_1$, the m-th entry reads $\langle \boldsymbol{z}^m, \boldsymbol{u}_1 \rangle = \sigma_1 v_{1,m}$. (iv) $\tilde{\boldsymbol{G}} = \mathbf{D}\boldsymbol{G}\mathbf{D}$ is a non-negative symmetric matrix with the eigenvalues of $\boldsymbol{G}$ and eigenvectors $\mathbf{D}\boldsymbol{v}_j$; by the Perron–Frobenius theorem its spectral radius λ_1 is an eigenvalue with a non-negative eigenvector, so $\mathbf{D}\boldsymbol{v}_1$ can be chosen non-negative, and

when $\tilde{\boldsymbol{G}}$ is entrywise positive this eigenvalue is simple and the eigenvector strictly positive. (v) $\sum_m v_{1,m}\boldsymbol{q}^m = \Pi_\perp \boldsymbol{Z}\boldsymbol{v}_1 = \sigma_1 \Pi_\perp \boldsymbol{u}_1 = \mathbf{0}$. □

A.3 Proof of Proposition 3

Let $\mathbf{D} = \mathrm{diag}(\mathrm{sign}(\boldsymbol{v}_1))$ (zero entries may be given either sign), $\tilde{\boldsymbol{z}}^m = D_{mm}\boldsymbol{z}^m$ and $\tilde{\boldsymbol{G}} = \mathbf{D}\boldsymbol{G}\mathbf{D}$, the Gram matrix of the $\tilde{\boldsymbol{z}}^m$. From $\boldsymbol{Z}\boldsymbol{v}_1 = \sigma_1\boldsymbol{u}_1$ and $\boldsymbol{Z}^\top\boldsymbol{u}_1 = \sigma_1\boldsymbol{v}_1$ we get $\boldsymbol{u}_1 = \sigma_1^{-1}\sum_m |v_{1,m}|\tilde{\boldsymbol{z}}^m$ and $\langle \boldsymbol{u}_1, \tilde{\boldsymbol{z}}^m\rangle = \sigma_1|v_{1,m}| \geq 0$ for every m, and $\lambda_1 = \sum_m \langle \boldsymbol{u}_1, \tilde{\boldsymbol{z}}^m\rangle^2$ by Lemma 1. Suppose $\tilde{\boldsymbol{G}}_{mm'} = \langle \tilde{\boldsymbol{z}}^m, \tilde{\boldsymbol{z}}^{m'}\rangle = g < 0$ and write $p = \langle \boldsymbol{u}_1, \tilde{\boldsymbol{z}}^m\rangle \geq 0$, $q = \langle \boldsymbol{u}_1, \tilde{\boldsymbol{z}}^{m'}\rangle \geq 0$. The Gram matrix of $(\boldsymbol{u}_1, \tilde{\boldsymbol{z}}^m, \tilde{\boldsymbol{z}}^{m'})$ is positive semidefinite, so its determinant $1-p^2-q^2-g^2+2pqg$ is non-negative, whence $p^2+q^2 \leq 1-g^2+2pqg \leq 1-g^2$ because $pq \geq 0 > g$. The remaining $k-2$ terms of λ_1 are at most 1 each, so $\lambda_1 \leq k-1-g^2$. Consequently, if $\lambda_1 > k-1$, i.e., $\rho > 1-1/k$, then $\tilde{\boldsymbol{G}}$ has no negative entry, $\boldsymbol{G}$ is balanced with this $\mathbf{D}$, and Lemma 2(iv) applied to $\tilde{\boldsymbol{G}}$ gives a non-negative leading eigenvector $\tilde{\boldsymbol{v}}$ of $\tilde{\boldsymbol{G}}$ with $\boldsymbol{v}_1 = \pm\mathbf{D}\tilde{\boldsymbol{v}}$; the sign convention $\sum_m v_{1,m} > 0$ selects $\mathbf{D}\boldsymbol{v}_1 = \tilde{\boldsymbol{v}} \geq \mathbf{0}$. For $k=3$, a sign matrix $\mathbf{D}$ with $\mathbf{D}\boldsymbol{G}\mathbf{D} \geq 0$ exists iff the number of negative off-diagonal entries is even, i.e., iff $\boldsymbol{G}_{12}\boldsymbol{G}_{13}\boldsymbol{G}_{23} \geq 0$ (a balanced signed triangle); if the triangle is frustrated, every sign matrix, in particular $\mathbf{D} = \mathrm{diag}(\mathrm{sign}(\boldsymbol{v}_1))$, leaves a negative entry g, so $\lambda_1 \leq 2-g^2 < 2$. The bound is tight: the planar configuration $\boldsymbol{z}^1 = (1,0)$, $\boldsymbol{z}^2 = (\cos\theta, \sin\theta)$, $\boldsymbol{z}^3 = (\cos\theta, -\sin\theta)$ has $\boldsymbol{G}_{23} = \cos 2\theta \to 0^-$ and $\lambda_1 \to 2$ as $\theta \downarrow \pi/4$.

The conclusion is about balance and cannot be strengthened to non-negativity of $\boldsymbol{v}_1$ in the original frame. The instance $\boldsymbol{Z} = [\boldsymbol{u}, \boldsymbol{u}, -\boldsymbol{u}]$ has $\boldsymbol{G}$ with entries $\boldsymbol{G}_{12} = 1$, $\boldsymbol{G}_{13} = \boldsymbol{G}_{23} = -1$, hence $\boldsymbol{G}_{12}\boldsymbol{G}_{13}\boldsymbol{G}_{23} = 1 \geq 0$: it is balanced, $\rho = 1$, and $\boldsymbol{v}_1 = (1,1,-1)^\top/\sqrt{3}$ with $\mathbf{D} = \mathrm{diag}(1,1,-1)$, so $\mathbf{D}\boldsymbol{v}_1 \geq \mathbf{0}$ as the proposition states while $\boldsymbol{v}_1$ itself has a negative entry. The global orientation rule $\sum_m v_{1,m} = 1/\sqrt{3} > 0$ does not detect it, which is why the objective contains the per-pair hinge $\sum_{m<m'}[-\boldsymbol{G}_{mm'}]_+$ rather than a single sign convention; a frustrated configuration, in contrast, is excluded by the spectral bound itself. □

A.4 Proof of Proposition 6

Fix a modality m and an instance i with specific variance $v_{m,i}$. Under Assumption 3 the complement score $t = \boldsymbol{b}^{m\top}\boldsymbol{q}^m = e^m + \nu$ with $e^m \sim \mathcal{N}(0, v_{m,i})$ and $\nu \sim \mathcal{N}(0, \varsigma^2)$ independent, so $\mathbb{E}[e^m \mid t] = \alpha_{m,i}t$ with $\alpha_{m,i} = v_{m,i}/(v_{m,i}+\varsigma^2) = \gamma_{m,i}/(1+\gamma_{m,i})$ and $\mathbb{E}[t^2] = v_{m,i}+\varsigma^2$. For any gate $\bar{\alpha}_m$, the orthogonality of the Bayes error gives $\mathbb{E}[(e^m - \bar{\alpha}_m t)^2 \mid i] = \mathbb{E}[(e^m - \alpha_{m,i}t)^2 \mid i] + (\alpha_{m,i}-\bar{\alpha}_m)^2(v_{m,i}+\varsigma^2)$. The term enters the prediction of y through $\beta_m e^m$ and is independent of the other terms of the Bayes predictor, so averaging over i yields the stated additional risk. □

A.5 Proof of Corollary 8

(i) Plug-in direction. Write the additive Bayes predictor of Theorem 5 as $\hat{y}^\star = \boldsymbol{w}^\top\phi(\boldsymbol{Z}; \Pi, \boldsymbol{w}_c)$ with features $\phi(\boldsymbol{Z}; \Pi, \boldsymbol{w}_c) = \big(\Pi\boldsymbol{Z}\boldsymbol{w}_c,\ (\boldsymbol{I}-\Pi)\boldsymbol{z}^1, \ldots, (\boldsymbol{I}-\Pi)\boldsymbol{z}^k\big)$, weights $\boldsymbol{w} = \big(\beta_0\bar{\gamma}_{s,k}\boldsymbol{a}/\sqrt{k},\ \beta_1\bar{\gamma}_1\boldsymbol{b}^1, \ldots, \beta_k\bar{\gamma}_k\boldsymbol{b}^k\big)$, where the factor $1/\sqrt{k}$ is that of the consensus term $\boldsymbol{a}^\top\boldsymbol{c}/\sqrt{k}$ in Theorem 5, the uniform aggregation $\boldsymbol{w}_c = \mathbf{1}/\sqrt{k}$ and the oracle projector $\Pi = \boldsymbol{a}\boldsymbol{a}^\top$; the plug-in predictor is $\hat{y} = \boldsymbol{w}^\top\phi(\boldsymbol{Z}; \hat{\Pi}, \boldsymbol{w}_c)$ with $\hat{\Pi} = \boldsymbol{u}_1\boldsymbol{u}_1^\top$. Since $\hat{y}^\star$ is the Bayes predictor and $\hat{y} - \hat{y}^\star$ is a function of $\boldsymbol{Z}$, $\mathbb{E}[(y-\hat{y})^2] = R_{PRISM} + \mathbb{E}[(\hat{y}-\hat{y}^\star)^2]$ and $(\hat{y}-\hat{y}^\star)^2 \leq \|\boldsymbol{w}\|^2\|\phi(\boldsymbol{Z}; \hat{\Pi}, \boldsymbol{w}_c) - \phi(\boldsymbol{Z}; \Pi, \boldsymbol{w}_c)\|^2$. For rank-one projectors $\|\hat{\Pi} - \Pi\|_2 = \sin\angle(\boldsymbol{u}_1, \boldsymbol{a})$, and since $\|\boldsymbol{Z}\boldsymbol{w}_c\| \leq \|\boldsymbol{Z}\|_2 \leq \|\boldsymbol{Z}\|_F$ and $\sum_m \|\boldsymbol{z}^m\|^2 = \|\boldsymbol{Z}\|_F^2$, the squared feature difference is at most $2\|\boldsymbol{Z}\|_F^2 \sin^2\angle(\boldsymbol{u}_1, \boldsymbol{a})$, while $\|\boldsymbol{w}\|^2 \leq \beta_0^2/k + \sum_m \beta_m^2$ because $\bar{\gamma} \leq 1$ and $\boldsymbol{a}, \boldsymbol{b}^{1:k}$ are unit vectors; taking expectations gives the stated bound. We deliberately keep $\|\boldsymbol{Z}\|_F^2$ inside the expectation instead of replacing it by k: under Assumption 3 the representations are Gaussian with $\mathbb{E}\|\boldsymbol{Z}\|_F^2 = k(v_s + \bar{v} + d\varsigma^2)$, $\bar{v} = \frac{1}{k}\sum_m v_m$, and are not unit vectors, whereas $\|\boldsymbol{Z}\|_F^2 = k$ holds for the normalised representations of Section 3; substituting a norm bound from the implementation into the analysis model would be exactly the step we avoid.

(ii) Data-dependent weights. Replacing $\boldsymbol{w}_c = \mathbf{1}/\sqrt{k}$ by $\boldsymbol{v}_1$ changes the prediction by $\beta_0\bar{\gamma}_{s,k}\boldsymbol{a}^\top\boldsymbol{Z}(\boldsymbol{v}_1 - \mathbf{1}/\sqrt{k})/\sqrt{k}$, whose square is at most $\frac{\beta_0^2}{k}\|\boldsymbol{Z}\|_2^2\|\boldsymbol{v}_1 - \mathbf{1}/\sqrt{k}\|^2 = \frac{\beta_0^2}{k}\lambda_1\|\boldsymbol{v}_1 -$

$\mathbf{1}/\sqrt{k}\|^2$. Fix the sign of $\boldsymbol{v}_1$ so that $\langle \boldsymbol{v}_1, \mathbf{1}\rangle \geq 0$ and let $\theta = \angle(\boldsymbol{v}_1, \mathbf{1}/\sqrt{k})$; then $\|\boldsymbol{v}_1 - \mathbf{1}/\sqrt{k}\|^2 = 2(1-\cos\theta) \leq 2\sin^2\theta$. Writing $\mathbf{1}/\sqrt{k} = \cos\theta\, \boldsymbol{v}_1 + \sin\theta\, \mathbf{p}$ with $\mathbf{p} \perp \boldsymbol{v}_1$, $\|\mathbf{p}\| = 1$, the Rayleigh quotient satisfies $\mathbf{1}^\top \boldsymbol{G}\mathbf{1}/k = \cos^2\theta\,\lambda_1 + \sin^2\theta\, \mathbf{p}^\top \boldsymbol{G}\mathbf{p} \leq \cos^2\theta\,\lambda_1 + \sin^2\theta\,\lambda_2$, i.e., $\sin^2\theta \leq (\lambda_1 - \mathbf{1}^\top \boldsymbol{G}\mathbf{1}/k)/(\lambda_1 - \lambda_2)$, which together with the trivial bound $\|\boldsymbol{v}_1 - \mathbf{1}/\sqrt{k}\|^2 \leq 2$ (both are unit vectors) is the stated bound; the truncation matters when $\lambda_1 \to \lambda_2$, and the bound vanishes in the aligned regime $\boldsymbol{G} \to \mathbf{1}\mathbf{1}^\top$. Combining (i) and (ii) with $(a+b)^2 \leq 2a^2 + 2b^2$ gives at most twice the sum, as claimed.

(iii) Angle bound. Theorem 7 applied to the rank-one matrix $\boldsymbol{Z}^\star = s\boldsymbol{a}\mathbf{1}^\top$, whose leading left singular vector is $\boldsymbol{a}$, whose leading singular value is $|s|\sqrt{k}$ and whose gap is $\Delta(\boldsymbol{Z}^\star) = s^2$, perturbed by $\boldsymbol{E} = \boldsymbol{Z} - \boldsymbol{Z}^\star = \sum_m (e^m \boldsymbol{b}^m + \boldsymbol{\epsilon}^m)\mathbf{e}_m^\top$, gives the stated angle bound. The truncation $\min\{1, \cdot\}$ is not cosmetic: the untruncated right-hand side behaves as s^{-2} near $s = 0$, and $\mathbb{E}[s^{-2}] = \infty$ under $s \sim \mathcal{N}(0, v_s)$, whereas $\sin^2\angle(\boldsymbol{u}_1, \boldsymbol{a}) \leq 1$ always, so the truncated bound is integrable and the excess risk in (i) is finite. □

Remark on the energy factor. The energy factor stays explicit because the two models live on different scales. For the normalised representations of Section 3, $\|\boldsymbol{Z}\|_F^2 = k$. Under Assumption 3, the representations are Gaussian with $\mathbb{E}\|\boldsymbol{Z}\|_F^2 = k(v_s + \bar{v} + d\varsigma^2)$, and a norm bound from one model must not be used inside the other. The second term vanishes as $\boldsymbol{G}$ approaches the aligned regime, where $\boldsymbol{v}_1 \to \mathbf{1}/\sqrt{k}$.

A.6 Proof of Theorem 4

Recall. Under Assumption 2 and full alignment ($\boldsymbol{z}^1 = \cdots = \boldsymbol{z}^k = \bar{\boldsymbol{z}}$ a.s.), every measurable predictor $\hat{y} = \varphi(\bar{\boldsymbol{z}})$ satisfies $\mathbb{E}[(y - \hat{y})^2] \geq \mathrm{Var}(\eta) + \sum_m \beta_m^2 \mathrm{Var}(e^m)$.

Proof. Step 1 (common information). Since $\bar{\boldsymbol{z}} = f^m(\boldsymbol{x}^m)$ for every m, $\bar{\boldsymbol{z}}$ is $\sigma(\boldsymbol{x}^m)$-measurable for every m, hence measurable with respect to $\bigcap_m \sigma(\boldsymbol{x}^m)$. Because h^m is injective, $\sigma(\boldsymbol{x}^m) = \sigma(s, e^m)$. We claim that any random variable W measurable w.r.t. $\sigma(s, e^1) \cap \sigma(s, e^2)$ is a.s. equal to a function of s. Write $W = F_1(s, e^1) = F_2(s, e^2)$ a.s. Conditioning on s, and using that e^1 and e^2 are independent of s and of each other, $F_1(s, e^1) = F_2(s, e^2)$ a.s. implies that for P_s-a.e. s the function $F_1(s, \cdot)$ is a.s. constant under P_{e^1} (a function of e^1 equal a.s. to a function of an independent variable must be a.s. constant). Hence $W = \mathbb{E}[W \mid s]$ a.s. Applying the claim pairwise shows that $\bar{\boldsymbol{z}}$ is a.s. a function of s; write $\bar{\boldsymbol{z}} = \zeta(s)$.

Step 2 (risk decomposition). For any measurable φ, with $\hat{y} = \varphi(\zeta(s))$, decompose $y - \hat{y} = \big(g(s) - \varphi(\zeta(s))\big) + \sum_m \beta_m e^m + \eta$. The three terms are pairwise uncorrelated: the first is a function of s, while e^m and η are zero-mean and independent of s and of each other. Therefore $\mathbb{E}[(y - \hat{y})^2] = \mathbb{E}[(g(s) - \varphi(\zeta(s)))^2] + \sum_m \beta_m^2 \mathrm{Var}(e^m) + \mathrm{Var}(\eta) \geq \sum_m \beta_m^2 \mathrm{Var}(e^m) + \mathrm{Var}(\eta)$, with equality iff $\varphi(\zeta(s)) = g(s)$ a.s., e.g., when ζ is injective. The Bayes-optimal predictor from all raw modalities is $\mathbb{E}[y \mid \boldsymbol{x}^{1:k}] = g(s) + \sum_m \beta_m e^m$ with risk $\mathrm{Var}(\eta)$. □

A.7 Proof of Theorem 5

Recall. Under Assumption 3, the minimum mean-squared errors are

$$\begin{aligned}
&\text{(a) consensus only, } \Pi_{\mathcal{C}}\boldsymbol{Z}: && R_{\mathcal{C}} = v_\eta + \beta_0^2 \tfrac{v_s}{1+k\gamma_s} + \textstyle\sum_m \beta_m^2 v_m,\\
&\text{(b) a single modality } \boldsymbol{z}^m: && R_m = v_\eta + \beta_0^2 \tfrac{v_s}{1+\gamma_s} + \beta_m^2 \tfrac{v_m}{1+\gamma_m} + \textstyle\sum_{m' \neq m} \beta_{m'}^2 v_{m'},\\
&\text{(c) consensus and complements } (\boldsymbol{c}, \boldsymbol{q}^{1:k}): && R_{PRISM} = v_\eta + \beta_0^2 \tfrac{v_s}{1+k\gamma_s} + \textstyle\sum_m \beta_m^2 \tfrac{v_m}{1+\gamma_m},
\end{aligned}$$

and (c) coincides with the Bayes risk of the full representation $\boldsymbol{Z}$.

Proof. Throughout, all variables are jointly Gaussian, so conditional expectations are linear and the minimum MSE of predicting y from a set of observations O equals $\mathrm{Var}(y \mid O)$. Because $\boldsymbol{a}, \boldsymbol{b}^1, \ldots, \boldsymbol{b}^k$ are orthonormal and the noise is isotropic, the coordinates $\boldsymbol{a}^\top \boldsymbol{\epsilon}^m$, $\boldsymbol{b}^{m\top} \boldsymbol{\epsilon}^m$ and the components orthogonal to $\mathrm{span}(\boldsymbol{a}, \boldsymbol{b}^{1:k})$ are independent $\mathcal{N}(0, \varsigma^2)$ variables, and the latter carry no information about $(s, e^{1:k})$. Hence, for any set of observed vectors, the relevant statistics are the projections on $\boldsymbol{a}$ and $\boldsymbol{b}^{1:k}$. We repeatedly use the scalar identity: if $t \sim \mathcal{N}(0, v)$ is observed through

n independent copies $t + \nu_i$, $\nu_i \sim \mathcal{N}(0, \varsigma^2)$, then $\mathrm{Var}(t \mid \text{obs}) = (1/v + n/\varsigma^2)^{-1} = v/(1 + n\gamma)$ with $\gamma = v/\varsigma^2$, and the posterior mean is $\frac{n\gamma}{1+n\gamma}$ times the sample mean of the observations.

(a) Consensus only. $\Pi_{\mathcal{C}}\boldsymbol{Z} = \boldsymbol{a}\boldsymbol{a}^\top\boldsymbol{Z}$ is determined by the k scalars $\boldsymbol{a}^\top\boldsymbol{z}^m = s + \boldsymbol{a}^\top\boldsymbol{\epsilon}^m$, k independent noisy copies of s; it contains no information about e^m since $\boldsymbol{a} \perp \boldsymbol{b}^m$. Therefore $\mathrm{Var}(y \mid \Pi_{\mathcal{C}}\boldsymbol{Z}) = \beta_0^2 \frac{v_s}{1+k\gamma_s} + \sum_m \beta_m^2 v_m + v_\eta$. With uniform weights $\boldsymbol{w} = \mathbf{1}/\sqrt{k}$, $\boldsymbol{a}^\top\boldsymbol{c} = \sqrt{k}\, s + \sum_m \boldsymbol{a}^\top\boldsymbol{\epsilon}^m/\sqrt{k}$ is a sufficient statistic for s within $\Pi_{\mathcal{C}}\boldsymbol{Z}$, with SNR $k\gamma_s$, which is the k-fold gain.

(b) Single modality. $\boldsymbol{z}^m$ provides one copy $\boldsymbol{a}^\top\boldsymbol{z}^m = s + \boldsymbol{a}^\top\boldsymbol{\epsilon}^m$ of s and one copy $\boldsymbol{b}^{m\top}\boldsymbol{z}^m = e^m + \boldsymbol{b}^{m\top}\boldsymbol{\epsilon}^m$ of e^m; $\boldsymbol{b}^{m'\top}\boldsymbol{z}^m$ for $m' \neq m$ is pure noise. Hence $\mathrm{Var}(y \mid \boldsymbol{z}^m) = \beta_0^2 \frac{v_s}{1+\gamma_s} + \beta_m^2 \frac{v_m}{1+\gamma_m} + \sum_{m' \neq m} \beta_{m'}^2 v_{m'} + v_\eta$.

(c) Consensus and complements. With the oracle projector, $\boldsymbol{q}^m = \Pi_\perp \boldsymbol{z}^m = e^m\boldsymbol{b}^m + \Pi_\perp\boldsymbol{\epsilon}^m$, whose projection on $\boldsymbol{b}^m$ is $e^m + \boldsymbol{b}^{m\top}\boldsymbol{\epsilon}^m$; together with $\boldsymbol{a}^\top\boldsymbol{c}$ from (a), the features $(\boldsymbol{c}, \boldsymbol{q}^{1:k})$ contain k copies of s and one copy of each e^m, all with independent noise. Thus $\mathrm{Var}(y \mid \boldsymbol{c}, \boldsymbol{q}^{1:k}) = \beta_0^2 \frac{v_s}{1+k\gamma_s} + \sum_m \beta_m^2 \frac{v_m}{1+\gamma_m} + v_\eta$. The full matrix $\boldsymbol{Z}$ provides exactly the same statistics (its remaining coordinates are noise), so $(\boldsymbol{c}, \boldsymbol{q}^{1:k})$ is sufficient and (c) equals the Bayes risk of $\boldsymbol{Z}$. The two gains in Eq. (2) follow by subtraction, using $v - \frac{v}{1+\gamma} = v\frac{\gamma}{1+\gamma}$ and $\frac{v_s}{1+\gamma_s} - \frac{v_s}{1+k\gamma_s} = v_s \frac{(k-1)\gamma_s}{(1+\gamma_s)(1+k\gamma_s)}$.

Shrinkage. $(e^m\boldsymbol{b}^m, \boldsymbol{q}^m)$ is jointly Gaussian with $\mathrm{Cov}(e^m\boldsymbol{b}^m, \boldsymbol{q}^m) = v_m\boldsymbol{b}^m\boldsymbol{b}^{m\top}$ and $\mathrm{Cov}(\boldsymbol{q}^m) = v_m\boldsymbol{b}^m\boldsymbol{b}^{m\top} + \varsigma^2\Pi_\perp$; on $\mathrm{range}(\Pi_\perp)$ the posterior mean is

$$\mathbb{E}[e^m \mid \boldsymbol{q}^m] = \alpha_m^\star\, \boldsymbol{b}^{m\top}\boldsymbol{q}^m, \qquad \mathbb{E}[e^m\boldsymbol{b}^m \mid \boldsymbol{q}^m] = \alpha_m^\star\, \boldsymbol{b}^m\boldsymbol{b}^{m\top}\boldsymbol{q}^m, \qquad \alpha_m^\star = \tfrac{v_m}{v_m+\varsigma^2} = \tfrac{\gamma_m}{1+\gamma_m}. \tag{7}$$

The projector $\boldsymbol{b}^m\boldsymbol{b}^{m\top}$ is part of the estimator and cannot be dropped: the component of $\boldsymbol{q}^m$ orthogonal to $\boldsymbol{b}^m$ is pure noise, and the scaled complement $\alpha_m^\star\boldsymbol{q}^m$ retains it. What is true, and what PRISM uses, is the weaker statement about the *predictor*: the Bayes predictor of y enters through the scalar $\beta_m\alpha_m^\star\boldsymbol{b}^{m\top}\boldsymbol{q}^m$, and for a head that is linear in its argument, $g_m(\boldsymbol{q}) = \boldsymbol{w}_m^\top\boldsymbol{q}$, one has $g_m(\alpha_m^\star\boldsymbol{q}^m) = \alpha_m^\star\boldsymbol{w}_m^\top\boldsymbol{q}^m$, which equals the Bayes term when $\boldsymbol{w}_m \propto \boldsymbol{b}^m$; the projection is then performed by the head and the gate only carries the shrinkage. Two consequences are worth stating. First, $\alpha_m^\star\boldsymbol{q}^m$ is *not* the MMSE estimator of $e^m\boldsymbol{b}^m$, only its unprojected scaling, so the optimality claim is about the coefficient, not about the vector. Second, for a head trained freely, any *constant* $\alpha_m > 0$ is absorbed into $\boldsymbol{w}_m$, so the constant part of the gate is not identified by risk at all; only its instance-dependent variation changes the attainable risk, which is the content of Proposition 6 and the reason the ablation of constant versus certificate gates measures a small effect (Appendix C.1).
Editing. The editing objective is strictly convex in $\boldsymbol{\delta}$; its gradient $\boldsymbol{\delta} + \kappa_m(\boldsymbol{q}^m + \boldsymbol{\delta})$ vanishes at $\boldsymbol{\delta}^\star = -\frac{\kappa_m}{1+\kappa_m}\boldsymbol{q}^m \in \mathrm{range}(\Pi_\perp)$, giving $\boldsymbol{q}^m + \boldsymbol{\delta}^\star = \frac{1}{1+\kappa_m}\boldsymbol{q}^m$. □

A.8 Properties of the Consensus-preserving Edit

Proposition 9 (Properties of the consensus-preserving edit)**.** *Let $T_{\boldsymbol{\alpha}}(\boldsymbol{Z}) = \Pi_{\mathcal{C}}\boldsymbol{Z} + \Pi_\perp\boldsymbol{Z}\,\mathrm{diag}(\boldsymbol{\alpha})$ with $\boldsymbol{\alpha} \in [0,1]^k$. Then (i) $\Pi_{\mathcal{C}}T_{\boldsymbol{\alpha}}(\boldsymbol{Z}) = \Pi_{\mathcal{C}}\boldsymbol{Z}$ (consensus preservation); (ii) $\|T_{\boldsymbol{\alpha}}(\boldsymbol{Z})\|_F \le \|\boldsymbol{Z}\|_F$ and, for fixed projectors, $T_{\boldsymbol{\alpha}}$ is 1-Lipschitz (non-expansiveness); (iii) the energy fraction $\rho(\mathbf{M}) = \lambda_1(\mathbf{M})/\|\mathbf{M}\|_F^2$, which extends the consensus ratio to matrices with unnormalised columns and equals λ_1/k for $\boldsymbol{Z}$, satisfies $\rho(T_{\boldsymbol{\alpha}}(\boldsymbol{Z})) \ge \rho(\boldsymbol{Z})$, with equality iff $\alpha_m = 1$ whenever $r_m > 0$.*

Proof. (i) $\Pi_{\mathcal{C}}\Pi_\perp = \mathbf{0}$, so $\Pi_{\mathcal{C}}T_{\boldsymbol{\alpha}}(\boldsymbol{Z}) = \Pi_{\mathcal{C}}\boldsymbol{Z}$. (ii) By orthogonality of the two projectors, $\|T_{\boldsymbol{\alpha}}(\boldsymbol{Z})\|_F^2 = \|\Pi_{\mathcal{C}}\boldsymbol{Z}\|_F^2 + \sum_m \alpha_m^2\|\boldsymbol{q}^m\|^2 \le \|\Pi_{\mathcal{C}}\boldsymbol{Z}\|_F^2 + \sum_m \|\boldsymbol{q}^m\|^2 = \|\boldsymbol{Z}\|_F^2$; the same computation applied to $\boldsymbol{Z} - \boldsymbol{Z}'$ (with the projectors held fixed) gives the Lipschitz bound. (iii) Let $\boldsymbol{Z}' = T_{\boldsymbol{\alpha}}(\boldsymbol{Z})$. Since $\Pi_{\mathcal{C}}\boldsymbol{Z}' = \Pi_{\mathcal{C}}\boldsymbol{Z} = \sigma_1\boldsymbol{u}_1\boldsymbol{v}_1^\top$ and the rank-one projection can only decrease the spectral norm, $\lambda_1(\boldsymbol{Z}') = \|\boldsymbol{Z}'\|_2^2 \ge \|\Pi_{\mathcal{C}}\boldsymbol{Z}'\|_2^2 = \lambda_1(\boldsymbol{Z})$, while $\|\boldsymbol{Z}'\|_F^2 = \lambda_1(\boldsymbol{Z}) + \sum_m \alpha_m^2 r_m \le k$. Hence $\lambda_1(\boldsymbol{Z}')/\|\boldsymbol{Z}'\|_F^2 \ge \lambda_1(\boldsymbol{Z})/k = \rho(\boldsymbol{Z})$, with equality iff $\alpha_m^2 r_m = r_m$ for all m. □

A.9 Proof of Theorem 7

Proof. Let $\boldsymbol{A} = \boldsymbol{Z}\boldsymbol{Z}^\top$ and $\tilde{\boldsymbol{A}} = \tilde{\boldsymbol{Z}}\tilde{\boldsymbol{Z}}^\top = \boldsymbol{A} + \boldsymbol{Z}\boldsymbol{E}^\top + \boldsymbol{E}\boldsymbol{Z}^\top + \boldsymbol{E}\boldsymbol{E}^\top$, so $\|\tilde{\boldsymbol{A}} - \boldsymbol{A}\|_2 \le 2\|\boldsymbol{Z}\|_2\|\boldsymbol{E}\|_2 + \|\boldsymbol{E}\|_2^2 = 2\sigma_1\|\boldsymbol{E}\|_2 + \|\boldsymbol{E}\|_2^2$. The Davis–Kahan $\sin\Theta$ theorem in the form of Yu

et al. (2015) gives, for the top eigenvectors of $\boldsymbol{A}$ and $\tilde{\boldsymbol{A}}$, $\sin\angle(\boldsymbol{u}_1, \tilde{\boldsymbol{u}}_1) \le 2\|\tilde{\boldsymbol{A}} - \boldsymbol{A}\|_2/(\lambda_1 - \lambda_2)$, and $\lambda_1 - \lambda_2 = k\Delta(\boldsymbol{Z})$. For rank-one projectors $\|\tilde{\Pi}_{\mathcal{C}} - \Pi_{\mathcal{C}}\|_2 = \sin\angle(\boldsymbol{u}_1, \tilde{\boldsymbol{u}}_1)$. For the second claim write $T_{\boldsymbol{\alpha}}(\tilde{\boldsymbol{Z}}) - T_{\boldsymbol{\alpha}}(\boldsymbol{Z}) = \tilde{\Pi}_{\mathcal{C}}\boldsymbol{E} + \tilde{\Pi}_{\perp}\boldsymbol{E}\,\mathrm{diag}(\boldsymbol{\alpha}) + (\tilde{\Pi}_{\mathcal{C}} - \Pi_{\mathcal{C}})\boldsymbol{Z}(\boldsymbol{I} - \mathrm{diag}(\boldsymbol{\alpha}))$, using $\tilde{\Pi}_{\perp} - \Pi_{\perp} = -(\tilde{\Pi}_{\mathcal{C}} - \Pi_{\mathcal{C}})$. The first two terms are orthogonal pieces of $\boldsymbol{E}$ scaled by factors in $[0, 1]$, so $\|\tilde{\Pi}_{\mathcal{C}}\boldsymbol{E} + \tilde{\Pi}_{\perp}\boldsymbol{E}\,\mathrm{diag}(\boldsymbol{\alpha})\|_F^2 = \|\tilde{\Pi}_{\mathcal{C}}\boldsymbol{E}\|_F^2 + \|\tilde{\Pi}_{\perp}\boldsymbol{E}\,\mathrm{diag}(\boldsymbol{\alpha})\|_F^2 \le \|\tilde{\Pi}_{\mathcal{C}}\boldsymbol{E}\|_F^2 + \|\tilde{\Pi}_{\perp}\boldsymbol{E}\|_F^2 = \|\boldsymbol{E}\|_F^2$; the last term has Frobenius norm at most $\|\tilde{\Pi}_{\mathcal{C}} - \Pi_{\mathcal{C}}\|_2\|\boldsymbol{Z}\|_F\|\boldsymbol{I} - \mathrm{diag}(\boldsymbol{\alpha})\|_2 \le \sqrt{k}\sin\angle(\boldsymbol{u}_1, \tilde{\boldsymbol{u}}_1)$ because $\|\boldsymbol{Z}\|_F^2 = k$ and $\alpha_m \in [0, 1]$. The triangle inequality yields the stated bound. □

A.10 Gradient Analysis of the Gap Regulariser

With $\lambda_j = \boldsymbol{v}_j^\top \boldsymbol{G}\boldsymbol{v}_j$ and $\boldsymbol{G} = \boldsymbol{Z}^\top\boldsymbol{Z}$, the Hellmann–Feynman identity gives $\partial\lambda_j/\partial\boldsymbol{G} = \boldsymbol{v}_j\boldsymbol{v}_j^\top$ at an eigenpair, hence $\partial\lambda_j/\partial\boldsymbol{Z} = 2\boldsymbol{Z}\boldsymbol{v}_j\boldsymbol{v}_j^\top$ and

$$\frac{\partial\mathcal{L}_{\mathrm{gap}}}{\partial\boldsymbol{Z}} = -\frac{2}{k}\,\boldsymbol{Z}\big(\boldsymbol{v}_1\boldsymbol{v}_1^\top - \boldsymbol{v}_2\boldsymbol{v}_2^\top\big) = -\frac{2}{k}\big(\sigma_1\boldsymbol{u}_1\boldsymbol{v}_1^\top - \sigma_2\boldsymbol{u}_2\boldsymbol{v}_2^\top\big). \tag{8}$$

For the m-th modality, $\partial\mathcal{L}_{\mathrm{gap}}/\partial\boldsymbol{z}^m = -\frac{2}{k}(\sigma_1 v_{1,m}\boldsymbol{u}_1 - \sigma_2 v_{2,m}\boldsymbol{u}_2)$: gradient descent moves $\boldsymbol{z}^m$ towards the leading direction in proportion to its loading $v_{1,m}$ and away from the second direction in proportion to $v_{2,m}$, and leaves the components along $\boldsymbol{u}_3, \ldots, \boldsymbol{u}_k$ untouched. In contrast, the singular-value objective $\mathcal{L}_{\mathrm{align}}$ has gradient $\frac{1}{\tau}[(p_1 - 1)\boldsymbol{u}_1\boldsymbol{v}_1^\top + \sum_{j\ge 2} p_j\boldsymbol{u}_j\boldsymbol{v}_j^\top]$ with softmax weights p_j, which suppresses *every* non-leading direction, and the GRAM log-volume objective $\log\det\boldsymbol{G} = 2\sum_j \log\sigma_j$, which we use as an unconstrained per-instance penalty without the contrastive normalisation of Cicchetti et al. (2025), has gradient $2\sum_j \boldsymbol{u}_j\boldsymbol{v}_j^\top/\sigma_j$, which is dominated by the smallest singular value, unbounded as $\sigma_j \to 0$, and drives it to zero first; Figure 4(a)–(b) shows the three behaviours empirically. When λ_1 and λ_2 are nearly degenerate, the eigenvectors are ill defined but the Rayleigh-quotient gradient remains bounded, which is why we treat the eigenvectors as constants (Section 4.2); back-propagating through `eigh` instead is compared in Appendix C.1.

B Implementation Details

B.1 Datasets and Features

- **CMU-MOSI** (Zadeh et al., 2016) contains 2,199 opinion utterances from 93 YouTube monologue videos (1,284/229/686 train/validation/test) labelled with sentiment intensity in $[-3, 3]$. Features: 768-dimensional BERT word-piece states (50 tokens), 5-dimensional COVAREP acoustic descriptors and 20-dimensional Facet visual descriptors, word-aligned (`aligned_50.pkl`).
- **CMU-MOSEI** (Zadeh et al., 2018b) contains 22,856 utterances from 3,228 YouTube videos (16,326/1,871/4,659) with the same label range. Features: BERT states (50 tokens), 74-dimensional COVAREP acoustics and 35-dimensional Facet visual descriptors, word-aligned.
- **CH-SIMS** (Yu et al., 2020) contains 2,281 Chinese utterances from 60 movies and TV series (1,368/456/457) labelled in $[-1, 1]$, together with unimodal labels that we use only in one ablation. Features: Chinese BERT states (39 tokens), 33-dimensional acoustic descriptors (400 frames) and 709-dimensional OpenFace visual descriptors (55 frames), unaligned (`unaligned_39.pkl`).

All features are the processed files distributed with the MMSA benchmark (Mao et al., 2022), exactly as used by EMOE and ALMT. Acoustic and visual features are standardised per dimension with training-set statistics.

B.2 Baselines

- **TFN** (Zadeh et al., 2017) and **LMF** (Liu et al., 2018) model unimodal, bimodal and trimodal interactions with an outer product of the modality representations, computed exactly or with a low-rank factorisation.
- **MulT** (Tsai et al., 2019a) fuses unaligned sequences with directional pairwise cross-modal Transformers.
- **MISA** (Hazarika et al., 2020) projects every modality into a modality-invariant and a modality-specific subspace with distributional similarity, orthogonality and reconstruction losses.

- **MAG-BERT** (Rahman et al., 2020) injects acoustic and visual information into BERT through a multimodal adaptation gate.
- **Self-MM** (Yu et al., 2021) generates unimodal pseudo-labels in a self-supervised manner to train unimodal heads jointly with the fusion head.
- **MMIM** (Han et al., 2021b) hierarchically maximises the mutual information between unimodal inputs and between the fused representation and the unimodal ones.
- **DMD** (Li et al., 2023b) decouples homogeneous and heterogeneous features and distils knowledge across modalities with graph distillation units.
- **ALMT** (Zhang et al., 2023a) learns a language-guided hyper-modality representation that suppresses irrelevant acoustic and visual information.
- **EMOE** (Fang et al., 2025) uses a mixture of modality experts with a router that learns sample-specific modality weights, balanced by a router-entropy and a weight–importance similarity loss, and distils unimodal predictions into the multimodal ones. It is the state-of-the-art dynamic fusion method and our closest baseline.

For the re-runs we use the MMSA implementations (TFN, LMF, MulT, MISA, MAG-BERT, Self-MM, MMIM, ALMT) with their released per-dataset hyper-parameters, and the official code of DMD and EMOE; see Appendix C.2.

B.3 Evaluation Metrics

For MOSI/MOSEI, Acc-7 and Acc-5 are the accuracies after rounding predictions and labels clipped to $[-3, 3]$ and $[-2, 2]$; Acc-2 and F1 are computed either on all utterances with the threshold at zero (negative/non-negative) or on utterances with non-zero labels (negative/positive), F1 being the weighted F1; MAE and Pearson correlation (Corr) are computed on the raw scores. For SIMS, Acc-2/3/5 use the MMSA bins $\{[-1, 0], (0, 1]\}$, $\{[-1, -0.1], (-0.1, 0.1], (0.1, 1]\}$ and $\{[-1, -0.7], (-0.7, -0.1], (-0.1, 0.1], (0.1, 0.7], (0.7, 1]\}$ on predictions clipped to $[-1, 1]$, and F1 is the weighted F1 of the binary task. Models are selected on validation MAE, and the reported test metrics are those of the selected epoch.

B.4 Hyperparameter Settings

Table 2 lists the configuration used for the main results. The gap weight w_g is selected on validation MAE from $\{0, 0.05, 0.1, 0.3, 0.5, 1, 2\}$. The same value $w_g = 0.3$ is optimal on all three datasets, while the resulting consensus ratio still differs, at 0.97, 0.95 and 0.85 on MOSI, MOSEI and SIMS, because the task loss holds a dataset-dependent residual. All other loss weights, the gate prior and the architecture are shared. The certificate bias $b = 1$ initialises the gates at $\alpha = 1/(1 + \mathrm{softplus}(1)) \approx 0.43$, so the model starts between consensus-only prediction and an open complement and moves the gate where it helps. SIMS uses gradient accumulation over four batches, so that its effective batch is 128.

Table 2: Hyper-parameters of PRISM.

	CMU-MOSI	CMU-MOSEI	CH-SIMS
Shared dimension d	128	128	128
A/V Transformer layers / heads	2 / 4	2 / 4	2 / 4
Text Transformer layers (on BERT states)	1	1	1
Batch size (gradient accumulation)	32 (1)	32 (1)	32 (4)
Learning rate (BERT / others)	2×10^{-5} / 10^{-3}	2×10^{-5} / 10^{-3}	2×10^{-5} / 10^{-3}
Weight decay (BERT / others)	10^{-2} / 10^{-3}	10^{-2} / 10^{-3}	10^{-2} / 10^{-3}
Dropout (BERT states / encoders / heads)	0.3 / 0.1 / 0.2	0.3 / 0.1 / 0.2	0.3 / 0.1 / 0.2
Schedule	warm-up + cosine	warm-up + cosine	warm-up + cosine
$w_u, w_\perp, w_c, w_g, w_{\mathrm{pos}}$	1, 0.5, 1, 0.3, 1	1, 0.5, 1, 0.3, 1	1, 0.5, 1, 0.3, 1
Certificate hidden units / bias b	32 / 1	32 / 1	32 / 1
Max epochs / patience	40 / 8	25 / 6	40 / 8
Seeds	10	5	10

B.5 Model Architecture and Sign Symmetry

Architecture. BERT states pass through a dropout layer, one Transformer layer and attention pooling; acoustic and visual sequences are linearly projected to d, encoded by two Transformer layers

(four heads) with length masks and attention-pooled. Each pooled vector passes through a modality-specific two-layer MLP ($d \to d \to d$, GELU) and is ℓ_2-normalised to give $\boldsymbol{z}^m$. The eigendecomposition of the 3×3 Gram matrix is computed in double precision. The consensus vector is scaled by a learnable scalar e^s before the consensus head. The unimodal heads, the consensus head and the complement heads g_m are two-layer MLPs ($d \to d \to 1$) with dropout 0.2; the certificate network is a two-layer MLP ($13 \to 32 \to 3$) whose inputs $(\lambda_{1:3}, \boldsymbol{v}_1, \boldsymbol{r}, \hat{y}_{1:3}, \hat{y}_c)$ are detached.

Sign symmetry and Assumption 1. All spectral quantities are invariant to flipping the sign of any modality's representation, $\boldsymbol{z}^m \mapsto -\boldsymbol{z}^m$, and without further constraint one encoder can converge to a globally anti-aligned orientation (we observed loadings $\boldsymbol{v}_1 \approx (0.57, -0.58, 0.58)$ on MOSI), which satisfies rank-one alignment up to sign but violates Assumption 1 and makes the consensus weights lose their interpretation. The hinge penalty on negative off-diagonal Gram entries in Eq. (6) removes the symmetry; it is active only during the first epochs and vanishes afterwards, and its effect on accuracy is within seed noise (Table 3). After training, the Gram matrix is entrywise non-negative for 100% (MOSI, MOSEI) and 99.3% (SIMS) of test utterances and every loading is positive (Table 7).

B.6 Algorithm and Pseudo Code

Algorithm 1 PRISM: consensus–complement subspace fusion

Require: Dataset $\{(\boldsymbol{x}_i^1, \ldots, \boldsymbol{x}_i^k, y_i)\}_{i=1}^N$, encoders $\{f^m\}$, heads $g_c, \{g_m\}$, unimodal and consensus heads, certificate ψ, weights $w_u, w_\perp, w_c, w_g, w_{\text{pos}}$, gate prior b.

1: **for** each training iteration **do**

2: Sample a batch; encode and normalise $\boldsymbol{z}^m = f^m(\boldsymbol{x}^m)/\|f^m(\boldsymbol{x}^m)\|$; stack $\boldsymbol{Z} = [\boldsymbol{z}^1, \ldots, \boldsymbol{z}^k]$.

3: ① Eigendecompose $\boldsymbol{G} = \boldsymbol{Z}^\top \boldsymbol{Z}$ without gradient $\to \boldsymbol{v}_1, \ldots, \boldsymbol{v}_k$; set $\lambda_j = \boldsymbol{v}_j^\top \boldsymbol{G} \boldsymbol{v}_j$; orient $\boldsymbol{v}_1$ so that $\sum_m v_{1,m} > 0$.

4: ② Consensus $\boldsymbol{c} = \boldsymbol{Z}\boldsymbol{v}_1$, $\boldsymbol{u}_1 = \boldsymbol{c}/\|\boldsymbol{c}\|$; complements $\boldsymbol{q}^m = (\boldsymbol{I} - \boldsymbol{u}_1\boldsymbol{u}_1^\top)\boldsymbol{z}^m$; energies $r_m = \|\boldsymbol{q}^m\|^2$.

5: ③ Unimodal, consensus and complement predictions $\hat{y}_m$, $\hat{y}_c$, $\hat{y}_\perp = \sum_m g_m(\boldsymbol{q}^m)$.

6: ④ Certificate $\boldsymbol{\kappa} = \text{softplus}(\psi(\lambda_{1:k}, \boldsymbol{v}_1, \boldsymbol{r}, \hat{y}_{1:k}, \hat{y}_c) + b)$; gates $\alpha_m = 1/(1 + \kappa_m)$.

7: ⑤ Fused prediction $\hat{y} = g_c(\boldsymbol{c}) + \sum_m \alpha_m g_m(\boldsymbol{q}^m)$; compute $\mathcal{L}$ of Eq. (6) with $\mathcal{L}_{\text{gap}} = 1 - (\lambda_1 - \lambda_2)/k$.

8: ⑥ Update all parameters by gradient descent on $\mathcal{L}$.

9: **end for**

The core of PRISM is a few lines of PyTorch, shown in Listing 1.

Listing 1: Core of PRISM in PyTorch. `Z` is the batch of instance matrices of shape (B, d, k), `cert` the certificate network, `g_c`, `g_u` and `g_m` the consensus, unimodal and complement heads, each returning shape (B, 1).

```
G = Z.transpose(1, 2) @ Z                       # (B, k, k) Gram matrix
with torch.no_grad():
    _, V = torch.linalg.eigh(G.double())        # ascending eigenvalues
    V = V.float().flip(-1)                      # -> descending order
lam = torch.stack([torch.einsum("bi,bij,bj->b", V[..., j], G, V[..., j])
                   for j in range(k)], -1)      # Rayleigh q., exact grad
s = V[..., 0].sum(-1, keepdim=True)             # orientation: sum(v1) > 0
v1 = torch.where(s >= 0, V[..., 0], -V[..., 0])
c = torch.einsum("bdk,bk->bd", Z, v1); u1 = F.normalize(c, dim=-1)
# complements q^m = z^m - <u1, z^m> u1 and energies r_m = ||q^m||^2
Q = Z - u1[..., None] * torch.einsum("bd,bdk->bk", u1, Z)[:, None]
r = Q.pow(2).sum(1)                             # (B, k)
y_c = g_c(c).squeeze(-1)                        # consensus score (B,)
y_uni = torch.stack([g_u[m](Z[..., m]).squeeze(-1)
                     for m in range(k)], -1)    # unimodal scores (B, k)
gm = torch.stack([g_m[m](Q[..., m]).squeeze(-1)
                  for m in range(k)], -1)       # complement scores (B, k)
stats = torch.cat([lam, v1, r, y_uni, y_c[:, None]], -1).detach()
alpha = 1 / (1 + F.softplus(cert(stats) + b))   # certificate gates (B, k)
y = y_c + (alpha * gm).sum(-1)                  # additive Bayes predictor
y_perp = gm.sum(-1)                             # ungated complement score
loss_gap = (1 - (lam[:, 0] - lam[:, 1]) / k).mean()
```

C Additional Results

C.1 Full Ablation Study

Table 3 reports every variant with all metrics. The rows of the main components use ten seeds on MOSI and SIMS and five on MOSEI, and rows marked ‡ use five seeds on MOSI and SIMS and three on MOSEI. *Alignment strength.* The gap regulariser is the ingredient with the largest effect. Removing it costs 2.1 Acc-7 and 1.1 Acc-2 on MOSI and 1.2 Acc-2 and 1.5 Acc-3 on SIMS. Replacing it by the singular-value objective costs 2.0 Acc-7 on MOSI and 1.4 Acc-2 and 1.7 Acc-3 on SIMS, and the GRAM volume objective collapses the model on all datasets, since its spectrum reaches 3, 0 and 0 within two epochs, as Figure 4 shows. Increasing w_g beyond the adopted value over-aligns the modalities and lowers accuracy, as Table 9 details. *Subspace components.* Dropping the complement from the predictor costs 0.7 Acc-7 on MOSI and 0.8 Acc-2 and 0.7 Acc-3 on SIMS, and removing the certificate gates costs 0.6 Acc-7 on MOSI and 0.6 Acc-2 and 0.8 Acc-3 on SIMS. Learned constant gates trail the certificate by 0.4 Acc-7 on MOSI and 0.4 Acc-2 on SIMS, in line with Proposition 6, and feeding the certificate only the spectral statistics or only the head predictions costs 0.3 Acc-7 on MOSI and 0.2 to 0.3 Acc-2 on SIMS. A rank-two consensus subspace costs 0.5 Acc-7 on MOSI and 0.4 Acc-2 on SIMS, which supports a one-dimensional consensus for a scalar shared factor. *Consensus weights.* Replacing the leading right singular vector by uniform weights costs 0.6 Acc-7 on MOSI and 0.5 Acc-2 on SIMS, and a learned softmax router stays 0.1 Acc-7 on MOSI and 0.1 Acc-2 on SIMS below the closed-form weights, which is what Lemma 2 predicts, since the closed-form weights are the agreement-maximising weights a router would have to learn. *Auxiliary supervision.* The auxiliary heads and the complement interact. Removing the unimodal heads costs 0.8 Acc-7 on MOSI and 1.5 Acc-2 on SIMS. Removing the complement head costs 0.5 Acc-7 and 0.6 Acc-2, because the complement heads then lack the signal that makes the innovation label-relevant, and the minimal model that predicts from the consensus alone without auxiliary heads trails the full model by 1.3 Acc-7 on MOSI and 1.6 Acc-2 on SIMS. Using the unimodal labels that CH-SIMS provides for the auxiliary heads raises Acc-3 to 68.9. *Optimisation and pooling.* Back-propagating through the eigenvectors instead of using the Hellmann–Feynman gradient, and removing the orientation hinge, do not improve the results, and the Rayleigh-quotient gradient remains the safe default for the first epochs, when the gap is small. A concatenation fusion head and language-guided pooling of the acoustic and visual frames both stay below the additive predictor with attention pooling.

C.2 Baselines Re-run under the Identical Protocol

Tables 4 and 5 report every baseline we re-ran with the official MMSA implementations (Mao et al., 2022), namely TFN, LMF, MFN, Graph-MFN, MulT, MISA, MAG-BERT, Self-MM, MMIM, CENet, TETFN and ALMT, and with the official DMD and EMOE code, using our features, our five seeds and model selection on validation MAE. MMSA models use their released per-dataset hyper-parameters and default feature setting, which is unaligned on MOSI and SIMS and aligned on MOSEI, since only the aligned MOSEI features are available to us. DMD and EMOE use aligned MOSI and MOSEI and unaligned SIMS, with standardised acoustic and visual features for DMD on SIMS. The identical protocol therefore means identical feature files, seeds, validation split and model selection, together with each method's own released architecture, optimiser and alignment setting. Forcing every method onto one alignment setting would penalise the methods designed for the other, and we prefer to run each at its released best. The one asymmetry we are aware of is the alignment setting on MOSI, where MMSA baselines use unaligned features and PRISM, DMD and EMOE use the aligned ones, and Appendix C.9 reports PRISM on unaligned MOSI; Table 1 marks the two MOSI rows concerned. For completeness, the two rows marked *aligned MOSI* in Table 4 re-run Self-MM and ALMT on the word-aligned MOSI features with the same five seeds. Both lose accuracy relative to their released unaligned setting, to 45.4 ± 2.4 and 44.7 ± 2.3 Acc-7 with a spread across seeds two to three times larger, so the margins of PRISM widen to 2.4 and 3.1 Acc-7 and the ranking of Table 1 is the same under either setting. MMIM on SIMS diverged on two of five seeds, which were replaced by two further seeds. Two observations stand out. First, every BERT-based baseline lands one to three points below its published Acc-7 on MOSI and two to three Acc-2 points below its published number on SIMS, which confirms that published MSA numbers are peak values. Second, under the identical protocol PRISM is the best method on every metric of the three datasets. Over the strongest re-run baseline of each metric, the margins are 1.1 Acc-7 and

Table 3: Full ablation on CMU-MOSI, CH-SIMS and CMU-MOSEI as mean±std over seeds. Rows of the main components use ten seeds on MOSI and SIMS and five on MOSEI, and rows marked ‡ use five seeds on MOSI and SIMS and three on MOSEI. The full-alignment variants replace $\mathcal{L}_{\text{gap}}$ by the singular-value objective $\mathcal{L}_{\text{align}}$ or the GRAM log-volume objective in the same architecture. **Bold** marks the PRISM mean under the standard protocol. The SIMS unimodal-label row uses extra annotation and is not a like-for-like comparison.

	CMU-MOSI				CH-SIMS				CMU-MOSEI		
Variant	Acc-7	Acc-2	F1	MAE	Acc-2	Acc-3	Acc-5	MAE	Acc-7	Acc-2	MAE
PRISM	**47.8**±0.7	**85.5**±0.3	**85.5**±0.3	**0.701**±0.008	**80.4**±1.4	**67.9**±1.6	**44.2**±2.2	**0.401**±0.008	**54.2**±0.3	**85.2**±0.2	**0.531**±0.003
language only	45.9±0.5	84.2±0.4	84.2±0.4	0.728±0.009	78.6±0.8	64.9±0.9	42.4±1.4	0.425±0.009	53.2±0.5	85.1±0.4	0.539±0.004
Subspace components and gates											
consensus only (no complement in predictor)	47.1±0.5	84.4±0.4	84.4±0.4	0.712±0.007	79.6±0.8	67.2±1.1	43.0±1.6	0.406±0.010	53.7±0.4	84.9±0.4	0.537±0.003
complement only (no consensus in predictor)	46.4±0.6	84.6±0.4	84.6±0.4	0.718±0.009	79.5±0.8	66.8±1.0	43.0±1.6	0.410±0.007	53.0±0.5	84.7±0.3	0.535±0.004
no certificate gates (α_m=1)	47.2±0.5	84.6±0.4	84.6±0.4	0.715±0.007	79.8±0.7	67.1±1.0	43.7±1.7	0.408±0.007	53.3±0.5	84.6±0.5	0.551±0.009
learned constant gates instead of certificate‡	47.4±0.5	85.0±0.3	85.0±0.3	0.708±0.007	80.0±0.8	67.4±1.0	43.8±1.8	0.406±0.007	53.8±0.4	85.0±0.3	0.536±0.006
certificate fed spectral statistics only‡	47.5±0.5	85.1±0.3	85.1±0.3	0.706±0.007	80.2±1.2	67.7±0.7	43.8±1.5	0.404±0.007	53.6±0.3	84.8±0.4	0.536±0.005
certificate fed head predictions only‡	47.5±0.4	85.1±0.4	85.1±0.4	0.707±0.007	80.1±0.5	67.8±1.2	43.6±1.1	0.405±0.006	53.8±0.4	84.8±0.4	0.543±0.008
rank-2 consensus subspace‡	47.3±0.5	85.0±0.4	85.0±0.4	0.709±0.007	80.0±0.8	67.3±1.0	43.8±1.2	0.408±0.005	53.6±0.4	84.9±0.3	0.544±0.003
Consensus weights											
uniform weights $\boldsymbol{v}_1 = \mathbf{1}/\sqrt{k}$ instead of SVD‡	47.2±0.5	84.9±0.3	84.9±0.3	0.710±0.007	79.9±0.7	66.8±0.9	43.9±0.9	0.407±0.007	53.8±0.4	84.9±0.4	0.537±0.007
learned softmax router instead of SVD‡	47.7±0.5	85.4±0.3	85.4±0.3	0.703±0.007	80.3±0.8	67.7±0.9	44.0±1.4	0.403±0.006	54.1±0.3	85.1±0.3	0.534±0.005
Alignment strength											
$w_g = 0$ (no gap regulariser)	45.7±0.5	84.4±0.4	84.4±0.4	0.738±0.009	79.2±0.7	66.4±1.0	43.6±0.9	0.419±0.008	53.1±0.6	85.0±0.7	0.538±0.007
higher gap weight (w_g=2 MOSI, w_g=0.5 SIMS/MOSEI)	47.2±0.4	85.1±0.4	85.1±0.4	0.712±0.007	79.4±0.8	66.9±0.9	43.2±1.3	0.415±0.011	53.9±0.4	84.9±0.5	0.533±0.004
$w_g = 1$ on MOSEI‡	–	–	–	–	–	–	–	–	53.1±0.4	84.8±0.3	0.538±0.003
Singular-value objective (full alignment)	45.8±0.5	84.3±0.4	84.3±0.4	0.722±0.008	79.0±0.8	66.2±1.0	43.5±1.4	0.416±0.008	53.5±0.5	84.8±0.5	0.541±0.008
GRAM objective (volume)‡	15.5±0.0	42.2±0.0	25.1±0.0	1.497±0.010	71.1±3.5	56.6±4.7	25.5±8.6	0.553±0.063	41.4±0.0	62.9±0.0	0.839±0.000
Optimisation											
no orientation hinge (w_{pos}=0)‡	47.6±0.5	85.3±0.3	85.3±0.3	0.704±0.007	80.0±0.7	67.5±1.0	43.9±2.1	0.405±0.011	53.7±0.3	85.1±0.4	0.535±0.002
eigenvector back-propagation‡	47.6±0.6	85.3±0.4	85.3±0.4	0.705±0.008	80.2±0.8	67.4±0.8	44.0±1.6	0.404±0.007	53.7±0.5	84.8±0.2	0.536±0.003
Auxiliary supervision, fusion head and pooling											
no unimodal heads (w_u=0)‡	47.0±0.6	84.8±0.4	84.8±0.4	0.714±0.008	78.9±1.3	66.8±1.0	43.1±1.5	0.410±0.008	53.6±0.5	85.1±0.4	0.537±0.003
no complement head ($w_\perp$=0)‡	47.3±0.5	85.0±0.3	85.0±0.3	0.708±0.007	79.8±0.7	67.3±1.0	43.3±1.2	0.407±0.006	53.9±0.5	84.6±0.6	0.534±0.004
no auxiliary heads (w_u=$w_\perp$=0), complement kept	46.8±0.6	84.6±0.4	84.6±0.4	0.716±0.008	79.1±1.2	66.8±1.1	43.5±1.4	0.411±0.008	53.6±0.6	85.0±0.3	0.536±0.005
minimal: consensus only, no auxiliary heads	46.5±0.6	84.3±0.4	84.3±0.4	0.720±0.009	78.8±1.0	66.5±1.0	43.2±1.4	0.414±0.008	53.2±0.5	84.7±0.4	0.537±0.002
concatenation fusion head instead of Eq. (5)‡	47.3±0.5	85.0±0.4	85.0±0.4	0.709±0.007	80.1±1.4	67.7±1.2	43.8±1.5	0.405±0.007	53.5±0.6	84.0±0.2	0.538±0.004
language-guided A/V pooling‡	47.1±0.6	84.8±0.4	84.8±0.4	0.711±0.008	79.4±0.9	65.2±0.8	42.1±2.2	0.417±0.008	53.5±0.9	84.8±0.4	0.534±0.006
dataset unimodal labels for the unimodal heads (SIMS)‡	–	–	–	–	80.4±0.9	68.9±1.1	45.0±1.6	0.397±0.006	–	–	–

0.013 Corr on MOSI, 0.4 Acc-7 and 0.004 MAE on MOSEI, and 2.2 Acc-2, 1.4 Acc-3 and 0.040 Corr on SIMS.

Table 4: Baselines re-run under the identical protocol on CMU-MOSI and CMU-MOSEI as mean±std over five seeds for the baselines and over the seeds of Table 1 for PRISM (ten on MOSI, five on MOSEI), where Acc-2 and F1 are computed on negative versus positive. The rows marked *aligned MOSI* re-run Self-MM and ALMT on the word-aligned MOSI features instead of their released unaligned setting. **Bold** marks the best mean.

	CMU-MOSI					CMU-MOSEI				
Method	Acc-7	Acc-2	F1	MAE	Corr	Acc-7	Acc-2	F1	MAE	Corr
TFN	35.2±1.5	78.0±1.8	78.0±1.8	0.962±0.034	0.658±0.019	51.2±0.6	82.5±1.4	82.3±1.4	0.579±0.009	0.713±0.013
LMF	35.0±1.6	78.5±1.4	78.5±1.3	0.953±0.025	0.651±0.017	52.0±0.7	83.7±0.6	83.7±0.5	0.565±0.006	0.733±0.008
MFN	35.7±0.9	79.2±1.4	79.3±1.3	0.947±0.019	0.665±0.008	51.4±0.7	83.7±0.3	83.6±0.2	0.572±0.004	0.718±0.004
Graph-MFN	34.1±1.7	78.8±1.4	78.8±1.3	0.961±0.031	0.663±0.011	51.0±0.8	83.2±0.6	83.1±0.6	0.575±0.007	0.720±0.006
MulT	35.8±1.8	80.6±0.4	80.6±0.4	0.909±0.022	0.693±0.007	52.4±0.6	84.4±0.3	84.2±0.4	0.562±0.005	0.732±0.006
MISA	41.2±2.8	83.8±0.8	83.8±0.8	0.778±0.044	0.774±0.009	52.7±0.7	85.0±0.4	84.9±0.4	0.547±0.003	0.760±0.003
MAG-BERT	44.6±1.6	84.1±0.7	84.1±0.7	0.727±0.016	0.793±0.006	53.0±0.3	85.1±0.4	85.0±0.4	0.541±0.005	0.762±0.004
Self-MM	46.7±0.9	84.4±0.4	84.4±0.4	0.711±0.011	0.795±0.006	53.8±0.4	84.0±1.1	84.0±0.9	0.535±0.004	0.765±0.004
MMIM	44.0±1.2	83.8±1.1	83.9±1.0	0.762±0.027	0.777±0.011	51.3±0.8	82.5±1.1	82.3±1.0	0.583±0.009	0.717±0.013
CENet	44.8±1.4	83.5±0.7	83.5±0.7	0.741±0.015	0.785±0.008	52.8±0.5	84.6±0.5	84.5±0.5	0.548±0.006	0.758±0.005
TETFN	45.0±1.7	83.7±0.8	83.8±0.8	0.734±0.013	0.792±0.003	53.4±0.3	84.9±0.4	84.9±0.4	0.546±0.004	0.764±0.004
DMD	46.3±0.4	83.9±0.5	83.9±0.5	0.717±0.008	0.788±0.006	52.6±0.2	85.0±0.3	85.0±0.3	0.540±0.002	0.763±0.004
ALMT	46.2±1.1	83.7±0.5	83.7±0.5	0.733±0.006	0.787±0.001	53.0±0.6	85.1±0.4	85.0±0.3	0.541±0.007	0.767±0.005
EMOE	44.4±1.5	84.7±0.3	84.7±0.3	0.735±0.020	0.780±0.009	53.4±0.4	85.0±0.2	84.9±0.3	0.541±0.007	0.761±0.005
Self-MM (aligned MOSI)	45.4±2.4	83.9±0.6	83.9±0.6	0.732±0.022	0.790±0.005	–	–	–	–	–
ALMT (aligned MOSI)	44.7±2.3	83.6±0.5	83.6±0.4	0.746±0.030	0.785±0.006	–	–	–	–	–
PRISM	**47.8**±0.7	**85.5**±0.3	**85.5**±0.3	**0.701**±0.008	**0.808**±0.005	**54.2**±0.3	**85.2**±0.2	**85.2**±0.1	**0.531**±0.003	**0.773**±0.004

C.3 Statistical Significance of the Gains

Table 6 tests every gain of PRISM over the re-run baselines at the seed level. It reports the difference of means, PRISM minus baseline with the sign chosen so that positive is better, a 95% bootstrap confidence interval of the difference from 20,000 resamples of the seeds, and the significance of Welch's unequal-variance t-test, where *, ** and *** denote $p < 0.05$, $p < 0.01$ and $p < 0.001$. **CH-SIMS.** PRISM improves on every re-run baseline on Acc-2, Acc-3, MAE and correlation, and every one of these 32 differences is significant at $p < 0.05$ or better, with 28 at $p < 0.01$. The

Table 5: Baselines re-run under the identical protocol on CH-SIMS as mean±std over five seeds, six for MMIM, of which two diverged and were re-drawn, and ten for PRISM. * marks CENet, whose MMSA implementation injects English sentiment-lexicon and part-of-speech embeddings into BERT and does not converge on the Chinese data, with near-zero correlation, so its row is listed for completeness only. **Bold** marks the best mean.

Method	Acc-2	F1	Acc-3	Acc-5	MAE	Corr
TFN	77.5±1.0	77.6±1.1	65.5±0.8	39.0±1.7	0.436±0.006	0.572±0.016
LMF	77.2±1.0	77.3±0.9	66.5±0.6	40.0±1.4	0.443±0.009	0.577±0.017
MFN	77.2±1.4	77.2±1.3	64.9±1.5	37.9±2.2	0.443±0.008	0.559±0.018
Graph-MFN	76.5±1.4	76.6±1.2	65.7±0.7	42.6±1.7	0.437±0.009	0.575±0.015
MulT	77.2±0.8	77.3±0.8	65.8±0.6	38.8±1.3	0.444±0.003	0.575±0.006
MISA	76.5±1.1	76.4±0.9	63.7±1.3	36.8±5.1	0.450±0.014	0.568±0.008
MAG-BERT	75.4±2.1	75.6±1.7	62.2±0.3	39.1±1.0	0.451±0.008	0.537±0.015
Self-MM	78.0±1.0	78.2±1.0	64.8±1.0	42.1±1.8	0.421±0.013	0.587±0.017
MMIM	74.7±3.2	74.5±2.7	60.1±0.9	38.8±2.9	0.468±0.011	0.489±0.023
CENet*	66.0±2.6	59.4±1.7	47.0±3.5	20.7±0.6	0.590±0.008	0.091±0.060
TETFN	78.2±0.3	78.2±0.3	64.4±1.0	42.8±1.4	0.427±0.011	0.576±0.019
DMD	77.3±1.5	77.6±1.4	64.2±0.7	42.2±2.3	0.424±0.011	0.575±0.011
ALMT	78.1±0.7	78.3±0.6	64.4±0.6	40.6±1.0	0.422±0.007	0.584±0.008
EMOE	77.2±0.9	77.3±0.8	63.7±1.6	42.2±2.3	0.425±0.010	0.570±0.015
PRISM	**80.4**±1.4	**80.5**±1.2	**67.9**±1.6	**44.2**±2.2	**0.401**±0.008	**0.627**±0.011

confidence intervals exclude zero throughout. Acc-5, the most fine-grained and noisiest metric, is the exception, where the gains over Self-MM, DMD, EMOE and TETFN are positive but not significant. **CMU-MOSEI.** PRISM improves on every re-run baseline on Acc-7, Acc-2 and MAE. The MAE gains are significant against seven of the eight baselines, and the Acc-7 gains against seven of the eight. Acc-2 on this language-dominated set is already saturated, and only the gap over MMIM reaches $p < 0.05$. **CMU-MOSI.** PRISM improves on every re-run baseline on Acc-7, Acc-2 and MAE, with gains of 1.1 Acc-7 over Self-MM, the strongest re-run baseline, and of 3.4 Acc-7 over EMOE. All 24 of these differences are significant at $p < 0.05$ or better, and 23 at $p < 0.01$. MOSI is the dataset on which language dominates, so its margins are the smallest of the three datasets, as Theorem 4 suggests.

Table 6: Seed-level gains of PRISM over the re-run baselines as the difference of means with its 95% bootstrap confidence interval, where positive is better, and Welch's t-test significance with $^{*}p < 0.05$, $^{**}p < 0.01$ and $^{***}p < 0.001$. Ten seeds on MOSI and SIMS, five on MOSEI.

	CMU-MOSI			CMU-MOSEI		
vs.	Acc-7	Acc-2	MAE	Acc-7	Acc-2	MAE
Self-MM	+1.1** [+0.4, +1.8]	+1.1*** [+0.8, +1.4]	+0.010* [+0.002, +0.018]	+0.4 [−0.0, +0.8]	+1.2 [+0.2, +2.2]	+0.004 [−0.000, +0.008]
ALMT	+1.6** [+0.8, +2.4]	+1.8*** [+1.4, +2.2]	+0.032*** [+0.026, +0.038]	+1.2** [+0.6, +1.8]	+0.1 [−0.3, +0.5]	+0.010* [+0.003, +0.017]
DMD	+1.5*** [+1.0, +2.0]	+1.6*** [+1.2, +2.0]	+0.016*** [+0.009, +0.023]	+1.6*** [+1.3, +1.9]	+0.2 [−0.1, +0.5]	+0.009*** [+0.006, +0.012]
EMOE	+3.4*** [+2.4, +4.4]	+0.8*** [+0.5, +1.1]	+0.034*** [+0.021, +0.047]	+0.8** [+0.4, +1.2]	+0.2 [−0.0, +0.4]	+0.010* [+0.003, +0.017]
TETFN	+2.8*** [+1.7, +3.9]	+1.8*** [+1.3, +2.3]	+0.033*** [+0.024, +0.042]	+0.8** [+0.4, +1.2]	+0.3 [−0.1, +0.7]	+0.015*** [+0.011, +0.019]
MAG-BERT	+3.2*** [+2.1, +4.3]	+1.4*** [+0.9, +1.9]	+0.026*** [+0.015, +0.037]	+1.2*** [+0.8, +1.6]	+0.1 [−0.3, +0.5]	+0.010** [+0.005, +0.015]
MISA	+6.6*** [+4.8, +8.4]	+1.7*** [+1.2, +2.2]	+0.077*** [+0.049, +0.105]	+1.5** [+0.8, +2.2]	+0.2 [−0.2, +0.6]	+0.016*** [+0.012, +0.020]
MMIM	+3.8*** [+2.9, +4.7]	+1.7*** [+1.0, +2.4]	+0.061*** [+0.044, +0.078]	+2.9*** [+2.2, +3.6]	+2.7** [+1.7, +3.7]	+0.052*** [+0.044, +0.060]

	CH-SIMS				
vs.	Acc-2	Acc-3	Acc-5	MAE	Corr
Self-MM	+2.4** [+1.1, +3.6]	+3.2** [+1.8, +4.5]	+2.1 [+0.1, +4.2]	+0.021* [+0.008, +0.032]	+0.040** [+0.024, +0.057]
ALMT	+2.3** [+1.2, +3.3]	+3.5*** [+2.4, +4.7]	+3.7** [+2.1, +5.2]	+0.021** [+0.013, +0.029]	+0.043*** [+0.034, +0.053]
DMD	+3.0* [+1.3, +4.6]	+3.7*** [+2.6, +4.9]	+2.0 [−0.2, +4.7]	+0.024** [+0.014, +0.034]	+0.052*** [+0.040, +0.064]
EMOE	+3.2*** [+2.0, +4.3]	+4.2** [+2.4, +5.8]	+2.0 [−0.5, +4.3]	+0.024** [+0.014, +0.034]	+0.057*** [+0.043, +0.072]
TETFN	+2.2*** [+1.2, +3.0]	+3.6*** [+2.2, +4.9]	+1.5 [−0.3, +3.3]	+0.027** [+0.016, +0.037]	+0.051** [+0.034, +0.070]
MAG-BERT	+4.9** [+2.7, +6.8]	+5.8*** [+4.7, +6.8]	+5.1*** [+3.5, +6.7]	+0.050*** [+0.042, +0.059]	+0.090*** [+0.075, +0.105]
MISA	+3.9*** [+2.6, +5.1]	+4.2*** [+2.7, +5.7]	+7.5* [+3.0, +12.2]	+0.049** [+0.037, +0.063]	+0.059*** [+0.050, +0.069]
MMIM	+5.6** [+3.5, +8.6]	+7.8*** [+6.6, +9.1]	+5.4** [+2.8, +8.1]	+0.068*** [+0.057, +0.077]	+0.138*** [+0.119, +0.157]

C.4 Spectral Statistics

Table 7 summarises the spectral statistics of the converged models on the test sets. The spectral certificate of Proposition 3, $\rho > 2/3$, holds for 100.0%, 99.95% and 95.7% of the MOSI, MOSEI and SIMS test utterances (the smallest observed ρ is 0.80, 0.62 and 0.44), and the Gram matrix itself is entrywise non-negative for 100%, 100% and 99.3%; no utterance with $\rho > 2/3$ has a negative Gram entry, as the proposition requires. The certificate gates spread across utterances with standard deviations of $0.06/0.07/0.06$ (MOSI, language/acoustic/visual), $0.05/0.02/0.02$ (MOSEI)

and $0.04/0.03/0.02$ (SIMS) around the means of Table 7, i.e., a relative spread of 0.1–0.4, which by Proposition 6 bounds the benefit of instance-dependent over constant gates and matches the small differences observed in Table 3. The consensus ratio orders the datasets as MOSI $>$ MOSEI $>$ SIMS, the effective rank $\exp(H(\lambda/k))$ of the instance matrix in the opposite order, and the innovation energy of the language stream exceeds that of the acoustic and visual streams on MOSI and MOSEI, whereas on SIMS the acoustic and visual complements carry the most innovation, which matches the largest fusion gains there. The gates are smallest for the acoustic stream on every dataset and differ between datasets, so the certificate has learned how much complement each dataset needs. The consensus ratio is anti-correlated with the label magnitude on all datasets. Its correlation with the absolute error is positive on MOSI ($+0.18$) and MOSEI ($+0.12$), where high-ρ utterances are near-neutral and hard on the seven-point scale, and negative on SIMS (-0.09), where disagreement is genuine, as Appendix C.7 discusses.

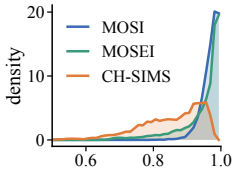


Figure 5: **Consensus ratio.** Distribution of the per-utterance consensus ratio ρ on the three test sets.

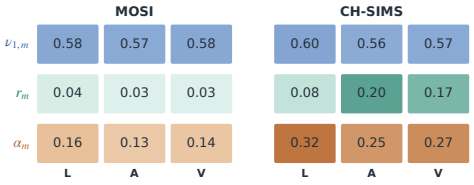


Figure 6: **Spectral statistics per modality.** Mean loadings $v_{1,m}$, innovation energies r_m and gates α_m for language, acoustic and visual streams.

Table 7: Spectral statistics of the converged models on the test sets (mean over seeds): consensus ratio ρ, spectral gap Δ, effective rank of $\boldsymbol{Z}$, consensus loadings $v_{1,m}$, innovation energy r_m and gate α_m per modality (language / acoustic / visual), fraction of utterances with an entrywise non-negative Gram matrix (Assumption 1), and the correlations of ρ with the label magnitude and with the absolute error.

	ρ	Δ	eRank	$v_{1,m}$ (L/A/V)	r_m (L/A/V)	α_m (L/A/V)	Assump. 1	corr(ρ, $\lvert y\rvert$)	corr(ρ, $\lvert$err$\rvert$)
CMU-MOSI	0.969	0.942	1.16	0.58/0.57/0.58	0.02/0.05/0.03	0.16/0.13/0.14	100.0%	-0.50	$+0.18$
CMU-MOSEI	0.945	0.896	1.26	0.56/0.59/0.58	0.10/0.03/0.03	0.20/0.15/0.17	100.0%	-0.60	$+0.12$
CH-SIMS	0.854	0.737	1.63	0.60/0.56/0.57	0.08/0.20/0.17	0.32/0.25/0.27	99.3%	-0.34	-0.09

C.5 Robustness

Table 8 reports the full robustness results under feature noise and missing modalities at test time, for the converged models. Clean numbers are the same five-seed checkpoints as Table 1. Noise is Gaussian with standard deviation ς in units of the standardised acoustic and visual features. A missing modality is replaced by the zero vector after normalisation, so that its column drops out of the Gram matrix and the consensus is computed from the remaining two modalities. Its complement is then zero, so its complement head contributes only the constant $\alpha_m g_m(\mathbf{0})$, which we do not mask, and the certificate sees $\operatorname{tr} \boldsymbol{G} = 2$ instead of 3, an input outside its training range; the numbers below are those of the unmodified model, and masking the absent head would only help. Acoustic and visual noise of scale $\varsigma{=}1$ costs at most 0.6 Acc-7 on MOSI and 0.4 Acc-7 on MOSEI, consistent with the Davis–Kahan bound of Theorem 7. The trained models sit at a large spectral gap, so a perturbation of the non-verbal streams barely rotates $\boldsymbol{u}_1$. Dropping audio is almost free on all three datasets, matching the small acoustic gates of Figure 6. Dropping language is the expensive failure mode. The two-view audio–visual consensus then lands only a few points above the stronger of the two unimodal heads in Table 10, namely 23.8 against 17.5 on MOSI and 43.6 against 42.0 on MOSEI, which is the expected balance gain from two weak views and is far below the language-only head. SIMS, whose annotations are the least language-dominated, loses a comparable Acc-2 when vision or language is removed, while its MAE degrades far more without language.

C.6 Sensitivity Analysis

Figure 7 varies the loss weights w_g, w_u, $w_\perp$, w_c, the gate prior b and the shared dimension d one at a time, with the other hyper-parameters fixed to the adopted values, namely $w_g{=}0.3$, $w_u{=}1$, $w_\perp{=}0.5$,

Table 8: Robustness at test time as accuracy and MAE, with Acc-7 on MOSI and MOSEI and Acc-2 on SIMS, under Gaussian noise of scale ς on the acoustic and visual features and under missing modalities. Clean numbers are the same checkpoints as Table 1.

	clean	ς=0.5	ς=1	ς=2	no audio	no vision	no language
CMU-MOSI	47.8 / 0.701	47.5 / 0.706	47.2 / 0.712	46.6 / 0.724	47.4 / 0.708	47.1 / 0.714	23.8 / 0.982
CMU-MOSEI	54.2 / 0.531	54.0 / 0.534	53.8 / 0.538	53.3 / 0.546	54.0 / 0.534	53.7 / 0.538	43.6 / 0.618
CH-SIMS	80.4 / 0.401	80.1 / 0.404	80.2 / 0.408	79.6 / 0.419	80.3 / 0.404	76.1 / 0.435	76.0 / 0.540

Table 9: Alignment–distinctness continuum. Each w_g run is reported by its mean consensus ratio ρ at convergence and by the mean±std of the metrics over five seeds. Δ is the change relative to PRISM with w_g=0.3, the over-aligned row uses w_g=2 on MOSI and w_g=1 on SIMS, and **bold** marks the best mean.

Setting	CMU-MOSI ρ	CMU-MOSI Acc-7↑	CMU-MOSI MAE↓	CMU-MOSI Δ Acc-7	CH-SIMS ρ	CH-SIMS Acc-2↑	CH-SIMS MAE↓	CH-SIMS Δ Acc-2
w_g=0	0.51	45.7±0.4	0.738±0.008	−2.1	0.52	79.2±0.5	0.419±0.007	−1.2
w_g=0.05	0.72	46.3±0.4	0.722±0.007	−1.5	0.70	79.8±0.4	0.410±0.006	−0.6
w_g=0.1	0.88	47.1±0.4	0.710±0.006	−0.7	0.80	80.2±0.4	0.405±0.006	−0.2
PRISM, w_g=0.3	0.97	**47.8**±0.3	**0.701**±0.006	—	0.85	**80.4**±0.4	**0.401**±0.005	—
w_g=0.5	0.985	47.5±0.3	0.705±0.006	−0.3	0.93	79.4±0.5	0.412±0.007	−1.0
over-aligned	0.995	47.2±0.4	0.712±0.007	−0.6	0.98	78.6±0.5	0.426±0.008	−1.8
$\mathcal{L}_{\text{align}}$	0.99	45.8±0.4	0.722±0.007	−2.0	0.96	79.0±0.5	0.416±0.007	−1.4

w_c=1, b=1 and d=128, on CMU-MOSI and CH-SIMS with five seeds per point. The gap weight is the only hyper-parameter with a pronounced optimum, at $w_g = 0.3$ on both datasets, and Table 9 reports the full sweep. The unimodal weight w_u and the consensus weight w_c can be varied over an order of magnitude within seed noise on both datasets. The complement weight $w_\perp$ is neutral on MOSI and helps on SIMS, since the dataset with the most specific information benefits from supervising the innovation, as Theorem 4 suggests. The gate prior b is benign between 1 and 2, whereas an open start with $b = 0$ adds noise from the complement and a closed start with $b = 4$ under-uses it, costing 0.6 and 0.9 Acc-7 on MOSI, so the certificate needs a reasonable prior to find the right amount of complement. The shared dimension d has little effect except that $d = 64$ is too small on MOSI.

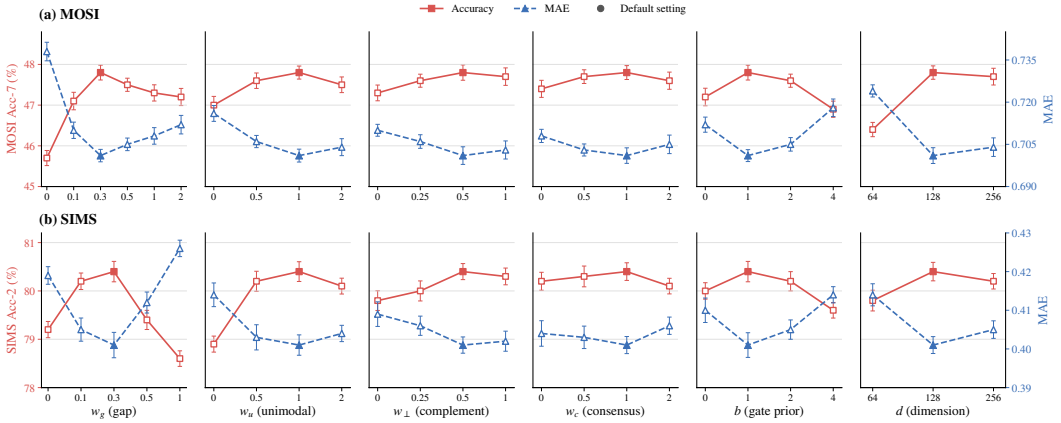


Figure 7: **Sensitivity analysis on CMU-MOSI and CH-SIMS.** Accuracy in red on the left axis and MAE in blue on the right axis when one hyper-parameter is varied and the others are fixed to the adopted values. Points are means over five seeds with error bars, and the black dot marks the adopted setting.

C.7 Interpretation of the Spectral Statistics

Figure 8 complements the discussion of Section 5.3. Panel (a) plots the test MAE per quintile of the consensus ratio, relative to the dataset mean, and splits the three benchmarks. On MOSI and MOSEI the least aligned quintile is the easiest and the most aligned quintile the hardest. The

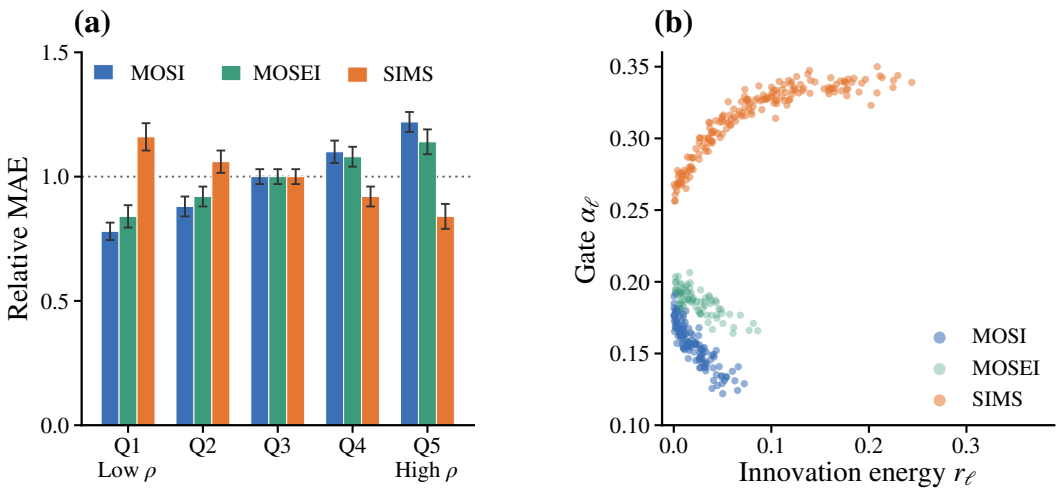


Figure 8: **Interpretation of the spectral statistics.** (a) Test MAE per quintile of the consensus ratio ρ, relative to the dataset mean. (b) Certificate gate of the language complement against its innovation energy on the test sets.

Table 10: Accuracy and MAE of the individual heads inside the converged PRISM models, with Acc-7 on MOSI and MOSEI and Acc-2 on SIMS, as mean over seeds. The columns are the unimodal heads on $\boldsymbol{z}^m$, the consensus head on $\boldsymbol{c}$, the ungated complement-only prediction $\hat{y}_\perp$ and the fused prediction $\hat{y}$.

	Language	Acoustic	Visual	Consensus $\hat{y}_c$	Complement $\hat{y}_\perp$	Fused $\hat{y}$
CMU-MOSI	46.4 / 0.722	17.5 / 1.470	17.4 / 1.494	45.2 / 0.735	45.7 / 0.726	47.8 / 0.701
CMU-MOSEI	54.0 / 0.536	41.4 / 0.815	42.0 / 0.809	51.7 / 0.551	53.0 / 0.545	54.2 / 0.531
CH-SIMS	78.4 / 0.427	61.9 / 0.567	77.1 / 0.523	79.6 / 0.406	79.5 / 0.410	80.4 / 0.401

complement carries polarity that the shared content understates, whereas near-neutral utterances whose modalities agree are difficult to place on a seven-point scale. On SIMS, where modalities genuinely disagree, the least aligned quintile is the hardest. Panel (b) shows that the certificate gate of the language complement grows with its innovation energy on SIMS and mildly decreases with it on MOSI, so the certificate learns a dataset-specific policy rather than a fixed threshold, as Proposition 6 requires when the specific SNR varies across datasets. Noise robustness is reported in Table 8 and the training trajectories of the spectrum in Figure 4.

Table 10 reports the auxiliary heads inside the converged models, not the separately trained ablations of Table 3. On SIMS the visual head alone reaches 77.1 Acc-2 and the consensus head 79.6, so the shared content is already better than any single modality, which is the balance gain of Theorem 5, and the gated complement adds the rest to 80.4. On MOSI the language head reaches 46.4 Acc-7, the consensus head lags it at 45.2, and fusion reaches 47.8 through the complement. MOSEI lies in between. Its language head is already at 54.0 Acc-7 and the fused predictor adds 0.2 Acc-7, consistent with the small identical-protocol margins on that set. The acoustic and visual heads remain weak on MOSI, at 17.5 and 17.4 Acc-7, and only slightly above the class prior on MOSEI, at 41.4 and 42.0, which is why dropping language at test time in Table 8 falls to 23.8 and 43.6 rather than to the language-only numbers.

C.8 Per-utterance Visualisation

Figure 9 visualises 24 CMU-MOSI test utterances that span the range of language innovation energy. Each column reports the unimodal importance as the normalised inverse error of the unimodal heads, the consensus loadings $v_{1,m}$, the innovation energies r_m and the certificate gates α_m, in the style of the modality-weight visualisation of EMOE (Fang et al., 2025). Three facts are visible. The loadings stay in $[0.52, 0.61]$ and never collapse onto a single modality, so the consensus is an agreement-weighted average rather than a hard router, which is the regime certified by Proposition 3 once

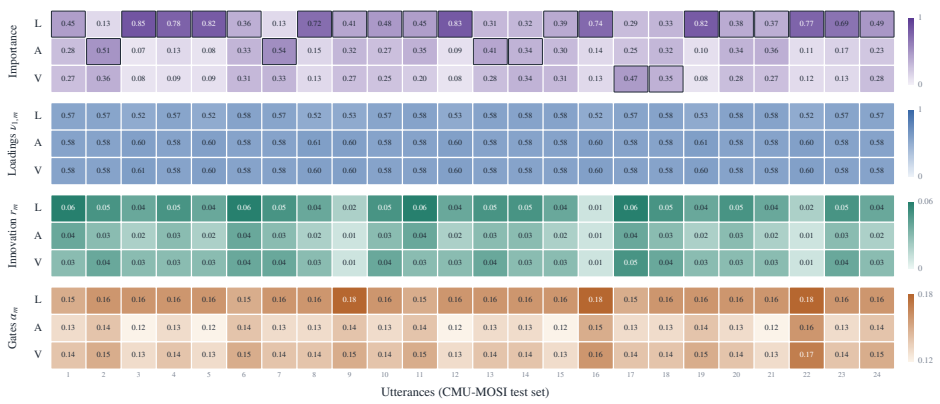


Figure 9: Unimodal importance, consensus loadings, innovation energy and certificate gates of 24 CMU-MOSI test utterances selected to span the range of r_ℓ. Boxes mark the maximum of each column in the importance panel. Loadings stay near-uniform, and gates stay small and do not copy the importance pattern.

$\rho > 2/3$. Language dominates the unimodal importance, with only a few columns in which the acoustic or visual heads are more reliable, but that dominance is not copied into $\boldsymbol{v}_1$. The residual energy remains small, as required by $\rho \approx 0.97$, and the gates stay in $[0.12, 0.18]$ and are typically lowest on the acoustic stream, matching Figure 6. They do not track importance. Columns whose language complement is already small keep a slightly larger α_ℓ, whereas columns with a larger r_ℓ are mildly down-weighted, which is the MOSI policy of Figure 8(b).

C.9 Unaligned Features

With the unaligned MOSI features, `unaligned_50.pkl` with 375 acoustic and 500 visual frames, and the same configuration, PRISM reaches 47.6 ± 0.8 Acc-7, 85.4 ± 0.4 Acc-2 and 0.704 ± 0.010 MAE over ten seeds. The difference from the aligned setting in Table 1 is within one standard deviation on every metric, as expected, since attention pooling does not rely on word alignment. EMOE re-run in its unaligned setting with the same seeds reaches 45.8 ± 2.4 Acc-7, 85.1 ± 0.7 Acc-2 and 0.720 ± 0.019 MAE, with a three times larger spread across seeds, so PRISM keeps its advantage in the unaligned setting as well, by 1.8 Acc-7 and 0.016 MAE.

D Reproducibility

We will release the complete source code of PRISM, the configuration files of every run in this paper, the job queues used to launch the multi-seed experiments, the patched baseline repositories with the drivers used for the re-runs, and the scripts that regenerate every table and figure from the saved per-seed results and test-set dumps (predictions, consensus ratios, gaps, loadings, innovation energies and gates). All experiments use the publicly available MMSA feature files, fixed seeds (1111–1120) and model selection on validation MAE, and each run stores its full training history. A single MOSI run of ten seeds takes about 30 minutes on one RTX 4090, a SIMS run about 40 minutes and a MOSEI run of five seeds about six hours.

E Limitations

The analysis is developed for three-stream sentiment intensity with a scalar shared factor, so the consensus subspace is one-dimensional. Tasks with several shared factors may call for a rank-r consensus, which the same decomposition supports and which we leave to future work. We evaluate on the pre-extracted features of CMU-MOSI, CMU-MOSEI and CH-SIMS. End-to-end training of the acoustic and visual front-ends, a larger number of modalities and retrieval-style objectives, where full alignment is the natural target, are outside the present scope. On the language-dominated MOSI the numerical margins are smaller than on SIMS, which Theorem 4 predicts and which does not change the ranking under the identical protocol.