RISK-CONTROLLED CORRECTION FOR FROZEN MULTIMODAL MODELS UNDER MISSING MODALITIES

Anonymous authors
Paper under double-blind review

ABSTRACT

A deployed multimodal predictor may face missing modalities after its weights can no longer be changed. This raises a question that training-time repair does not answer: *what part of the resulting error is still correctable from the information that remains?* We show that this part is characterized by the *Bayes gap*, the difference between the model's prediction and the Bayes prediction given the observed modalities. For cross-entropy and squared error, excess risk measures this gap in the geometry of the loss; ambiguity in the observed modalities cannot be removed. Thus, a badly degraded prediction is not necessarily a correctable one. Estimating the Bayes prediction normally requires labels. However, when a model observing more modalities is conditionally correct, averaging its outputs conditional on the remaining modalities recovers this target without retrieval labels; with the same fixed blend and neighbors, it is no worse in expected cross-entropy or Brier risk than averaging hard labels. Some such assumption is unavoidable: otherwise the target is not identifiable from unlabeled data alone. GUARD (Gated Utility-Aware Recalibration for Deployment) leaves the deployed predictor unchanged. It estimates a candidate correction by retrieval, learns from deployment-matched labeled data when correction is likely to help, and calibrates a threshold on a held-out split by conformal risk control. Under exchangeability, the marginal probability that GUARD applies a correction whose loss increase exceeds a prescribed tolerance is bounded by a preset level, regardless of the learned score. In controlled experiments, correction gain follows headroom rather than raw error; across nine frozen model–dataset pairs spanning six domains, GUARD retains most of blanket correction's gain within harm budgets that blanket correction often exceeds.

1 INTRODUCTION

Modalities can be missing at deployment. A diagnostic model trained on twelve-lead ECG (Wagner et al., 2020) may run where only a few leads are recorded (Reyna et al., 2021); an activity recognizer trained on a body-worn sensor array (Ordóñez & Roggen, 2016) may lose streams when sensors disconnect; an audio–visual model may receive only one of its two streams (Parthasarathy & Sundaram, 2020). Deployments may also fix the predictor because its weights are inaccessible (Sun et al., 2022). We assume query access to predictions and representations, without modifying the model. Output correction still requires validation and applicable change control (FDA, 2025).

Existing missing-modality methods reconstruct absent modalities (Wang et al., 2023b), distill complete-modality teachers (Li et al., 2024b; Zhuang et al., 2025b), align partial and complete representations (Lin & Hu, 2023; He et al., 2026), retrieve full-modality neighbors (Zhuang et al., 2025a; Zhi et al., 2025), or learn reliability-aware routing (Zou et al., 2026). Recent work also asks whether repair is needed at all (Chen et al., 2026; Madaan et al., 2026; Gao et al., 2026). These approaches generally modify or augment the prediction pipeline before deployment, even when the backbone remains frozen (Figure 1b). We instead study post-hoc output correction that leaves the deployed predictor unchanged (Figure 1c). This raises a deployment question: *once the model is frozen, what part of its error is still correctable from the modalities it observes?*

This paper studies a deployment regime in which the deployed predictor must remain fixed, but labeled data from the target deployment condition are available for post-hoc adaptation. We use these data to fit the correction mechanism and calibrate without modifying or retraining the predictor.

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.

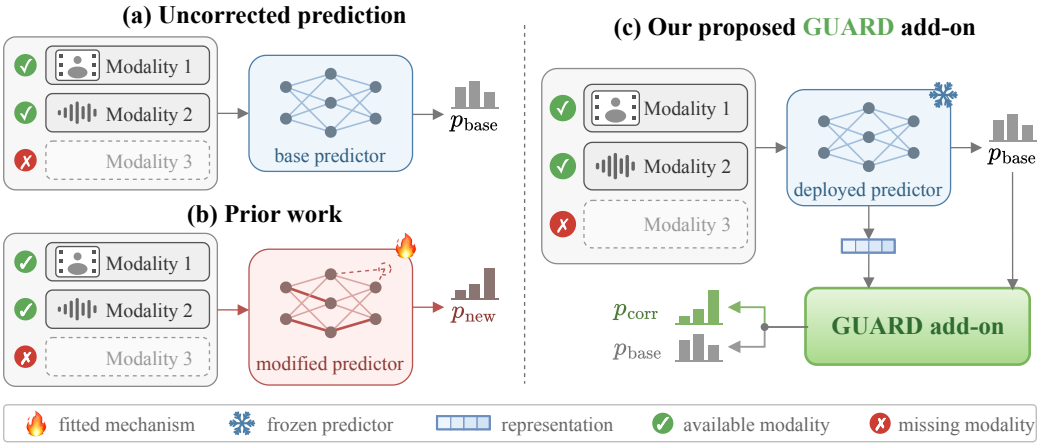

Figure 1: **Missing-modality prediction and post-hoc correction.** *(a)* Uncorrected prediction under missing modalities. *(b)* Prior methods insert a fitted mechanism into the prediction pipeline before deployment. *(c)* GUARD leaves the deployed predictor unchanged and selectively corrects its output.

The first distinction is between *error* and *correctable error*. Missingness can expose systematic bias, but it can also create ambiguity: different outcomes may be compatible with the same observed modalities. Only the former is reachable by a correction computed from what remains. We show that this reachable component is exactly the *Bayes gap*, the discrepancy between the frozen prediction and the best prediction available from the observed modalities. Consequently, a badly degraded prediction need not offer more correction headroom than a mildly degraded one.

The second distinction concerns where a correction target can come from. For nested observations, we derive an exact relation between their Bayes gaps. When a predictor with more modalities is conditionally correct, its outputs can replace retrieval labels: averaging richer-view outputs conditional on the remaining modalities recovers the poorer-view Bayes target and, for the same fixed blend and neighbors, is no worse in expected cross-entropy or Brier risk than averaging hard labels. This assumption cannot be dropped: without some link between richer outputs and labels, the poorer Bayes target is not identifiable from unlabeled paired data alone.

The third distinction is between *correction* and *intervention*. A candidate correction that helps on average can still harm individual predictions. Our method, **GUARD**, therefore separates the two decisions: RECALIBRATE constructs a candidate output correction, while CERTIFY decides whether that candidate may replace the frozen prediction. GUARD learns an action score from deployment-available quantities and calibrates a threshold on a separate held-out split by conformal risk control (Angelopoulos et al., 2024). Under exchangeability, the marginal probability that GUARD both intervenes and increases loss beyond a prescribed tolerance is bounded by a preset level, for any score fixed before calibration.

Our contributions are as follows.

- **Correctability.** For common canonical losses, we characterize the error that remains correctable after a multimodal predictor is frozen: deployable correction can act only on the Bayes gap, not on ambiguity induced by the observed modalities.
- **Cross-mask transfer.** We derive an exact relation between the Bayes gaps of nested modality sets. It yields a retrieval target from richer-view frozen outputs, establishes when it exactly recovers the poorer-view Bayes predictor, proves finite-sample dominance for matched fixed blends under a conditionally correct richer predictor, and explains why retrieval-label-free recovery needs an assumption.
- **Risk-controlled intervention.** GUARD separates constructing a correction from deciding whether to use it. Its action score ranks utility, while an independent conformal calibration controls the marginal probability of harmful intervention for any fixed scorer and corrector.
- **Empirical evaluation.** Controlled experiments show that gain follows correctable headroom rather than raw error, and across nine frozen model–dataset pairs in six domains GUARD keeps most of blanket correction's utility while controlling harmful intervention.

2 Related Work

Missing-modality learning. Methods for handling absent modalities include aligning representations (Lin & Hu, 2023), distilling from a complete-modality teacher (Li et al., 2024b; Zhuang et al., 2025b), completing inputs by prompting (Lee et al., 2023; Guo et al., 2024), retrieving across samples (Zhuang et al., 2025a), and building reliability-aware architectures (He et al., 2026; Zou et al., 2026; Zhang et al., 2024; Zhuang et al., 2026). Ma et al. (2022) document how sharply multimodal transformers degrade when a modality is withheld. A recent line asks instead whether a missing modality should be repaired at all (Chen et al., 2026; Madaan et al., 2026; Gao et al., 2026). Closest in motivation, Gao et al. (2026) learn a gate jointly with paired direct and repaired branches but provide no finite-sample guarantee on harmful interventions. GUARD addresses a related question under a different deployment constraint: the deployed weights remain unchanged, the gate is fitted post hoc and calibrated on separate data, and no modality is reconstructed.

Correction after a frozen predictor. Roth et al. (2026) also freeze the deployed predictor, but fit residual completion operators from paired complete and incomplete views. GUARD instead edits outputs directly and couples correction with finite-sample harm control; its default hard-label retrieval target does not require paired complete and incomplete views. Other ways to adapt a deployed system change the predictor or its outputs rather than deciding when to act: test-time adaptation on unlabeled target data (Wang et al., 2021; Ramazanova et al., 2025), label-shift correction (Lipton et al., 2018), selective recalibration (Zollo et al., 2024), and nearest-neighbor interpolation such as kNN-LM (Khandelwal et al., 2020), which motivates our default retrieval implementation.

Selective intervention and risk control. Deciding whether to act connects to selective prediction, where rejection is learned jointly with prediction (Geifman & El-Yaniv, 2019) or samples are deferred to another decision maker (Madras et al., 2018), and to distribution-free risk control (Angelopoulos et al., 2025; 2024), with loss-controlling and action-conditional extensions (Wang et al., 2025; Laufer-Goldshtein et al., 2023; Zhu et al., 2026). Conformal decision theory is closest in form, calibrating decisions directly against a risk budget rather than through prediction sets (Lekeufack et al., 2024); Moayedikia (2026) instead certifies the prediction sets of a multimodal fusion architecture. GUARD scores each candidate intervention on a frozen predictor and calibrates one threshold for it, which yields the finite-sample harm guarantee stated in §4.2.

3 Correctability After the Model is Frozen

3.1 Frozen deployment setting

A sample carries M modalities and a label Y. At deployment a mask $S \subseteq \{1, \dots, M\}$ indicates the observed ones, and $X = (x_S, S)$ collects what is observed. Everything the corrector may read at deployment, including any representation the frozen model computes, is a function of X. A frozen model turns X into scores $z \in \mathbb{R}^K$, where K is the output dimension ($K = C$ for C-class classification). In this paper, we consider *canonical* losses $\ell(y, z)$ whose score-space gradient has the form

$$\nabla_z \ell(y, z) = m(z) - t(y), \tag{1}$$

where $m(\cdot)$ is the prediction and $t(\cdot)$ is the label statistic. For example, cross-entropy loss with softmax satisfies $m(z) = \mathrm{softmax}(z)$, $t(y) = e_y$; half-squared error $\ell(y, z) = \frac{1}{2}\|z - y\|^2$ satisfies $m(z) = z$, $t(y) = y$. We write $\bar{\ell}(y, p)$ for the loss evaluated on predictions, e.g. $-\log p_y$ for cross-entropy; loss conventions are detailed in App. F.1.

3.2 The Bayes gap: correctable error is not raw error

Condition equation 1 makes the deployment question tractable: averaging over labels replaces $t(Y)$ by its conditional mean. Define

$$\pi(X) := \mathbb{E}[t(Y) \mid X], \qquad d(X) := m(z(X)) - \pi(X). \tag{2}$$

Here $\pi(X)$ is the conditional class distribution for classification or conditional mean for squared error, and $d(X)$ is the model's *Bayes gap* from that target.

Theorem 1 (Bayes-gap reduction)**.** *Let $\ell(y, \cdot)$ be differentiable and convex on $\mathbb{R}^K$, satisfy equation 1, and have finite conditional risk and integrable $t(Y)$. For any finite deployable output $z'(X)$,*

$$\nabla_{z'} \mathbb{E}[\ell(Y, z') \mid X] = \mathbb{E}[m(z') - t(Y) \mid X] = m(z') - \pi(X). \tag{3}$$

Hence every finite conditional optimum satisfies $m(z^\star(X)) = \pi(X)$, and every such stationary point is optimal. An X-measurable post-hoc correction can therefore remove only the excess risk carried by the Bayes gap $d(X)$. In particular, if $m(z) = \pi(X)$ almost surely, no deployable output correction based on X can reduce expected risk. Boundary class distributions may require limiting logits.

The proof is in App. F.1. The underlying geometry is classical (Banerjee et al., 2005; Gneiting & Raftery, 2007); the deployment consequence is not. The result is exact in output space: uncertainty remaining at $\pi(X)$ cannot be eliminated by a correction based on X. Representation-level corrections face the same information constraint and induce measurable output edits (Props. 3, 5). Thus, *degradation severity and correction headroom are different quantities*: a severely degraded predictor may have little systematic error left to remove, while a less degraded one may retain substantial bias. §5.2 examines this distinction.

3.3 Nested observations: when richer views can replace labels

Theorem 1 identifies $\pi(X)$, written π_S under an observed modality set S, as the target of any correction, but estimating π_S ordinarily requires labels. Multimodal data provide an additional source of information: the same sample may also be observed under a richer modality set $T \supseteq S$. This raises a natural question: *can predictions made from the richer view serve as targets for correcting the poorer one?* Write X_S, m_S, π_S, and $d_S = m_S - \pi_S$ for the quantities associated with the poorer mask S, and similarly for T. Assume $\sigma(X_S) \subseteq \sigma(X_T)$ under the same joint law, including any mask context. The tower property gives $\pi_S = \mathbb{E}[\pi_T \mid X_S]$ and hence

$$d_S = \underbrace{\big(m_S - \mathbb{E}[m_T \mid X_S]\big)}_{\text{estimable from paired views and outputs}} + \underbrace{\mathbb{E}[d_T \mid X_S]}_{\text{inherited richer-view error}}. \tag{4}$$

The first term can be estimated from paired views and frozen predictions, whereas the second measures the systematic error inherited from the richer-view predictor. If the richer-view error vanishes after conditioning on X_S, that is, $\mathbb{E}[d_T \mid X_S] = 0$, then $\pi_S = \mathbb{E}[m_T \mid X_S]$. Thus, the population conditional average recovers the Bayes target for S without retrieval labels; finite-neighbor estimates additionally incur estimation error. A sufficient special case is a richer-mask predictor that is itself Bayes-optimal, $m_T = \pi_T$; that stronger condition is also the one the dominance result of §4.1 uses. Some assumption of this kind is unavoidable: without one, π_S is not identifiable from the unlabeled paired law (App. F.4). Neither identity says whether a given correction should be applied.

4 GUARD: From Correction to Intervention

The theory answers what can in principle be corrected, but deployment still requires two separate decisions: how to construct a candidate correction, and whether that candidate should replace the frozen prediction. GUARD separates them: without touching model weights, RECALIBRATE turns the Bayes target into an estimable candidate prediction, while CERTIFY treats that candidate as an optional action and decides when it may be applied under a prescribed harm budget. Three disjoint splits carry the offline steps: D_{source} builds the target estimator, D_{fit} fits the blend weight $\widehat{\beta}$ and action score r_θ, and D_{conf} calibrates the threshold $\widehat{\lambda}$. Figure 2 illustrates the implementation of GUARD.

4.1 RECALIBRATE: construct a candidate correction

Given the Bayes target $\pi(X)$ identified in §3, GUARD treats its estimator as a frozen plug-in $\widehat{\pi} = \mathcal{A}(D_{\text{source}}, X)$ built from a source split. We instantiate $\mathcal{A}$ by averaging neighbors in the available feature space; its information limit is discussed in App. C. The default target averages their labels, while the inheritance identity equation 4 yields a second target that replaces labels with predictions from a richer modality set:

$$\text{(hard-label) } \widehat{\pi}^{(Y)}(X) = \sum_{j \in \mathcal{N}_k(X)} w_j \, t(Y_j), \quad \text{(cross-mask transfer) } \widehat{\pi}^{(T)}(X) = \sum_{j \in \mathcal{N}_k(X)} w_j \, m_T\big(X_T^{(j)}\big), \tag{5}$$

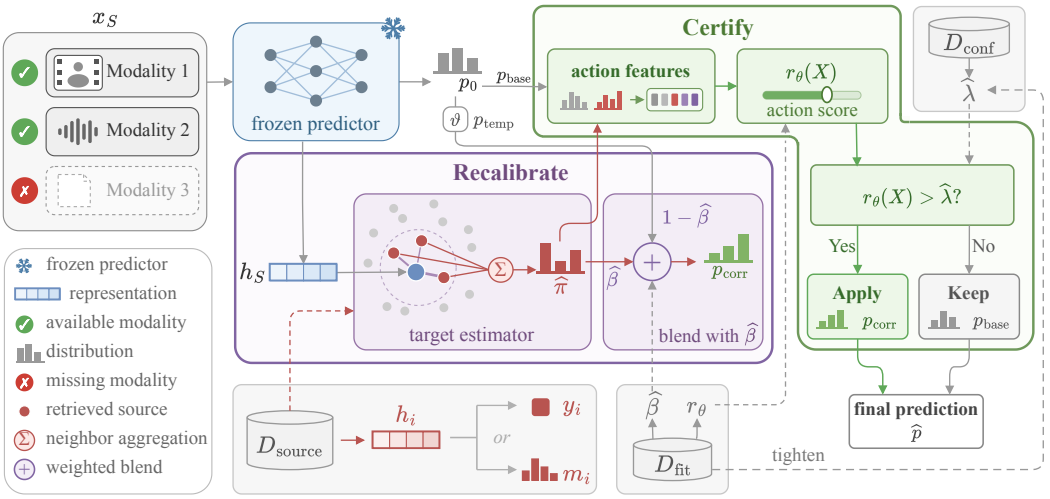

Figure 2: **GUARD at deployment.** The frozen output p_0 is temperature-scaled to p_{temp} before blending with $\widehat{\pi}$. The reference $p_{\text{base}} = p_0$ feeds the action features, defines harm on D_{conf}, and is kept when the gate declines (App. D). Offline, D_{source} builds the target estimator; D_{fit} selects ϑ and $\widehat{\beta}$, fits r_θ, and tightens the threshold that D_{conf} calibrates by conformal risk control.

where $w_j \geq 0$ and $\sum_j w_j = 1$. Let $p_{\text{temp}} = \text{Temp}_\vartheta(m(z))$ denote the frozen prediction after temperature scaling. The weights, temperature ϑ, k, target and retrieval space are selected on D_{fit} (App. D). Either target produces a candidate

$$p_{\text{corr}} = (1 - \widehat{\beta})\, p_{\text{temp}} + \widehat{\beta}\, \widehat{\pi}, \tag{6}$$

where $\widehat{\beta} \in [0, 1]$ minimizes mean prediction loss $\bar{\ell}$ on D_{fit} (App. F.6). The reference p_{base} used to measure harm and retained on rejected queries is the frozen output p_0 throughout (App. D). Selecting the candidate affects utility but not the validity of CERTIFY under its exchangeability assumption. The hard-label mode is nearest-neighbor interpolation in the spirit of kNN-LM (Khandelwal et al., 2020); App. B also evaluates a trained probe.

Why richer-mask outputs can replace labels. The two targets exchange different sources of error. Hard-label retrieval contains per-neighbor label noise; cross-mask transfer removes that noise but inherits any systematic error of the richer-view predictor. If that predictor is conditionally correct, $m_T = \pi_T$, its output is the conditional mean of the corresponding hard label. Prop. 9 then shows that, under cross-entropy or Brier score, the cross-mask blend has no larger expected risk for the same fixed $\beta \in [0, 1]$, admissible k, neighbors and feature-based weights. This compares fixed blends before gating; it does not order separately tuned or tempered targets. App. F.4 gives the proof, optimal blend, and label-efficiency trade-off.

4.2 CERTIFY: DECIDING WHEN TO APPLY IT

From correction to intervention. A correction that improves average risk may still harm individual predictions, so GUARD treats p_{corr} as a candidate action rather than an automatic replacement. For a labeled pair (X, Y), define the change in loss induced by correction as

$$\Delta_Y(X) = \bar{\ell}(Y, p_{\text{corr}}(X)) - \bar{\ell}(Y, p_{\text{base}}(X)). \tag{7}$$

Thus $\Delta_Y(X) < 0$ means that correction helps, while we call an intervention *harmful* when $\Delta_Y(X) > \delta$, for a tolerated increase $\delta \geq 0$. An ideal policy would intervene only where correction helps, but Δ_Y depends on Y, which is unavailable at deployment. We therefore use D_{fit} to learn a score $r_\theta(X)$ that ranks candidate interventions by their predicted utility. The score is trained with the binary target

$$Z = \mathbf{1}\{\bar{\ell}(Y, p_{\text{corr}}(X)) < \bar{\ell}(Y, p_{\text{base}}(X))\}, \tag{8}$$

but uses only features available at deployment. In our implementation, r_θ is a logistic model over five features derived from $(p_{\text{base}}, \widehat{\pi})$, listed in App. F.6. The score r_θ only ranks candidates: it is trained to predict whether correction improves loss ($\Delta_Y < 0$), whereas the guarantee below controls the distinct event of excessive harm ($\Delta_Y > \delta$).

Finite-sample harm control. By itself, this ranking provides no control over how often harmful corrections are applied. We therefore threshold it to obtain a nested family of intervention policies,

$$A_\lambda(X) = \mathbf{1}\{r_\theta(X) > \lambda\}, \qquad L_\lambda(X, Y) = A_\lambda(X)\,\mathbf{1}\{\Delta_Y(X) > \delta\}, \tag{9}$$

where A_λ determines whether to intervene and L_λ indicates a harmful intervention. As λ increases, the intervention set can only shrink, so $L_\lambda \in [0, 1]$ is non-increasing in λ. We calibrate λ on a held-out split $D_{\text{conf}} = \{(X_i, Y_i)\}_{i=1}^n$, which is not used to fit the corrector, $\widehat{\beta}$, or r_θ. Let $\widehat{R}_n(\lambda) = n^{-1}\sum_i L_\lambda(X_i, Y_i)$ be its empirical harmful-intervention rate, and let $\alpha \in (0, 1)$ be the prescribed harm budget. Conformal risk control (Angelopoulos et al., 2024) selects

$$\widehat{\lambda} = \inf\Big\{\lambda : \frac{n}{n+1}\widehat{R}_n(\lambda) + \frac{1}{n+1} \le \alpha\Big\}, \tag{10}$$

with $\widehat{\lambda} = +\infty$ (no intervention) if the set is empty or $n = 0$. The selected threshold has the following guarantee:

Proposition 2 (Finite-sample harm control, after Angelopoulos et al., 2024)**.** *Condition on all pre-calibration fitting, fixing p_{base}, p_{corr}, r_θ, the loss and δ. If calibration and deployment samples are exchangeable under this conditional law, then $\widehat{\lambda}$ from equation 10 satisfies*

$$\mathsf{Pr}(\Delta_Y(X) > \delta \ \wedge\ \text{apply}) \le \alpha. \tag{11}$$

App. F.5 verifies the conditions required by conformal risk control. Proposition 2 has two consequences. First, the guarantee holds for *any* score fixed before D_{conf} is read: a poor ranking may reduce utility or intervention rate, but it does not invalidate the harm bound. Second, the guarantee is marginal over the calibration draw and deployment sample. It therefore does not certify any individual correction and controls joint harm, $\mathsf{Pr}(\Delta_Y > \delta \wedge \text{apply})$, rather than the conditional rate $\mathsf{Pr}(\Delta_Y > \delta \mid \text{apply})$. Conditional harm can be high when intervention is rare; App. B reports all three rates.

GUARD may further increase $\widehat{\lambda}$ using D_{fit} to remove low-ranked interventions that add little estimated utility (App. F.6). This cannot weaken the guarantee: for any $\lambda^\star \ge \widehat{\lambda}$, $L_{\lambda^\star} \le L_{\widehat{\lambda}}$ pointwise, and hence $\mathbb{E}[L_{\lambda^\star}] \le \alpha$. The resulting certified policy applies p_{corr} only when admitted by the gate and otherwise retains p_{base}; its *gated gain* is the corresponding change in the benchmark metric.

5 Experimental Results

5.1 Experimental setup

Our evaluation spans seven benchmark families across six application domains, covering nine frozen model–dataset pairs and forty-four degraded conditions. Each frozen model is anchored to its published implementation or protocol before GUARD is attached, and its parameters remain fixed throughout evaluation. The degradation axes reflect different forms of information loss covered by Theorem 1: entire modalities are withheld on multimodal benchmarks, sensor families, electrodes, or ECG leads are removed, and protein sequences are progressively truncated.

For every condition, D_{source}, D_{fit}, D_{conf}, and D_{test} have disjoint roles: IEMOCAP is split by video, OPPORTUNITY by time block and the others by row within the evaluated population, so exchangeability is claimed for held-out rows or blocks rather than an unseen subject (App. D). All retrieval choices and $\widehat{\beta}$ are selected before test evaluation, and CERTIFY is calibrated only on D_{conf}; the test set is read once after all decisions are fixed. Unless stated otherwise, we use harm budget $\alpha = 0.2$ and tolerance $\delta = 0.05$ and report averages over the prescribed seeds or deployment splits. The four subsections below test one claim each. App. D gives model provenance, degradation levels, data splits, retrieval grids, and baselines; App. B reports all per-condition results.

5.2 Does gain follow correctable headroom rather than raw error?

Theorem 1 identifies the target of correction, while realized gain also depends on its estimation. To isolate headroom from raw error, we need a known Bayes target and separate control over ambiguity and bias. We therefore construct a controlled synthetic classification with $\pi(X) = \text{softmax}(WX/\tau)$

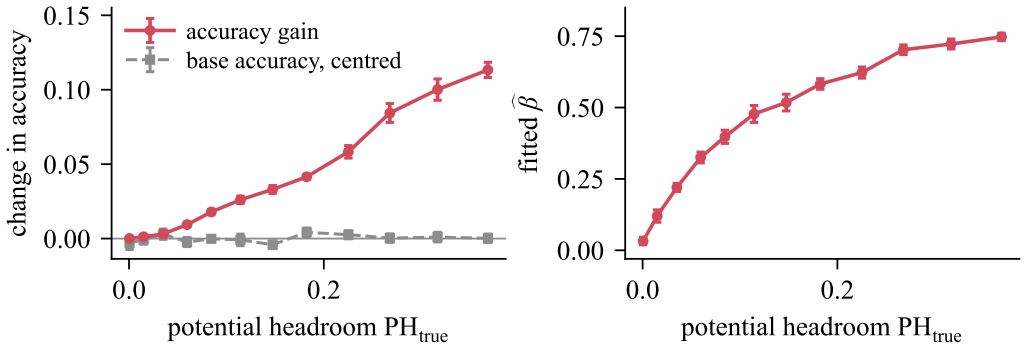

Figure 3: **Gain follows the gap, not the error** on 12 synthetic configurations and 10 seeds. For each ambiguity level τ, the bias parameter γ is chosen to keep base accuracy near 0.65, so the amount of error stays fixed. *Left:* accuracy gain against true potential headroom $\mathrm{PH}_{\mathrm{true}} = \mathbb{E}\|d(X)\|_2^2$, with base accuracy centered on its mean. *Right:* the blend weight selected on D_{fit}.

and frozen-model prediction $m(X) = \mathrm{softmax}(WX/\tau + \gamma VX)$. Here τ controls conditional ambiguity, while γ controls systematic model bias. For each τ, we choose γ so that base accuracy remains near 0.65, producing 12 configurations with similar raw error but different correctable headroom. Because $\pi(X)$ is known in this setting, we can measure the true potential headroom $\mathrm{PH}_{\mathrm{true}} := \mathbb{E}\|d(X)\|_2^2$, the reducible component of the unhalved Brier risk. Accuracy gain need not be monotone in this quantity; the relation below is empirical. Across 10 seeds, correction gain tracks $\mathrm{PH}_{\mathrm{true}}$ with correlation 0.99 and vanishes as the gap disappears; the fitted blend weight decreases with it. At matched base accuracy the most bias-dominated configuration gains $+0.11$ in accuracy and $+0.86$ in loss, while the ambiguity-dominated one gains nothing on either scale.

5.3 Does post-hoc correction generalize across frozen models?

We next ask whether the same frozen-deployment policy survives changes in architecture, modality type and application domain. On CMU-MOSEI, attaching GUARD to three published frozen models improves all three, by up to 2.1 accuracy points on average; the quoted rows are context, not a ranking: they use their own protocols and no deployment-matched labels. The gain concentrates on the subsets that lose language, averaging 3.5 points there and reaching 5.8, and stays within a point when language is retained. Table 1 gives the full published comparison; the main question here is whether the same behavior extends beyond sentiment.

We then apply the same policy across the whole suite. Every benchmark mean improves, with joint harm inside the budget. Gain ranges from 19.0 points where sensor loss exposes large correctable headroom to 0.4 on a model that largely absorbs the degradation, a gain within two standard errors of zero (App. B). Low joint harm need not mean reliable interventions: 83.15% of applied occurrences on IEMOCAP audio plus text are harmful, with a slight accuracy loss (App. B).

5.4 Why not correct every prediction?

A correction can be useful on average and still be a poor blanket deployment policy. Figure 4 isolates this distinction by applying the same corrector either everywhere or only where CERTIFY admits it. The utility curves nearly coincide, but the harm curves separate sharply: the ungated corrector crosses the budget on all six benchmarks, whereas GUARD stays under it on all six.

Table 3 compares scoring, calibration and tightening with the same frozen model and corrector. Among the calibrated rules, realized harm stays within budget except two without tightening: *GUARD w/o tighten* on CMU-MOSEI (0.202) and the confidence score on OPPORTUNITY (0.205), both allowed by an expectation bound. First, against *GUARD w/o tighten*, the learned score improves or matches gain on six of the seven benchmarks, IEMOCAP excepted, but does not uniformly lower harm: model confidence is lower or equal on four of the seven. Second, under the same comparison Learn-then-Test records lower realized harm on all seven. The two control different probabilities, so neither reading dominates. Third, the fit-split tightening lowers joint harm on all seven benchmarks;

Table 1: **CMU-MOSEI under different modality combinations.** Binary accuracy and weighted F1 (ACC/F1, %) on non-neutral samples. L, A and V denote observed language, acoustic and visual modalities. Unmarked rows are quoted from Zhuang et al. (2025b) and Zhuang et al. (2026); † marks our frozen reproductions. Reproduced rows average five deployment splits, except the unchanged text-only CMAD rows, which use the full test set. Averages are taken before rounding; bold marks the best entry in each column. Harm and apply rates are in App. B.

Method	L	A	V	L, A	L, V	A, V	L, A, V	**Avg.**
Published under their own training protocols; context, not a controlled comparison								
MCTN (Pham et al., 2019)	82.6/82.8	62.7/54.5	62.6/57.1	83.5/83.3	83.2/83.2	63.7/62.7	84.2/84.2	74.6/72.5
MMIN (Zhao et al., 2021)	82.3/82.4	58.9/59.5	59.3/60.0	83.7/83.3	83.8/83.4	63.5/61.9	84.3/84.2	73.7/73.5
GCNet (Lian et al., 2023)	83.0/83.2	60.2/60.3	61.9/61.6	84.3/84.4	84.3/84.4	64.1/57.2	85.2/85.1	74.7/73.7
IMDer (Wang et al., 2023b)	84.5/84.5	63.8/60.6	63.9/63.6	85.1/85.1	85.0/85.0	64.9/63.5	85.1/85.1	76.0/75.3
DiCMoR (Wang et al., 2023a)	84.2/84.3	62.9/60.4	63.6/63.6	85.0/84.9	84.9/84.9	65.2/64.4	85.1/85.1	75.8/75.4
MPLMM (Guo et al., 2024)	84.6/84.6	61.8/49.4	62.2/53.2	85.1/85.1	85.2/85.1	63.2/51.0	85.3/85.3	75.3/70.5
UMDF (Li et al., 2024a)	85.3/85.2	62.9/59.4	62.1/62.4	85.3/85.3	85.3/85.2	65.4/62.8	85.7/85.6	76.0/75.1
CorrKD (Li et al., 2024b)	85.3/85.1	61.9/56.8	63.0/62.5	85.3/85.2	85.6/85.5	64.5/62.3	85.7/85.6	75.9/74.7
MMANet (Wei et al., 2023)	84.6/84.6	62.9/48.5	63.7/63.1	84.8/84.8	85.3/85.2	64.5/60.9	85.3/85.2	75.9/73.2
CMAD (Zhuang et al., 2025b)†	85.7/85.7	62.9/58.9	63.0/61.6	86.7/86.7	86.8/86.8	63.9/63.0	86.8/86.8	76.6/75.6
TMDC (Zhuang et al., 2026)†	85.6/85.6	73.5/72.9	67.1/67.1	85.1/85.2	85.5/85.6	67.6/67.8	84.6/84.7	78.4/78.4
MoMKE (Xu et al., 2024)†	86.5/86.5	71.1/71.1	70.2/69.4	87.0/87.0	87.1/87.0	72.7/72.4	**87.5/87.4**	80.3/80.1
Same frozen models as the three rows above, corrected post hoc under our protocol								
CMAD + GUARD	85.7/85.7	67.6/65.1	67.6/66.2	86.8/86.8	86.6/86.6	69.7/68.8	86.8/86.8	78.7/78.0
TMDC + GUARD	85.6/85.6	**74.0/73.4**	72.1/72.0	85.5/85.7	85.5/85.6	72.4/72.4	85.5/85.7	80.1/80.0
MoMKE + GUARD	**86.7/86.7**	72.2/72.0	**73.0/72.2**	**87.1/87.1**	**87.2/87.2**	**74.9/74.5**	87.2/87.2	**81.2/81.0**

Table 2: **Frozen model and GUARD at $\alpha = 0.2$ across six domains.** Metrics follow each benchmark. Harm and apply ranges cover incomplete conditions; CMU-MOSEI uses only language-absent masks (A, V, AV). *Cond.* is the number of degraded conditions. Harm is measured against the frozen output p_0 (App. D).

Benchmark (model)	**Metric**	**Cond.**	**frozen**	**GUARD**	**Joint harm**	**Apply (%)**
DrugBAN (3 datasets)	AUROC	9	0.718	**0.848**	0.049–0.165	43–74
OPPORTUNITY (DeepConvLSTM)	wtd. F1	4	0.563	**0.753**	0.148–0.186	56–71
CMU-MOSEI (CMAD)	binary acc.	3	0.637	**0.688**	0.094–0.116	30–37
IEMOCAP (MoMKE)	4-class WA	6	0.680	**0.684**	0.031–0.121	4–42
AVE (AV-att)	29-way acc.	2	0.464	**0.548**	0.177–0.178	57–65
NinaPro DB5 (sEMG CNN)	accuracy	10	0.663	**0.742**	0.118–0.147	75–98
PTB-XL (resnet1d_wang)	macro AUC	10	0.761	**0.852**	0.078–0.157	47–87

its cost is clearest on CMU-MOSEI, where gain falls from $+0.064$ to $+0.051$ as harm falls from 0.202 to 0.106, and is at most 0.006 elsewhere.

Table 3: **Scoring, calibration and tightening.** Every rule sees the same frozen model, corrector, splits and budget $\alpha = 0.2$. *Blanket* corrects everything: an ungated reference, not an upper bound. Against *GUARD w/o tighten*, *Conf. score* changes only the score (Hendrycks & Gimpel, 2017; Geifman & El-Yaniv, 2017) and *LTT calib.* uses Learn-then-Test (Angelopoulos et al., 2025), with a different entropy feature on PTB-XL; the last two columns differ only in the fit-split tightening. Cells give accuracy gain / joint harm rate. A star marks harm above α, and boldface marks the lowest joint harm in each row. Details are in App. D.

Benchmark	**Blanket**	**Conf. score**	**LTT calib.**	**GUARD w/o tighten**	**GUARD**
CMU-MOSEI	$+0.067/0.331^*$	$+0.055/0.198$	$+0.062/0.177$	$+0.064/0.202^*$	$+0.051/\mathbf{0.106}$
IEMOCAP	$+0.004/0.734^*$	$+0.008/0.180$	$+0.007/0.145$	$+0.004/0.178$	$+0.003/\mathbf{0.077}$
OPPORTUNITY	$+0.183/0.279^*$	$+0.171/0.205^*$	$+0.176/0.171$	$+0.178/0.187$	$+0.178/\mathbf{0.169}$
AVE	$+0.084/0.300^*$	$+0.073/0.191$	$+0.084/\mathbf{0.169}$	$+0.084/0.196$	$+0.084/0.178$
NinaPro DB5	$+0.092/0.125$	$+0.092/0.125$	$+0.092/0.124$	$+0.092/0.125$	$+0.092/\mathbf{0.072}$
PTB-XL	$+0.170/0.239^*$	$+0.164/0.193$	$+0.168/0.169$	$+0.171/0.194$	$+0.165/\mathbf{0.150}$
DrugBAN	$+0.085/0.275^*$	$+0.074/0.191$	$+0.084/0.171$	$+0.085/0.190$	$+0.084/\mathbf{0.150}$

5.5 When can richer-view outputs replace retrieval labels?

The cross-mask result identifies a trade-off: replacing hard labels removes retrieval-label noise but inherits error from the richer-view predictor. Table 4 studies its practical effect by holding the neighbors fixed and changing the target before fitting the blend and gate. When the richer predictor is

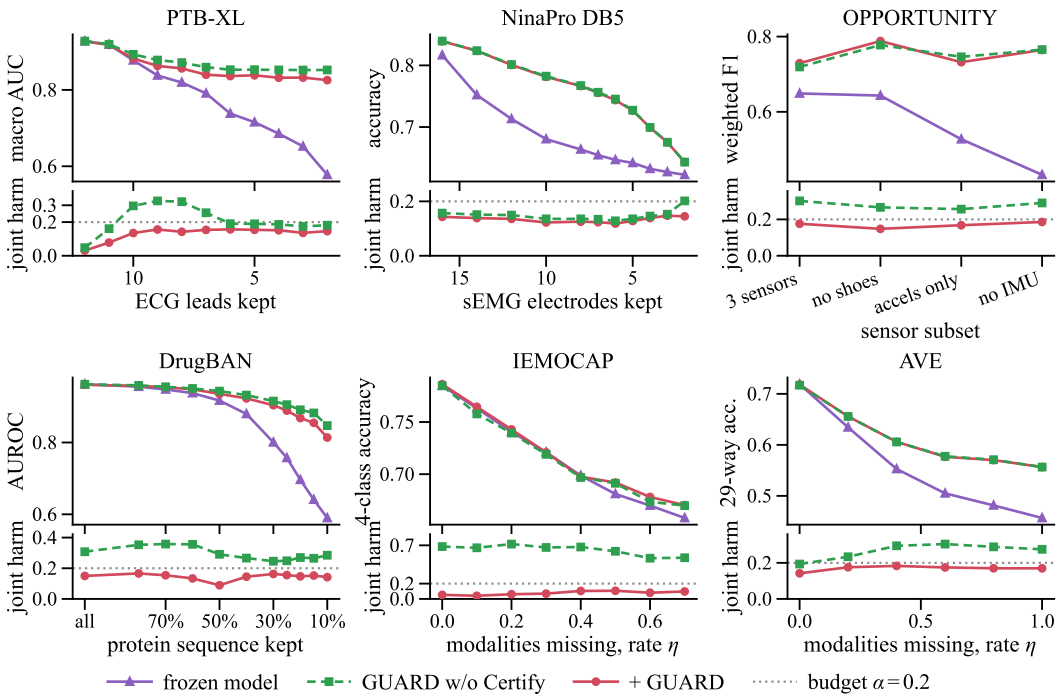

Figure 4: **Utility and harm as the degradation deepens.** Upper axis: the published metric. Lower strip: joint harm against the budget. Dashed *GUARD w/o* CERTIFY applies the corrector everywhere and overlaps GUARD where the gate admits everything. Blanket joint harm runs from 0.202 on NinaPro to 0.72 on IEMOCAP, against 0.200.

Table 4: **Certified gain under the two retrieval targets.** Targets from equation 5 share neighbors, with the blend and gate refitted. Accuracy gain is averaged over conditions and seeds; rows are ordered by richer-mask accuracy. Bold marks a gain larger by over two paired standard errors.

	richer-mask	**gated gain**	
Dataset, split	**accuracy**	**hard-label**	**cross-mask**
Human, random	0.931	+0.033	+0.035
BindingDB, random	0.906	+0.146	+0.143
BioSNAP, random	0.829	+0.138	+0.138
OPPORTUNITY	0.611	**+0.178**	+0.056
BioSNAP, cluster	0.556	**+0.112**	+0.028
BindingDB, cluster	0.551	**+0.234**	+0.028

strong, cross-mask transfer matches hard-label retrieval to within 0.003 without retrieval labels, also from a pool the model never trained on (App. B). When it is weak (0.551–0.611 accuracy), most of the gain is lost. Proposition 9 is conditional, so this is not a universal accuracy threshold; conditional correctness is not identifiable from unlabeled data (App. F.7).

6 Discussion and Conclusion

Missing modalities leave a frozen model with a Bayes gap that the observed inputs can reduce and ambiguity that they cannot; richer-view outputs can replace retrieval labels when the richer model is conditionally correct. Because a helpful correction can still hurt individual predictions, GUARD separates building a candidate from deciding when to apply it, keeping mean joint harm within budget on every benchmark. Its guarantee is marginal, assumes exchangeability and needs labeled deployment data. Except for the cross-mask target, the analysis covers any frozen model whose inputs change. Conditional risk control, shift-aware calibration and fewer labels could make certified correction a general maintenance layer for deployed models.

AI USE STATEMENT

This paper was prepared with the assistance of generative AI tools, and we disclose that use here.

The authors formulated the research questions and the theoretical framework. Generative AI tools assisted in checking mathematical statements, identifying missing assumptions and counterexamples, and revising derivations and proofs. The synthetic family behind Figure 3 (App. E) is specified analytically and was designed by the authors, and no data set used in this paper was generated by a generative AI tool.

Generative AI tools assisted the authors in implementing and debugging experiment code, including the sampling and evaluation code for that synthetic family and the scripts that render the figures. They assisted in drafting and editing the manuscript, and in searching for related work. The authors also used a large language model to obtain critical feedback on drafts of this manuscript; the authors decided which of that feedback to act on.

The authors retained responsibility for formulating the mathematical claims, choosing the experimental methodology, and interpreting the reported results. Generative AI tools were not used to propose research hypotheses, independently design experiments, clean or reformat the benchmark data, or translate research material. Qualitative or thematic data analysis was not applicable to this work.

The authors are responsible for verifying AI-assisted contributions before submission and for the final text, claims and artifacts. Mathematical arguments and proofs were reviewed by the authors, and experimental results were checked against the saved experiment outputs. The verification script described below checks reported numbers against saved experiment outputs; agreement with those outputs does not itself validate the experimental protocol or theoretical assumptions.

ETHICS STATEMENT

This work develops post-hoc correction for already-deployed multimodal models and evaluates it on existing research datasets (BioSNAP, BindingDB, Human, OPPORTUNITY, CMU-MOSEI, IEMOCAP, AVE, NinaPro DB5 and PTB-XL), accessed under their respective terms; no new data were collected from human subjects. Several of these datasets record people and PTB-XL contains clinical ECGs, so we note that Proposition 2 bounds a marginal rate over a deployment distribution: it certifies no individual prediction and does not by itself bound harm within a subgroup (§4.2, App. C). GUARD changes the end-to-end output even when the backbone is frozen. Clinical use would therefore require validation of the corrected system and assessment under applicable regulatory change control; unchanged weights alone imply neither regulatory exemption nor clinical safety (FDA, 2025).

REPRODUCIBILITY STATEMENT

An anonymized implementation accompanies this submission at `https://github.com/anonymous221206/GUARD`. Assumptions and complete proofs for the theoretical claims are given in App. F.1 onwards, and the conditions under which each result applies are collected in App. C. The experimental protocol, the provenance of every frozen model and the definition of each degradation are in App. D; the per-condition numbers behind every summary table are in App. B. In the repository, `docs/PAPER_MAP.md` maps each table and figure to the driver and the saved outputs that produce it. Those outputs ship with the code, so every reported cell can be recomputed and every figure redrawn from a clean clone without a GPU; the verification scripts in `scripts/tables/` check them against the manuscript source.

REFERENCES

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations*, 2024.

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2):1641–1662, 2025. doi: 10.1214/24-AOAS1998.

Peizhen Bai, Filip Miljković, Bino John, and Haiping Lu. Interpretable bilinear attention network with domain adaptation improves drug–target prediction. *Nature Machine Intelligence*, 5(2):126–136, 2023. doi: 10.1038/s42256-022-00605-1.

Arindam Banerjee, Xin Guo, and Hui Wang. On the optimality of conditional expectation as a Bregman predictor. *IEEE Transactions on Information Theory*, 51(7):2664–2669, 2005. doi: 10.1109/TIT.2005.850145.

Rongfei Chen, Tingting Zhang, Xiaoyu Shen, and Wei Zhang. Evaluation before generation: A paradigm for robust multimodal sentiment analysis with missing modalities. *arXiv preprint arXiv:2604.05558*, 2026.

FDA. Marketing submission recommendations for a predetermined change control plan for artificial intelligence-enabled device software functions. U.S. Food and Drug Administration, final guidance for industry and FDA staff, 2025. URL `https://www.fda.gov/media/166704/download`. Issued August 18, 2025.

Yubo Gao, Haotian Wu, Xiaoyu Xu, Yibo Yan, Hong Chen, Ruoshui Peng, Fei Pan, Puay Siew Tan, Zhuoran Gao, Yonghua Hei, Jie Zhang, and Xuming Hu. Should missing modalities always be necessary to repair for multi-modal sentiment analysis? *arXiv preprint arXiv:2607.17262*, 2026.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 4878–4887, 2017.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 2151–2159. PMLR, 2019.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437.

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

Kang He, Yuzhe Ding, Rao Fu, Yukang Feng, Kaipeng Zhang, Yiming Liu, Fei Li, Chong Teng, and Donghong Ji. Uncertainty-calibrated elastic alignment for multimodal sentiment analysis with missing modalities. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 5268–5288, 2026. doi: 10.18653/v1/2026.findings-acl.260.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations (ICLR)*, 2017.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Generalization through memorization: Nearest neighbor language models. In *International Conference on Learning Representations*, 2020.

Bracha Laufer-Goldshtein, Adam Fisch, Regina Barzilay, and Tommi S. Jaakkola. Efficiently controlling multiple risks with pareto testing. In *International Conference on Learning Representations*, 2023.

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

Jordan Lekeufack, Anastasios N. Angelopoulos, Andrea Bajcsy, Michael I. Jordan, and Jitendra Malik. Conformal decision theory: Safe autonomous decisions from imperfect predictions. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 11668–11675. IEEE, 2024. doi: 10.1109/ICRA57147.2024.10610041.

Mingcheng Li, Dingkang Yang, Yuxuan Lei, Shunli Wang, Shuaibing Wang, Liuzhen Su, Kun Yang, Yuzheng Wang, Mingyang Sun, and Lihua Zhang. A unified self-distillation framework for multimodal sentiment analysis with uncertain missing modalities. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(9):10074–10082, 2024a. doi: 10.1609/aaai.v38i9.28871.

Mingcheng Li, Dingkang Yang, Xiao Zhao, Shuaibing Wang, Yan Wang, Kun Yang, Mingyang Sun, Dongliang Kou, Ziyun Qian, and Lihua Zhang. Correlation-decoupled knowledge distillation for multimodal sentiment analysis with incomplete modalities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12458–12468, 2024b.

Zheng Lian, Lan Chen, Licai Sun, Bin Liu, and Jianhua Tao. GCNet: Graph completion network for incomplete multimodal learning in conversation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):8419–8432, 2023. doi: 10.1109/TPAMI.2023.3234553.

Ronghao Lin and Haifeng Hu. MissModal: Increasing robustness to missing modality in multimodal sentiment analysis. *Transactions of the Association for Computational Linguistics*, 11:1686–1702, 2023. doi: 10.1162/tacl_a_00628.

Zachary C. Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and correcting for label shift with black box predictors. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3122–3130. PMLR, 2018.

Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. Are multimodal transformers robust to missing modality? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18177–18186, 2022.

Divyam Madaan, Sumit Chopra, and Kyunghyun Cho. Characterizing the predictive impact of modalities with supervised latent-variable modeling. In *International Conference on Machine Learning (ICML)*, 2026.

David Madras, Toni Pitassi, and Richard Zemel. Predict responsibly: Improving fairness and accuracy by learning to defer. In *Advances in Neural Information Processing Systems*, volume 31, pp. 6150–6160, 2018.

Alireza Moayedikia. Conformal fusion under missing modalities. *arXiv preprint arXiv:2608.07183*, 2026.

Ganesh R. Naik, Ali H. Al-Timemy, and Hung T. Nguyen. Transradial amputee gesture classification using an optimal number of sEMG sensors: An approach using ICA clustering. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 24(8):837–846, 2016. doi: 10.1109/TNSRE.2015.2478138.

Francisco Javier Ordóñez and Daniel Roggen. Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition. *Sensors*, 16(1):115, 2016. doi: 10.3390/s16010115.

Srinivas Parthasarathy and Shiva Sundaram. Training strategies to handle missing modalities for audio-visual expression recognition. In *Companion Publication of the 2020 International Conference on Multimodal Interaction (ICMI '20 Companion)*, pp. 400–404, 2020. doi: 10.1145/3395035.3425202.

Luis Pelaez Murciego, Mauricio C. Henrich, Erika G. Spaich, and Strahinja Dosen. Reducing the number of EMG electrodes during online hand gesture classification with changing wrist positions. *Journal of NeuroEngineering and Rehabilitation*, 19:78, 2022. doi: 10.1186/s12984-022-01056-w.

Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabás Póczos. Found in translation: Learning robust joint representations by cyclic translations between modalities. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6892–6899, 2019. doi: 10.1609/aaai.v33i01.33016892.

Stefano Pizzolato, Luca Tagliapietra, Matteo Cognolato, Monica Reggiani, Henning Müller, and Manfredo Atzori. Comparison of six electromyography acquisition setups on hand movement classification tasks. *PLoS ONE*, 12(10):e0186132, 2017. doi: 10.1371/journal.pone.0186132.

Merey Ramazanova, Alejandro Pardo, Bernard Ghanem, and Motasem Alfarra. Test-time adaptation for combating missing modalities in egocentric videos. In *International Conference on Learning Representations*, 2025.

Matthew A. Reyna, Nadi Sadr, Erick A. Perez Alday, Annie Gu, Amit J. Shah, Chad Robichaux, Ali Bahrami Rad, Andoni Elola, Salman Seyedi, Sardar Ansari, Hamid Ghanbari, Qiao Li, Ashish Sharma, and Gari D. Clifford. Will two do? varying dimensions in electrocardiography: The PhysioNet/Computing in cardiology challenge 2021. In *Computing in Cardiology*, volume 48, pp. 1–4, 2021. doi: 10.23919/CinC53138.2021.9662687.

Holger R. Roth, Ziyue Xu, and Peter Cnudde. FedCoRe: Target-adaptive completion for missing modalities in healthcare federated learning. *arXiv preprint arXiv:2608.18311*, 2026.

Ann M. Simon, Keira Newkirk, Laura A. Miller, Kristi L. Turner, Kevin Brenner, Michael Stephens, and Levi J. Hargrove. Implications of EMG channel count: enhancing pattern recognition online prosthetic testing. *Frontiers in Rehabilitation Sciences*, 5:1345364, 2024. doi: 10.3389/fresc.2024.1345364.

Nils Strodthoff, Patrick Wagner, Tobias Schaeffter, and Wojciech Samek. Deep learning for ECG analysis: Benchmarks and insights from PTB-XL. *IEEE Journal of Biomedical and Health Informatics*, 25(5):1519–1528, 2021. doi: 10.1109/JBHI.2020.3022989.

Tianxiang Sun, Yunfan Shao, Hong Qian, Xuanjing Huang, and Xipeng Qiu. Black-box tuning for language-model-as-a-service. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, pp. 20841–20855. PMLR, 2022.

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

Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Dieter Kreiseler, Fatima I. Lunze, Wojciech Samek, and Tobias Schaeffter. PTB-XL, a large publicly available electrocardiography dataset. *Scientific Data*, 7:154, 2020. doi: 10.1038/s41597-020-0495-6.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021.

Di Wang, Ping Wang, Zhong Ji, Xiaojun Yang, and Hongyue Li. Conformal loss-controlling prediction. *IEEE Transactions on Neural Networks and Learning Systems*, 36(2):2973–2983, 2025. doi: 10.1109/TNNLS.2024.3356512.

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

Yuanzhi Wang, Yong Li, and Zhen Cui. Incomplete multimodality-diffused emotion recognition. In *Advances in Neural Information Processing Systems*, volume 36, pp. 17117–17128, 2023b. doi: 10.52202/075280-0748.

Shicai Wei, Chunbo Luo, and Yang Luo. MMANet: Margin-aware distillation and modality-aware regularization for incomplete multimodal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20039–20049, 2023.

Wenxin Xu, Hexin Jiang, and Xuefeng Liang. Leveraging knowledge of modality experts for incomplete multimodal learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 438–446, 2024. doi: 10.1145/3664647.3681683.

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

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

Wugeng Zheng, Ziwen Kan, Tianlong Chen, Chen Chen, and Song Wang. MuteBench: Modality unavailability tolerance evaluation for incomplete multimodal fusion. *arXiv preprint arXiv:2605.15235*, 2026.

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

Zihan Zhu, Shayan Kiyani, George Pappas, and Hamed Hassani. Conformal risk-averse decision making with action conditional guarantee. In *International Conference on Machine Learning (ICML)*, 2026.

Yan Zhuang, Minhao Liu, Wei Bai, Yanru Zhang, Wei Li, Jiawen Deng, and Fuji Ren. Hypermodality enhancement for multimodal sentiment analysis with missing modalities. In *Advances in Neural Information Processing Systems*, volume 38, pp. 157679–157708, 2025a. doi: 10.52202/085713-4754.

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

Yan Zhuang, Minhao Liu, Yanru Zhang, Jiawen Deng, and Fuji Ren. TMDC: A two-stage modality denoising and complementation framework for multimodal sentiment analysis with missing and noisy modalities. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(3):2281–2289, 2026. doi: 10.1609/aaai.v40i3.37212.

Thomas P. Zollo, Zhun Deng, Jake C. Snell, Toniann Pitassi, and Richard Zemel. Improving predictor reliability with selective recalibration. *Transactions on Machine Learning Research*, 2024.

Shihao Zou, Wei Wei, and Yongshuo Zhang. DEAR: Distributional error-aware reliability for robust multimodal sentiment analysis with missing modalities. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 30344–30361, 2026. doi: 10.18653/v1/2026.findings-acl.1517.

APPENDIX: TABLE OF CONTENTS

A Appendix Guide and Notation **15**

B Additional Results **16**

C Scope and Limitations **25**

D Experimental Setup **26**

E Controlled Separation of Error, Ambiguity and Gap **28**

F Proofs for the Correction Reduction **29**
F.1 The correction reduction 29
F.2 Output-space dominance 30
F.3 Headroom and ambiguity 30
F.4 Cross-mask inheritance and transfer 31
F.5 Finite-sample harm control 34
F.6 Selecting the blend weight 35
F.7 The cross-mask accuracy screen 35

A APPENDIX GUIDE AND NOTATION

This appendix has four parts, ordered by what a reader is most likely to want first. §B gives every per-condition number behind the summary tables of §5, including the intact-input controls, where the frozen model sees every modality and correction is least expected to help. §C states where each result stops. §D–§E give the experimental setup and a synthetic family in which gap and ambiguity are controlled separately. §F.1–§F.7 prove the claims: the reduction to the Bayes gap, why editing outputs gives up nothing, how much of the observed error is reachable, the cross-mask inheritance identity and its closed form, the finite-sample harm bound, how the blend weight is chosen, and the one-sided accuracy screen used to study when the cross-mask target is plausible.
A reader checking the guarantee needs §F.5 alone. A reader checking that the target is well posed needs §F.4 and §F.7. Table 5 collects the notation.

Table 5: Notation used throughout the paper and this appendix.

Symbol	Meaning
Y	the label
$S \subseteq T$	modality masks; S observes a subset of what T observes
$X = (x_S, S)$	everything about a sample the corrector may read at deployment
$z(X)$	the frozen model's scores
$\ell(y, z)$, $\bar{\ell}(y, p)$	score-space loss and prediction loss, respectively
$m(z)$	the prediction in mean-parameter form; $\mathrm{softmax}(z)$ for cross-entropy
$t(y)$	the label statistic; e_y for cross-entropy, y for squared error
$\pi(X)$	$\mathbb{E}[t(Y) \mid X]$, the best prediction the observed inputs allow
$d(X)$	$m(z(X)) - \pi(X)$, the *Bayes gap*: the part of the error a corrector can reach
π_S, d_S, m_S	the same quantities under mask S
b	$\mathbb{E}[d_T \mid X_S]$, the richer mask's gap inherited by the poorer one
$\widehat{\pi}$	the frozen plug-in estimate of π, built from D_{source}
$k, \mathcal{N}_k(X)$	neighborhood size and the retrieved neighbors
p_0, p_{base}	the frozen output $m(z)$, and the harm reference/fallback, equal to p_0 throughout
$p_{\mathrm{temp}}, p_{\mathrm{corr}}$	temperature-scaled frozen output, and the candidate $(1 - \widehat{\beta})p_{\mathrm{temp}} + \widehat{\beta}\widehat{\pi}$
ϑ	temperature; $\vartheta = 1$ leaves probabilities unchanged
β, $\widehat{\beta}$	blend weight in $[0, 1]$, and its fitted value
$r_\theta(X)$	learned action score, fitted on D_{fit}
A_λ, L_λ, $\widehat{\lambda}$	threshold policy, its harmful-intervention loss, and the threshold calibrated by CRC
Δ_Y, α, δ	loss change $\bar{\ell}(Y, p_{\mathrm{corr}}) - \bar{\ell}(Y, p_{\mathrm{base}})$; harm budget and tolerance, 0.2 and 0.05
D_{source}	builds the target estimator
D_{fit}	selects k, target, space, weighting, temperature and $\widehat{\beta}$
D_{conf}, D_{test}	calibrates the gate; read once after every choice is fixed

B Additional Results

Tables 6–10 give every condition behind the summary of Table 2, including the intact-input conditions where the frozen model has nothing to correct. This section collects the per-condition figures behind the summary statements of §5.

Table 6: **DrugBAN under degraded inputs.** AUROC and accuracy for the frozen model and for GUARD, on the two in-domain *random* splits and the cross-domain *cluster* protocol defined by the model's authors. AUROC is one of the metrics DrugBAN reports and is threshold-free, so a gain there cannot come from moving a decision boundary. Mean over three seeds, except the two BindingDB *cluster* rows, which are single-seed results. † Protein truncated to 25%; ‡ Murcko scaffold with the protein truncated to 50%. *Retrain residual* is the part of the frozen model's accuracy drop that retraining does not recover: the shortfall a model retrained on the same degraded input still shows against one trained on intact input, mean over two initializations. It is an empirical residual under our recipe, not a measurement of intrinsic information loss; values near zero sit inside the spread of the two initializations. It is omitted on the *cluster* rows, where the total drop is smaller than its own noise. The *random* rows retrieve from the source population the released split defines; the *cluster* rows retrieve from the deployment domain, since under the authors' cross-domain protocol a source pool supplies neighbors from the wrong distribution and the gate declines most of them, leaving BindingDB *cluster*† at 0.536 AUROC with a 30% intervention rate instead of 0.790 at 70%.

		AUROC		Accuracy				
Dataset	**Split**	frozen	GUARD	frozen	GUARD	**Joint harm**	**Apply (%)**	**Retrain resid.**
BioSNAP	random†	0.680	**0.808**	0.571	**0.746**	0.165	54	+0.007
	random‡	0.750	**0.836**	0.675	**0.780**	0.159	49	+0.034
	cluster†	0.560	**0.635**	0.505	**0.602**	0.156	43	—
BindingDB	random†	0.758	**0.887**	0.678	**0.835**	0.142	71	+0.004
	random‡	0.800	**0.908**	0.722	**0.853**	0.163	67	+0.002
	cluster†	0.503	**0.790**	0.529	**0.731**	0.137	70	—
	cluster‡	0.522	**0.850**	0.528	**0.794**	0.111	74	—
Human	random†	0.948	**0.961**	0.869	**0.902**	0.049	59	−0.001
	random‡	0.944	**0.960**	0.861	**0.896**	0.108	62	+0.021

Table 7: **OPPORTUNITY under sensor loss.** Weighted F1, the metric DeepConvLSTM is published with, and accuracy, for the frozen model and for GUARD over four sensor configurations, mean over three seeds.

	Weighted F1		Accuracy			
Condition	frozen	GUARD	frozen	GUARD	**Joint harm**	**Apply (%)**
low-cost accels only	0.527	**0.732**	0.538	**0.741**	0.168	64
no IMU family	0.433	**0.764**	0.502	**0.783**	0.186	69
no shoe sensors	0.643	**0.788**	0.646	**0.796**	0.148	56
three sensors only	0.649	**0.729**	0.660	**0.735**	0.176	71

Harm among applied interventions. A small joint harm rate need not mean that most applied corrections help. For a fixed policy with positive apply probability, conditional harm equals joint harm divided by apply probability. If the apply probability is at most α, the joint constraint alone permits every applied correction to be harmful. Table 11 makes this distinction explicit on IEMOCAP, using the same 25 fold–cut runs per condition as Table 8.

For run r, let a_r count applied test predictions and h_r count those whose cross-entropy increase exceeds $\delta = 0.05$ relative to the raw frozen output. We report the pooled occurrence rate $\sum_r h_r / \sum_r a_r$ and, separately, the mean of h_r/a_r over runs with $a_r > 0$; the conditional rate is undefined when $a_r = 0$. Counts are recovered from the full-precision saved rates and the original test-split sizes, with an integer consistency check for every run. The repeated cuts reuse examples, so these are descriptive rates over evaluated occurrences, not independent-sample estimates with binomial confidence intervals.

Audio plus text is the clearest low-intervention case: only 3.74% of predictions are corrected on average, yet 296 of the 356 applied occurrences are harmful (83.15%), and mean accuracy falls

Table 8: **Certified correction across five further benchmarks.** The frozen model against GUARD at $\alpha = 0.2$, each block scored with the metric its dataset is published under, named beneath the dataset. *Intact* feeds the undegraded input back in: it isolates the headroom a frozen model retains when nothing is missing, which the theory predicts need not be zero. CMU-MOSEI binary accuracy is computed on non-neutral samples ($y \neq 0$), the standard protocol for that dataset; its rows average ten deployment splits, not the five of Table 1. Per-benchmark replication and protocol detail are in App. D.

Dataset (model)	**Available**	frozen	GUARD	**Joint harm**	**Apply (%)**
CMU-MOSEI (CMAD)	audio + visual	0.645	**0.703**	0.116	37
binary accuracy	audio only	0.631	**0.679**	0.094	30
	visual only	0.637	**0.683**	0.107	32
	intact	0.863	**0.864**	0.027	25
IEMOCAP (MoMKE)	audio only	0.617	**0.637**	0.121	42
4-*class WA*	text only	**0.745**	0.743	0.058	8
	visual only	**0.547**	0.542	0.093	36
	audio + text	**0.793**	0.791	0.031	4
	audio + visual	0.629	**0.640**	0.117	21
	text + visual	**0.750**	0.748	0.067	12
	intact	0.785	**0.786**	0.053	19
AVE (AV-att)	audio only	0.465	**0.600**	0.177	65
29-*way acc.*	visual only	0.462	**0.496**	0.178	57
	intact	**0.719**	0.718	0.169	85
NinaPro DB5 (sEMG CNN)	14 electrodes	0.752	**0.824**	0.139	97
accuracy	12 electrodes	0.713	**0.801**	0.136	97
	10 electrodes	0.680	**0.781**	0.122	97
	8 electrodes	0.664	**0.767**	0.125	97
	7 electrodes	0.654	**0.756**	0.123	97
	6 electrodes	0.647	**0.744**	0.118	97
	5 electrodes	0.642	**0.727**	0.128	98
	4 electrodes	0.632	**0.699**	0.139	98
	3 electrodes	0.627	**0.675**	0.147	98
	2 electrodes	0.623	**0.643**	0.145	75
	intact	0.817	**0.839**	0.143	97
PTB-XL (resnet1d_wang)	11 leads	**0.919**	0.918	0.078	47
macro AUC	10 leads	0.877	**0.882**	0.135	62
	9 leads	0.838	**0.863**	0.156	68
	8 leads	0.819	**0.856**	0.143	63
	7 leads	0.791	**0.840**	0.154	70
	6 leads	0.738	**0.836**	0.157	87
	5 leads	0.715	**0.838**	0.154	87
	4 leads	0.685	**0.832**	0.151	86
	3 leads	0.652	**0.832**	0.136	86
	2 leads	0.578	**0.826**	0.146	84
	intact	0.927	**0.928**	0.031	76

Table 9: **Joint harm and intervention rate on CMU-MOSEI.** Joint harm is the probability that a correction is applied and raises the loss by more than $\delta = 0.05$; the budget is $\alpha = 0.2$. *Blanket* corrects every prediction; *GUARD* corrects only where the calibrated action score admits the intervention. With every modality observed, the corrector has almost nothing to add: the gate admits about a quarter of the samples, accuracy moves by $+0.0005$, and joint harm is 0.027. These runs use the ten deployment splits behind Table 8, not the five behind Table 1.

Observed	**Blanket joint harm**	**GUARD joint harm**	**Apply (%)**
A	0.302	**0.094**	30
V	0.339	**0.107**	32
A, V	0.351	**0.116**	37
L, A, V	0.272	**0.027**	25

Table 10: **Effect of removing the gate.** Blanket correction is GUARD with CERTIFY removed: the same frozen corrector applied to every sample. Joint harm is given as the range over each benchmark's conditions, against the $\alpha = 0.2$ budget. Averaged over conditions, blanket joint harm runs from 1.6 to 9.5 times GUARD's, and breaches the budget on six of the seven benchmarks, where GUARD stays inside it on every condition of this comparison, all of which are run at $\alpha = 0.2$; where it happens to stay inside, nothing certifies that it will on the next distribution. The two policies stay closest at mild degradation, where the frozen model has little to correct. Ranges are over the conditions of the comparison in Table 3, and both tables are computed from the same per-condition file, so a mean here always lies inside its own range. That comparison selects on accuracy and, on some benchmarks, uses fewer conditions than Table 2, so its ranges need not match that table's. Δ is accuracy gain, with per-label accuracy on PTB-XL.

		Blanket correction		**GUARD**
Benchmark	**Conditions**	Δ	**joint harm**	**joint harm**
CMU-MOSEI	3	+0.067	0.302–0.351	**0.094–0.116**
IEMOCAP	7	+0.004	0.475–0.939	**0.031–0.121**
OPPORTUNITY	4	+0.183	0.256–0.302	**0.148–0.186**
AVE	2	+0.084	0.291–0.310	**0.177–0.178**
NinaPro DB5	4	+0.092	0.109–0.153	**0.066–0.084**
PTB-XL	6	+0.170	0.175–0.329	**0.126–0.164**
DrugBAN	3	+0.085	0.243–0.290	**0.135–0.187**

Table 11: **Joint and conditional harm on IEMOCAP.** Joint harm and apply rate are means over runs, as in Table 8. Each condition has 9,206 held-out occurrences across 25 runs. The pooled column divides the displayed harmful/applied counts; the run-mean column weights each nonempty run equally. *Zero* counts the runs with no intervention. Harm means a cross-entropy increase above 0.05, not necessarily an accuracy error.

	Joint	Apply	Harmful/	Conditional harm (%)		Zero
Available	harm	(%)	applied	pooled	run mean	runs
audio	0.121	42.18	1143/4018	28.45	30.05	0
text	0.058	7.84	535/727	73.59	67.12	3
visual	0.093	35.67	865/3230	26.78	36.13	0
audio + text	0.031	3.74	296/356	83.15	86.86	6
audio + visual	0.117	20.96	1097/1889	58.07	64.02	0
text + visual	0.067	11.90	613/1084	56.55	72.85	6
intact	0.053	19.28	478/1802	26.53	64.78	3

by 0.17 points. Text-only and text plus visual also combine majority-harmful interventions with small accuracy losses. These outcomes satisfy the stated joint-risk objective while offering poor conditional reliability. An application requiring reliable individual interventions would need an additional conditional-risk requirement and a separately calibrated policy; lowering the joint budget alone does not provide that guarantee.

Full CMU-MOSEI comparison. Table 1 gives the published comparison summarized in §5.3. The purpose of this table is contextual: the quoted methods are trained under their own missing-modality protocols, whereas GUARD leaves each reproduced model frozen and attaches the same post-hoc policy.

A released-code HME comparison on paired test partitions. Table 12 adds a direct run of the released HME model (Zhuang et al., 2025a) on the same word-aligned MOSEI archive as CMAD. HME is trained on the original training set and selected on the original validation set, then evaluated on exactly the held-out test quarters used by GUARD. All seven masks use these paired quarters, including text-only CMAD, whose unchanged predictions are recomputed here instead of using the full-test value in Table 1. The models and supervision regimes still differ: HME trains its representation and fusion layers, while GUARD keeps CMAD frozen and additionally uses labeled deployment pool, fit and calibration splits. This comparison supplies experimental context; the probe ablations below isolate the corrector under matched supervision.
The released-code run is competitive when language is present but weaker without it: mean accuracy over A, V, AV is 58.6%, below the corresponding published HME scores. The result therefore does

Table 12: **HME and CMAD on paired MOSEI test partitions.** Accuracy/weighted F1 (%), excluding neutral samples, averaged over five deployment partitions. HME uses one training seed and the released-code configuration described in App. D; the partitions are not independent training runs. HME receives no deployment labels, whereas CMAD + GUARD uses labeled pool/fit/conf splits. HME uses its own preprocessing and batch context.

Observed	HME (released-code run)	CMAD	CMAD + GUARD
L	86.0/85.9	86.6/86.6	86.6/86.6
A	58.7/52.6	62.9/58.9	67.6/65.1
V	59.1/59.7	63.0/61.6	67.6/66.2
LA	86.0/85.9	86.7/86.7	86.8/86.8
LV	86.7/86.6	86.8/86.8	86.6/86.6
AV	58.0/58.6	63.9/63.0	69.7/68.8
Six incomplete masks	72.4/71.6	75.0/73.9	77.5/76.7
Language absent (A, V, AV)	58.6/57.0	63.3/61.1	68.3/66.7
Intact (LAV)	86.6/86.6	86.8/86.8	86.8/86.8

not establish a ranking against HME under its original protocol. It also cannot attribute GUARD's gain to retrieval alone, since the deployment labels and model training differ.

Uncertainty of the summary gains. The per-cell files shipped with the code record every fold and deployment draw, so a summary gain can be read against its own spread. On IEMOCAP the six availability patterns of Table 2 give $+0.0035$ accuracy with a standard error of 0.0019 over twenty-five replicates, and eight of those replicates are negative: this benchmark's gain is not separated from zero at two standard errors. It is also the smallest gain in the suite, on the frozen model that absorbs most of its degradation, which is the outcome §5.2 leads us to expect. The same computation on the gate comparison of Table 3 gives $+0.178 \pm 0.045$ on OPPORTUNITY over three seeds of its four sensor configurations and $+0.051 \pm 0.003$ on CMU-MOSEI over its three availability patterns.

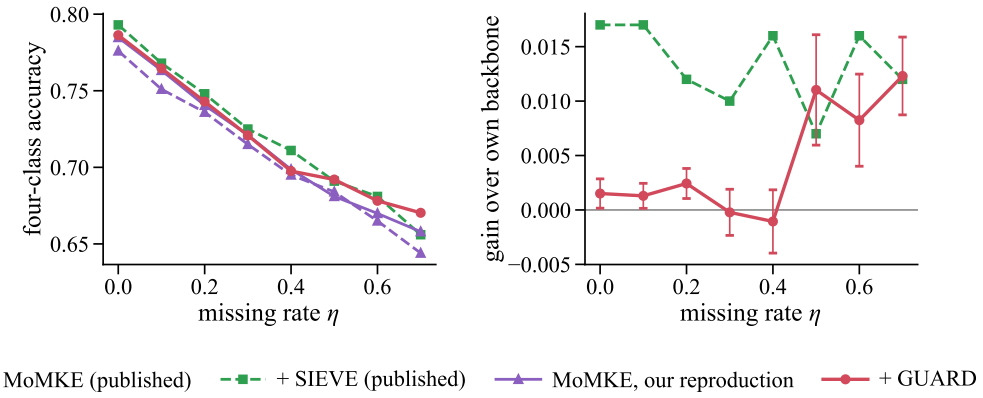

Figure 5: Accuracy and gain under different missing rates on IEMOCAP, both methods attached to the same backbone architecture. *Left:* four-class accuracy; the two MoMKE lines show the result quoted by Gao et al. (2026) and our reproduction. *Right:* gain over each method's own backbone. SIEVE trains the backbone and gains a roughly constant amount at every rate, including $\eta = 0$ where no modality is missing; GUARD leaves the weights frozen and gains only as the degradation grows. Underlying numbers are in Table 13.

Fitted blend weight on CMU-MOSEI. Over the three frozen models and seven modality subsets, $\widehat{\beta}$ averages 0.70 where language is missing against 0.12 where it is present, so the correction is scaled down rather than switched off: it is selected to zero on only two of the thirty-three language-keeping cells. The average gain ranges from 2.1 accuracy points on CMAD to 0.9 on MoMKE.

Effect of fit-split tightening. On AVE, where the two thresholds can be measured against each other directly, the conformal threshold alone realizes joint harm 0.196, against 0.178 once tightened at an intervention rate of 0.61; the step lowers harm on all six cells, and these are the numbers Table 3 reports.

Table 13: Accuracy under different missing rates on IEMOCAP. Four-class accuracy (Acc-4) is reported at missing rate η: every modality is absent independently with probability η, and a sample that would lose all three keeps one, the protocol the quoted rows are scored under. "Avg." refers to the average over all eight rates, and Δ is the gain of each method over the backbone directly above it. Our two rows are the mean over five folds and five deployment draws; the gain averaged over rates is $+.004$ with a standard error of .001, and it is concentrated at $\eta \geq 0.5$. Rows marked † are quoted from Gao et al. (2026) and update the deployed model's weights; GUARD leaves them frozen, so the two are not run under the same constraint and only the Δ column is comparable across blocks.

Method	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	**Avg.**	Δ
MoMKE (Xu et al., 2024)†	.776	.751	.736	.715	.695	.684	.665	.644	.708	—
+ SIEVE (Gao et al., 2026)†	.793	.768	.748	.725	.711	.691	.681	.656	.722	+.013
MoMKE, our reproduction (frozen)	.785	.763	.740	.721	.699	.681	.670	.658	.715	—
+ GUARD	.786	.765	.743	.721	.698	.692	.678	.670	.719	+.004

A trained probe as the corrector. §4.1 treats the corrector as a plug-in any algorithm may supply. Table 14 holds the frozen model and the splits fixed, re-runs the same fitting and calibration procedure for each corrector, since L_λ depends on the corrector and a threshold calibrated for one does not transfer to another, and replaces the neighbor average by a probe trained on the same D_{source}. Neither corrector dominates. The probe moves accuracy gain by at most 2.4 points in either direction; PTB-XL uses per-label accuracy. It has a higher harm rate on four of the five benchmarks. Both correctors stay inside the budget on every benchmark, and removing the gate puts both over it on four of the five. CERTIFY is therefore not tied to the estimator it certifies.

The probe is a multinomial logistic model on the pool labels for the hard-label target and a ridge regression onto the richer mask's outputs for the cross-mask target, each selecting its regularization on D_{fit} by the rule the retrieval grid uses. Recomputed on the same splits, its retrieval column reproduces the GUARD column of Table 3 on every benchmark. On DrugBAN's original mixture of source-pool conditions, the probe has lower gated gain and higher ungated harm (0.359 against retrieval's 0.275). That aggregate does not isolate a cross-domain advantage; the matched deployment-pool comparison below tests that setting separately. Ungated, the probe reaches 0.757 joint harm on IEMOCAP against a budget of 0.2.

Table 14: **The same certificate over two correctors.** Gated gain, joint harm and ungated joint harm at $\alpha = 0.2$, for the retrieval plug-in of §4.1 and for a probe trained on the same D_{source}. Everything but the corrector is held fixed, and the retrieval column is recomputed here rather than copied, so it reproduces the GUARD column of Table 3. Bold marks the better of the two; * marks harm above the budget. Five benchmarks from Table 3; Table 15 adds a matched CMU-MOSEI comparison and deployment-pool DrugBAN comparisons. OPPORTUNITY is not included in this probe ablation.

	Gated gain		Joint harm		Ungated harm	
Benchmark	Retrieval	Probe	Retrieval	Probe	Retrieval	Probe
IEMOCAP	$+0.003$	$\mathbf{+0.008}$	$\mathbf{0.077}$	0.089	0.734^*	0.757^*
AVE	$+0.084$	$\mathbf{+0.108}$	$\mathbf{0.178}$	0.184	0.300^*	0.285^*
NinaPro DB5	$+0.092$	$\mathbf{+0.093}$	$\mathbf{0.072}$	0.081	0.125	0.118
PTB-XL	$+0.165$	$\mathbf{+0.172}$	0.150	$\mathbf{0.140}$	0.239^*	0.215^*
DrugBAN	$\mathbf{+0.084}$	$+0.067$	$\mathbf{0.150}$	0.153	0.275^*	0.359^*

Matched supervision on MOSEI and deployment-pool DrugBAN. Table 15 extends the comparison to CMU-MOSEI and to DrugBAN's two cluster datasets with deployment-matched retrieval pools. Both correctors use the same labeled splits and the shared fitting and certification procedure, with separate hyperparameter selection on D_{fit}. Harm is measured against the raw frozen output. The MOSEI retrieval control reproduces all 30 saved cells behind Table 3. For DrugBAN, each fixed seed-42 host uses three partitions of its deployment calibration data into disjoint pool, fit and calibration thirds, keeping the original test set held out. These rows select by fit accuracy, so they are separate from the AUROC-selected comparison in Table 6.

Retrieval has the larger gain on MOSEI, while the probe intervenes less and has lower joint harm. On both DrugBAN cluster datasets, the deployment-pool probe has higher mean gain and lower mean joint harm. Thus the original source-pool aggregate does not establish a general advantage for

Table 15: **Retrieval and a probe with matched deployment supervision.** Accuracy gain, joint harm and apply rate at $\alpha = 0.2$. CMU-MOSEI averages A, V and AV over ten splits and scores accuracy on non-neutral samples. The DrugBAN rows retain 50% of the protein and average three calibration partitions of a fixed host, with both correctors trained on the same deployment pool.

	Gated gain		Joint harm		Apply (%)	
Benchmark	Retrieval	Probe	Retrieval	Probe	Retrieval	Probe
CMU-MOSEI	$+0.051$	$+0.026$	0.106	0.088	32.9	29.0
BioSNAP cluster	$+0.101$	$+0.130$	0.140	0.130	39.7	41.4
BindingDB cluster	$+0.213$	$+0.224$	0.151	0.129	77.1	74.1

retrieval under domain shift. The choice of estimator depends on the available pool and the task; the certificate applies to either.

Varying the retrieval pool. Table 16 varies one offline role at a time and keeps the others fixed. On CMU-MOSEI, D_{source} is carved from deployment data the models never trained on. There the unlabeled cross-mask target keeps more gain than hard labels at every pool size. Flipping pool labels removes the hard-label gain, while GUARD switches to cross-mask on D_{fit} and keeps its own. A smaller pool lowers gain and joint harm together, because the gate declines more often. On DrugBAN, hard-label gain grows steadily with pool size. Shrinking D_{conf} keeps mean joint harm inside the budget and barely moves gain, since the fit-split tightening often sets the threshold. Individual calibration draws spread more widely at small sizes, up to 0.32 at 50 points on DrugBAN, as a marginal bound allows. Under the cluster shift, origin outweighs size: 1.2k–2.4k deployment rows beat the 10k–12k-row source pool, and adding the source pool to them does not help.
On the DrugBAN random splits of Table 4, D_{source} is a subsample of the model's training data. The all-inputs model is nearly exact there (0.98–1.00 accuracy, against 0.83–0.94 on test), so its outputs partly restate training labels. We therefore also carve the pool from deployment data the model never trained on. Cross-mask still matches or exceeds hard-label gain: $+0.007$ against $+0.004$ on Human, $+0.075$ against $+0.071$ on BindingDB and $+0.077$ against $+0.076$ on BioSNAP, with every run inside the budget.

Correction latency and memory. A probe has an operational advantage when its utility is competitive. Table 17 measures correction overhead on saved frozen representations, including standardization, target construction, blending, scoring and gating. The controlled configuration uses hard-label targets, a standardized feature space, uniform $k = 20$ retrieval or a linear probe with regularization $C = 1$, and unit temperature. Each corrector fits its own blend and gate. These fixed configurations isolate implementation cost; they are not the per-cell selected winners of Table 14 and make no new accuracy claim.
All measurements use float64 on one pinned Intel Xeon Gold 6244 CPU thread, NumPy 2.5.2 and scikit-learn 1.9.0. After ten warm-up calls, we time 200 single-query calls and 50 batches of 64, using the same sampled inputs for both correctors. Frozen feature extraction, disk I/O, fitting and calibration are excluded. The timed path is checked against the canonical batch outputs before measurement. Memory denotes numerical arrays retained by the deployed corrector, transform and scorer, excluding the common host and Python runtime; it is not total process RSS.
The probe is approximately 14–31 times faster for single queries here. With batches of 64, median cost per query is 0.228–0.497 ms for retrieval and 0.0077–0.0168 ms for the probe. Exact retrieval stores $O(N(d + K))$ numerical values and searches all N pool rows per query; a linear probe needs $O(dK)$ coefficients and arithmetic independent of N. The current multi-label probe also retains its pool-label array in a closure, which is included in PTB-XL's measured state. The retrieval implementation recomputes pool norms on each call; these measurements include that work. Distance weighting additionally recomputes a pool bandwidth in the current implementation and is outside this uniform-weight benchmark. Caching fixed pool quantities or using an approximate index would change the latency comparison and has not been measured here. When utility is comparable, these results favor a probe for latency and memory; retrieval remains an empirical choice, not a requirement of the certificate.

Other action-score learners and feature subsets. Proposition 2 holds for any score fixed before D_{conf}, so this ablation can move gated gain and the intervention rate but not the bound. Table 18 holds the frozen model, the splits, the corrector, the blend and the budget fixed and varies only the map from the five deployment-computable features of App. F.6 to a score. Each learner keeps its

Table 16: **Varying one offline role at a time**, at $\alpha = 0.2$. (a) CMU-MOSEI: three frozen models, modality subsets A, V and AV, five splits; accuracy gain on non-neutral samples as D_{source} shrinks, and as a fraction of its labels is flipped at full size. (b) DrugBAN random splits, three datasets, three seeds and the two conditions of Table 4: hard-label AUROC gain as the source pool shrinks, three subsamples per size. (c) Gain and mean joint harm as D_{conf} shrinks, over 8 (MOSEI) or 20 (DrugBAN, seed 42) subsamples per size. (d) DrugBAN cluster splits: AUROC gain of GUARD by pool origin; *matched* subsamples the source pool to the size of the deployment pool, *source* uses all of it, and *both* adds it to the deployment pool.

(a) CMU-MOSEI $\lvert D_{\text{source}}\rvert$	50	150	400	full	
Hard-label gain	+0.001	−0.003	+0.004	+0.015	
Cross-mask gain	+0.001	+0.004	+0.014	+0.035	
GUARD gain	+0.000	+0.002	+0.013	+0.035	
GUARD joint harm	0.025	0.041	0.081	0.094	
Flipped pool labels	0	0.1	0.2	0.3	0.4
Hard-label gain	+0.015	+0.008	+0.005	+0.003	−0.000
GUARD gain	+0.035	+0.035	+0.034	+0.035	+0.035
GUARD picks cross-mask (%)	100	96	93	98	98
(b) DrugBAN, random $\lvert D_{\text{source}}\rvert$	100	300	1000	3000	full
Hard-label AUROC gain	+0.008	+0.016	+0.033	+0.053	+0.079
Joint harm	0.144	0.148	0.145	0.131	0.128
(c) Calibration size $\lvert D_{\text{conf}}\rvert$	50	100	200	full	
CMU-MOSEI gain	+0.033	+0.035	+0.035	+0.035	
CMU-MOSEI joint harm	0.086	0.091	0.093	0.094	
DrugBAN AUROC gain	+0.077	+0.076	+0.076	+0.076	
DrugBAN joint harm	0.136	0.138	0.132	0.133	
(d) DrugBAN, cluster Pool	deployment	matched	source	both	
BioSNAP	+0.091	+0.042	+0.048	+0.086	
BindingDB	+0.307	+0.043	+0.032	+0.303	

Table 17: **Measured correction overhead at fixed settings.** One representative condition per benchmark: audio-only AVE and IEMOCAP, the six limb leads on PTB-XL, and four-electrode NinaPro. N is the retrieval-pool size, d the feature dimension and K the output dimension. Latency is single-query median/95th percentile in milliseconds; retained state is in MiB.

				Latency (ms)		State (MiB)	
Benchmark	N	d	K	Retrieval	Probe	Retrieval	Probe
AVE	33280	64	29	6.786/6.985	0.398/0.430	23.614	0.016
PTB-XL	17084	256	5	17.517/17.900	0.898/0.938	34.023	0.666
IEMOCAP	4446	768	4	11.195/11.540	0.366/0.397	26.198	0.035
NinaPro	14259	128	41	5.544/5.768	0.403/0.424	18.387	0.043

library default apart from one size knob, so no variant is tuned. The logistic row is the deployed policy recomputed here; it reproduces the retrieval column of Table 14 on all five benchmarks and is this table's control.

Across these five benchmarks the choice of score family has limited effect on gated gain. No learner beats the deployed logistic by more than 0.002 gain anywhere, and the variant that ranks best by AUROC is usually not the one that gains most: they coincide only on NinaPro and PTB-XL. A better ranking moves the policy along the same trade-off, since the calibrated threshold sets how much harm is admitted. The random forest shows this most clearly: it is the most conservative variant on every benchmark and gives up a third of the gain on PTB-XL to stay there.

The two feature blocks should be read together. Removing any single feature costs almost nothing, with one exception, AVE without the ℓ_1 distance. Read alone, that block would suggest the features are redundant. The keep-one block shows instead that they are redundant *with each other*: no single feature reproduces the full score everywhere. The ℓ_1 distance alone matches on AVE and NinaPro but loses a fifth of the gain on DrugBAN and all of it on IEMOCAP, while base confidence alone matches on PTB-XL and IEMOCAP and loses a sixth on AVE and DrugBAN. The agreement flag is the exception, and it illustrates the degeneracy noted in App. F.6 rather than a property of the gate: on AVE and NinaPro the indicator is identically one, so the score is constant and conformal risk control has only two admissible policies. On AVE, where ungated harm is 0.300, it admits nothing and the gate closes; on NinaPro, where ungated harm is 0.125, it admits everything and the row reproduces blanket correction. In both cases the budget is respected, as Proposition 2 requires; an uninformative score reduces utility but does not affect validity.

Table 18: **Which learner, and which features, for the action score.** Gated gain and joint harm at $\alpha = 0.2$, written gain/harm. Everything but the score is held fixed, and the logistic row is recomputed rather than copied, so it reproduces the retrieval column of Table 14. The same five benchmarks as that table, over 6, 80, 18, 175 and 27 evaluation cells, respectively.

Action score	AVE	NinaPro	PTB-XL	IEMOCAP	DrugBAN
Score family, all five features					
logistic (deployed)	$+0.084/0.18$	$+0.092/0.07$	$+0.165/0.15$	$+0.003/0.08$	$+0.084/0.15$
decision tree, depth 3	$+0.083/0.17$	$+0.092/0.08$	$+0.162/0.15$	$+0.004/0.07$	$+0.079/0.12$
random forest, 200 trees	$+0.072/0.11$	$+0.086/0.04$	$+0.112/0.08$	$+0.003/0.05$	$+0.067/0.06$
kNN, $k = 25$	$+0.083/0.18$	$+0.092/0.08$	$+0.163/0.15$	$+0.003/0.09$	$+0.084/0.15$
MLP, one hidden layer of 16	$+0.084/0.17$	$+0.092/0.07$	$+0.164/0.15$	$+0.004/0.08$	$+0.084/0.15$
Logistic score, one feature removed					
without base confidence	$+0.084/0.17$	$+0.091/0.07$	$+0.162/0.16$	$+0.003/0.08$	$+0.084/0.15$
without base entropy	$+0.084/0.18$	$+0.092/0.07$	$+0.165/0.15$	$+0.004/0.07$	$+0.084/0.16$
without target confidence	$+0.084/0.18$	$+0.091/0.07$	$+0.164/0.15$	$+0.003/0.07$	$+0.083/0.16$
without agreement flag	$+0.084/0.18$	$+0.092/0.07$	$+0.165/0.15$	$+0.004/0.08$	$+0.084/0.15$
without ℓ_1 distance	$+0.050/0.18$	$+0.089/0.10$	$+0.165/0.16$	$+0.004/0.07$	$+0.082/0.16$
Logistic score, one feature only					
base confidence only	$+0.071/0.18$	$+0.090/0.11$	$+0.165/0.16$	$+0.004/0.03$	$+0.070/0.16$
base entropy only	$+0.071/0.18$	$+0.090/0.11$	$+0.158/0.17$	$+0.003/0.06$	$+0.070/0.16$
target confidence only	$+0.039/0.18$	$+0.087/0.10$	$+0.158/0.18$	$-0.001/0.08$	$+0.070/0.16$
agreement flag only	$+0.000/0.00$	$+0.092/0.12$	$+0.080/0.06$	$-0.000/0.02$	$+0.018/0.06$
ℓ_1 distance only	$+0.084/0.18$	$+0.091/0.07$	$+0.153/0.16$	$-0.000/0.08$	$+0.068/0.14$

Exchangeability stress tests. Calibrating on one subject and deploying on another makes the gate more conservative: OPPORTUNITY's intervention rate falls from 0.650 to 0.615 and joint harm with it from 0.169 to 0.147. Calibrating at a mild degradation and deploying at a severe one is the damaging direction, since calibration then looks easier than deployment: PTB-XL's rate rises from 0.739 to 0.844 and joint harm from 0.141 to 0.144, still inside the budget.

Budget sweep and the missing-rate protocol. Table 19 and Figure 6 sweep the budget α. Figure 5 and Table 13 compare GUARD with SIEVE on IEMOCAP under the missing-rate protocol of Gao et al. (2026).

Harm tolerance. Table 20 sweeps δ with the retrieval settings fixed on D_{fit}, so only the harm event moves. Gain and joint harm barely change between $\delta = 0$, where any loss increase counts as harm, and $\delta = 0.2$. Joint harm stays inside the budget in every run except two of 26 DrugBAN runs at $\delta = 0.2$. The flip columns count decisions rather than losses: applied corrections that turn a correct frozen decision wrong, and those that turn a wrong one correct. The second outnumbers the first by about 1.7 to one on CMU-MOSEI and 2.6 to one on DrugBAN.

Table 19: **What the budget costs.** Mean over 211 exchangeable runs at each budget, pooled across every benchmark but IEMOCAP. *Over* α is the share of individual runs whose realized joint harm exceeds the budget. The last row applies the same corrector everywhere and does not use the budget α. Per-benchmark curves are in Figure 6.

Budget α	**Joint harm**	**Gated gain**	**Apply**	**Over** α
0.05	0.041	+0.062	0.48	20.4%
0.10	0.081	+0.076	0.62	9.0%
0.20	0.135	+0.084	0.77	2.4%
0.30	0.154	+0.085	0.81	0.9%
0.50	0.169	+0.085	0.83	0.0%
ungated	0.221	+0.090	1.00	—

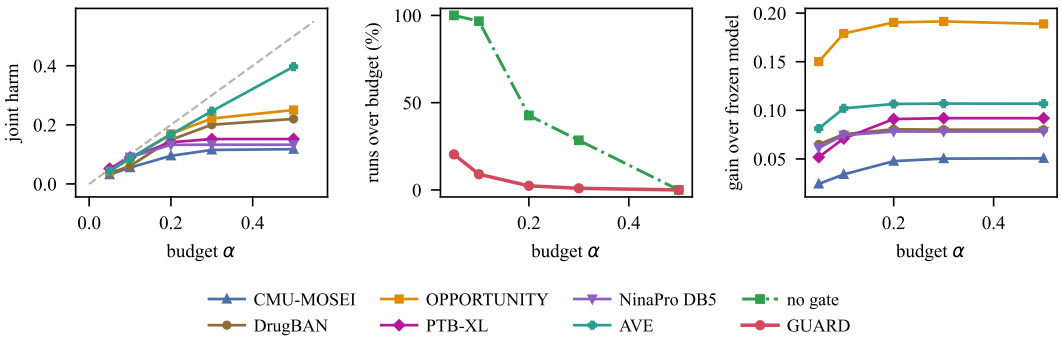

Figure 6: Risk and utility under different budgets α, over 1055 runs spanning every benchmark but IEMOCAP, which is not swept. *Left:* empirical joint harm against the budget, with the dashed diagonal at α; every per-benchmark mean stays below it. *Middle:* the percentage of runs over budget, for GUARD and for the same corrector ungated. *Right:* gated gain in each benchmark's own metric, which rises with the budget and levels off by $\alpha = 0.3$. Target, neighborhood and blend weight are selected once on D_{fit} and held fixed, so α alone varies and the ungated corrector is the same at every budget. Intervention rates are in Figure 7.

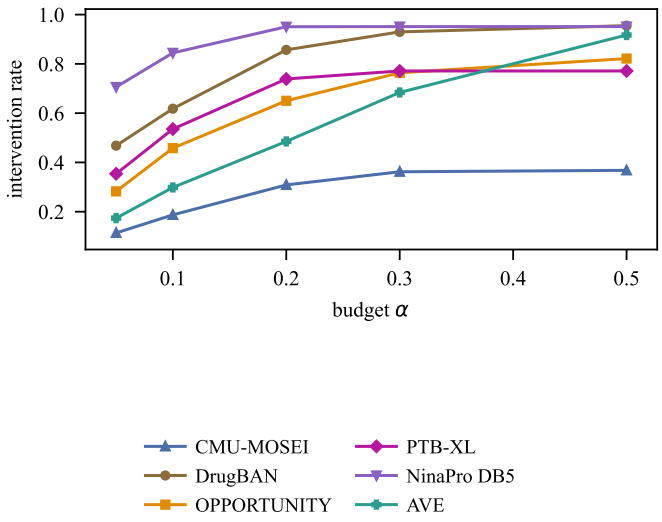

Figure 7: Intervention rate against the budget. Before tightening, a larger budget can only lower the CRC threshold for a fixed scorer and calibration sample. The plotted policy also applies fit-split tightening, so its intervention rate need not be monotone and may saturate below one.

Table 20: **Harm tolerance** δ at $\alpha = 0.2$. CMU-MOSEI: three frozen models, modality subsets A, V and AV, five splits; gain in accuracy on non-neutral samples. DrugBAN: the settings of Table 4 under the two conditions used there; gain in AUROC. ***Flip***$^-$ and ***flip***$^+$ are the shares of test samples on which an applied correction turns a correct frozen decision wrong, or a wrong one correct.

	CMU-MOSEI				DrugBAN			
δ	Gain	Joint harm	Flip$^-$	Flip$^+$	Gain	Joint harm	Flip$^-$	Flip$^+$
0	+0.034	0.093	0.034	0.059	+0.099	0.135	0.070	0.187
0.02	+0.035	0.097	0.036	0.062	+0.099	0.134	0.071	0.188
0.05	+0.035	0.094	0.037	0.063	+0.100	0.133	0.072	0.189
0.1	+0.036	0.087	0.040	0.067	+0.101	0.134	0.074	0.192
0.2	+0.037	0.059	0.043	0.071	+0.105	0.136	0.078	0.196

Retrieval without labels. Table 4 reports the empirical cost of replacing retrieval labels with richer-mask frozen outputs. App. F.7 analyzes the observed separation further. Population accuracy gives a necessary condition for conditional correctness; its finite-sample analogue is used here as a descriptive screen, not a formal rejection test or a sufficient deployment rule.

C Scope and Limitations

This appendix specifies the domain over which each result of the paper applies.

Frozen-model scope. Theorem 1 is exact for the information supplied to a corrector. If GUARD reads only $U = \phi(X)$, including the frozen prediction and retrieval features, its Bayes target is $\pi_U = \mathbb{E}[t(Y) \mid U]$. For square-integrable $t(Y)$ and a square-integrable U-measurable prediction p,

$$\mathbb{E}\|p - \pi(X)\|^2 = \mathbb{E}\|p - \pi_U\|^2 + \mathbb{E}\|\pi_U - \pi(X)\|^2. \tag{12}$$

The cross term vanishes by conditioning on U. The second term is information lost by the accessible representation and cannot be removed by a U-measurable correction. Retrieval in that space need not estimate $\pi(X)$ itself. Feature-space reduction concerns a local quadratic surrogate; output-space containment is exact for measurable heads and unrestricted measurable output edits (Props. 3, 5). Containment is an expressivity statement, not an algorithm for computing a feature repair without access to the head.

Data and access requirements. GUARD reads the deployed model's predictions and representations and modifies neither. Beyond that it assumes labeled data drawn from the deployment condition: D_{source} supplies the retrieval pool, D_{fit} fits the blend weight and the action score, and D_{conf} calibrates the threshold. The default hard-label target requires labels for the pool, and a pool drawn from a distribution far from deployment reduces utility without affecting validity. Cross-mask transfer removes labels from retrieval only; the accuracy screen, $\widehat{\beta}$-selection and CERTIFY continue to use them.

Scope of the statistical guarantee. The present instantiation of CERTIFY applies to classification. Nothing in conformal risk control is specific to that case, but the action score and its features are defined over a predicted distribution, so a regression instantiation would redefine both. Proposition 2 assumes the calibration split and the deployment sample are exchangeable, and it controls joint marginal harm: it bounds the rate at which a harmful correction is applied, not the safety of any single query and not the harm conditional on the decision to apply. Distribution shift after calibration falls outside the formal guarantee; App. B reports robustness under two deliberate violations, and App. F.5 gives the formal statement and the mask-conditional variant.

Scope of the cross-mask result. Population recovery holds when $\mathbb{E}[d_T \mid X_S] = 0$; recovery by finite-pool estimators additionally requires consistency. The observed sigma-fields, including mask context, must be nested. A conditionally correct richer predictor suffices for recovery, and Proposition 9 uses this stronger condition with matched weights and blend. It is a convex-risk comparison before adaptive selection and gating, and need not hold after tempering the teacher. It does not imply dominance in accuracy or gated gain. Because conditional correctness is not identifiable from unlabeled paired data alone (App. F.4), no assumption-free unlabeled diagnostic can certify it in general. GUARD selects its target on labeled D_{fit}; App. F.7 provides a descriptive accuracy screen. Realizing correction headroom additionally depends on the quality of the target estimator.

D EXPERIMENTAL SETUP

Table 2 names the nine frozen model–dataset pairs and the six domains they span. Each is anchored to its own published number before any correction is attached; the degradations applied, the provenance of each model and the details of its reproduction follow.

Each condition assigns disjoint roles to D_{source}, D_{fit}, D_{conf} and D_{test}, with every decision rule frozen before test evaluation. Unless stated otherwise $\alpha = 0.2$, $\delta = 0.05$, and $\widehat{\beta}$ is chosen on D_{fit} by the loss-minimizing rule of §4.1; every combination of dataset, condition and seed is reported. The neighborhood size k, the retrieval target, the retrieval space, the neighbor weighting and the temperature are all selected on D_{fit} by the metric the benchmark is reported under, except inside the rule comparison of Table 3, which selects on accuracy throughout; the test split is read once, after selection. Table 22 gives the size of each role and the grids searched.

Prediction loss and temperature reference. Write $p_0 = m(z)$ for the original frozen output. For categorical predictions, $\mathrm{Temp}_\vartheta(p)_c = p_c^{1/\vartheta} / \sum_j p_j^{1/\vartheta}$; for multi-label outputs it applies $\mathrm{sigmoid}(\mathrm{logit}(p_c)/\vartheta)$ coordinatewise. The selected temperature may also be applied to the richer-view retrieval values. All such choices are fixed before calibration.

Every evaluation in the paper uses $p_{\text{base}} = p_0$: the temperature belongs to the candidate, while the action features, the harm event and the prediction a rejected query keeps all read the raw frozen output. Positive temperature scaling preserves class decisions and rankings, so it leaves every frozen accuracy, F1 and AUROC entry unchanged. For the same reason, temperature scaling alone gains nothing under any metric reported here.

Released-code HME protocol. The additional HME run uses the CMAD archive of $16{,}265/1{,}869/4{,}643$ train/dev/test clips and the official HME model at commit `cdb1d60`, with pretrained BERT and Transformers 4.30.2/PyTorch 1.12.1. Following the fixed-missing setting, training observes all modalities; validation evaluates the six incomplete masks. We retain the released code's hidden size 192, transformer/hyper/latent depths $2/3/3$, batch size 256, learning rate $2 \cdot 10^{-5}$, objective $L_{\text{task}} + 0.8L_{\text{aux}} + 0.01L_{\text{info}}$, and AdamW reset each epoch. These settings follow the code rather than the different learning-rate and loss description in HME's paper. BERT activation checkpointing reduces memory without changing the batch. Seed 5576 trains for at most 100 epochs, with patience 10 on mean per-example validation loss over the six masks; epoch 54 is selected and training stops after epoch 64. Test metrics are computed only afterward. Each of five seed-0–4 deployment permutations uses the final quarter as test, with batch size 128 and fixed evaluation seeds. HME retains its stochastic VIB inference and within-batch context; pool/fit/conf examples do not enter its test batches. Source features and label order match the CMAD archive, and perturbing only masked inputs leaves HME's predictions unchanged. Table 12 reports a single training run, not an estimate of training-seed variability.

Degradation axes of Figure 4. The panels sweep, in turn, a nested lead ladder that removes precordial before limb leads, extending rather than reproducing the configurations of Reyna et al. (2021); the electrode budget; OPPORTUNITY's four sensor subsets; the fraction of target protein retained; and the deployment missing rate on IEMOCAP and AVE. The DrugBAN panel is BindingDB alone, scored by AUROC as in Table 2; `scripts/drugban_auroc_ladder.py` rescores the saved ladder under that metric.

Comparators of Table 3. Neither *Conf. score* nor *LTT calib.* carries the fit-split tightening of §4.2, so each differs from GUARD in that step as well as in the one it is named for. The two calibrations control different things: conformal risk control bounds the risk in expectation over the calibration draw, Learn-then-Test bounds the risk of the threshold it selects with probability $1 - \zeta$ under its i.i.d. sampling assumption ($\zeta = 0.05$), and neither bounds the harm realized on a given split. On PTB-XL, its separately fitted logistic score uses $-\sum_c p_c \log p_c$, whereas GUARD uses the sum of Bernoulli entropies; that row consequently compares both score construction and calibration. Because the table scores every benchmark in accuracy, its numbers differ from the published metrics of Table 2: OPPORTUNITY's weighted-F1 gain is $+.190$ against the $+.178$ in accuracy shown there, DrugBAN is reported elsewhere by AUROC and PTB-XL by macro AUC.

Choosing the degradation levels. PTB-XL's ladder extends the reduced-lead configurations of the PhysioNet/CinC challenge (Reyna et al., 2021). NinaPro's electrode ladder follows a literature that asks how few sEMG channels a gesture classifier can be built on, which reports usable accuracy well below the full array (Naik et al., 2016; Pelaez Murciego et al., 2022; Simon et al., 2024); we sweep

the range rather than adopt one recommended count. The remaining axes are the benchmarks' own availability patterns, in the spirit of tolerance evaluations such as Zheng et al. (2026).

Baselines. We compare against two families. *Without a guarantee*: the frozen model itself; blanket correction; the confidence scores standard in selective prediction, namely maximum softmax probability (Hendrycks & Gimpel, 2017), its margin, and negative entropy (Geifman & El-Yaniv, 2017); agreement between the model and the retrieval target; a gate fitted on D_{fit} to predict whether correction helps, which transposes the per-sample routing principle of Gao et al. (2026) to a frozen model; and published missing-modality methods quoted from their own tables. *With a guarantee*: two rules carry a finite-sample harm guarantee, possible because harmful application is observable on labeled calibration data. *Conf. score* applies conformal risk control to the model's own confidence, and *LTT calib.* calibrates a learned logistic score by Learn-then-Test (Angelopoulos et al., 2025), with Hoeffding–Bentkus p-values tested in a fixed order chosen on D_{fit}, stopping at the first non-rejection, at significance level $\zeta = 0.05$, which is a distinct quantity from the tolerated loss increase δ. Table 3 therefore compares guarantees as well as decision rules; what it cannot compare is the published missing-modality methods, which are quoted from their own tables and carry no such statement.

Table 21: **Benchmarks, frozen models, and degradations.** *Anchor* compares each frozen model against its published number before any correction is attached; where a model is published across many availability patterns we give the worst gap over all of them, and Human has no directly comparable published entry. * The only released weights we could use. † Our own model, so its anchor is the dataset paper's protocol rather than a reproduction.

Dataset	Frozen model	Anchor: published → ours	Degradations
BioSNAP	DrugBAN	AUROC .903 → .904	protein 25%, 50%; scaffold
BindingDB	DrugBAN	AUROC .960 → .963	protein 25%, 50%; scaffold
Human	DrugBAN	—	protein 25%, 50%
OPPORTUNITY	DeepConvLSTM	wtd. F1 .895 → .879	four sensor subsets
CMU-MOSEI	CMAD	acc. within .015	audio, visual, or both
IEMOCAP	MoMKE	WA .801 → .778 full; short sparse	audio only
AVE	AV-att*	acc. .7270 → .7274	audio only; visual only
NinaPro DB5	sEMG CNN†	acc. .690 → .814	16 down to 2 of 16 electrodes
PTB-XL	resnet1d_wang	mAUC .930 → .927	11 down to 2 leads, precordial first

Provenance of the frozen models. Every model we correct is published except NinaPro's sEMG CNN, which is our own and follows the dataset paper's protocol (Pizzolato et al., 2017). Only AVE ships usable weights (Tian et al., 2018); the rest we retrain from the authors' released code (Bai et al., 2023; Ordóñez & Roggen, 2016; Strodthoff et al., 2021; Xu et al., 2024; Zhuang et al., 2025b). IEMOCAP follows the same rule as every other benchmark: $\widehat{\beta}$ is selected on each fold's own fit third. The folds' evaluation videos are disjoint and each fold's fit third shares no video with its calibration or test third. Because drug–target interaction has no established missing-modality protocol, its degradations were fixed in advance.

Reproducing each published model. Table 21 compares each frozen model with its published reference performance. Two anchors fall short and we report them without further tuning. DrugBAN's BioSNAP *cluster* split reaches 0.591–0.619 AUROC over three seeds against a published 0.654 ± 0.023. MoMKE reproduces its full-modality number to within 2.4 points on each fold's whole evaluation set (.778); the intact rows of Tables 8 and 13 give .785 because they score only the held-out part GUARD evaluates on. It does not reproduce its sparse numbers, falling 13.8 points short on audio only; the availability-pattern results for that model should be read against our own frozen row, not against its published table.

Loading the AVE checkpoints. The AVE checkpoints were released for PyTorch 0.3.0 and are loaded here under PyTorch 2.13. The pickled LSTM predates the `_flat_weight_refs` attribute that current PyTorch reads on every forward pass; without it the model loads without error and emits a constant output. We restore the attribute from `_flat_weights` at load time and fill any missing `nn.Module` attributes from a freshly constructed template.

Table 22: Sample counts per data role, and the grids searched on D_{fit}. Roles are disjoint at the row level. Counts are per condition; IEMOCAP deployment roles are split by video and OPPORTUNITY roles by contiguous blocks, while the other deployment subdivisions are by row within the evaluated population. OPPORTUNITY counts are means over seeds; IEMOCAP counts are means over its five folds, whose evaluation sets range from 1023 to 1241 samples; NinaPro counts are per subject.

Benchmark	$\lvert D_{\text{source}}\rvert$	$\lvert D_{\text{fit}}\rvert$	$\lvert D_{\text{conf}}\rvert$	$\lvert D_{\text{test}}\rvert$	**Construction**
CMU-MOSEI	1161	1161	1161	1160	four disjoint quarters
IEMOCAP	4425	436	344	325	pool from other folds; fold split by video
AVE	33280	1340	1340	1340	released pool; deployment split in thirds
NinaPro	14259	2237	2237	2238	pool from other repetitions; rest in thirds
OPPORTUNITY	2772	2735	2574	2500	random 200-window blocks
PTB-XL	17084	719	719	720	official patient folds
DrugBAN (BindingDB)	12000	2460	2460	9840	released split

Grids searched on D_{fit}, scored by the metric the benchmark reports	
Neighborhood k	$\{3, 5, 8, 12, 20, 35, 50\}$
Retrieval target	hard label, cross-mask
Retrieval space	standardized, cosine
Neighbor weighting	uniform, distance kernel
Temperature	1.0, 2.0
Blend weight β	41 equally spaced values on $[0, 1]$, by loss on D_{fit}

E Controlled Separation of Error, Ambiguity and Gap

On real models, the claim that correction gain tracks the Bayes gap rather than raw error is correlational because gap, ambiguity, and error cannot be manipulated independently. We therefore construct a synthetic family in which they can be controlled: $X \sim \mathcal{N}(0, I_{10})$, $\pi(X) = \mathrm{softmax}(WX/\tau)$, $Y \sim \mathrm{Cat}(\pi(X))$, and model prediction $m(X) = \mathrm{softmax}(WX/\tau + \gamma VX)$. The temperature τ controls ambiguity ($\mathrm{tr}\,\mathrm{Cov}(e_Y \mid X)$), while γ controls systematic conditional bias. We tune three configurations to have matched base accuracy (≈ 0.63–0.66) but different sources of error. The full GUARD pipeline (a source-kNN corrector and $\widehat{\beta}$ fitted on held-out data) then runs identically in each configuration over three seeds.

Table 23: **Three matched-accuracy configurations.** A separate run of the family behind Figure 3, with configurations of matched base accuracy whose error comes from the Bayes gap in different proportions. $\mathrm{PH}_{\text{true}} = \mathbb{E}\|d(X)\|_2^2$ is potential headroom, the mean squared Bayes gap; *Loss gain* is $L_{\text{base}} - L_{\text{corrected}}$. Mean over three seeds; the family is specified in App. E.

Configuration	Base acc	$\mathrm{PH}_{\text{true}}$	$\widehat{\beta}$	Loss gain	Δacc
Bias-dominant	0.64	0.39	0.75	$\mathbf{+1.02}$	$+0.11$
Mixed	0.65	0.23	0.62	$+0.37$	$+0.05$
Ambiguity-dominant	0.65	0.00	0.03	-0.00	$+0.00$

Figure 3 sweeps this family densely: for each τ on a grid, γ is bisected until the base accuracy returns to the target, so accuracy is matched by construction at every point rather than at three tuned ones. Over twelve points and ten seeds the base accuracy is 0.650 with a standard deviation of 0.010, while potential headroom runs from 0.369 to 0.0005. Across the twelve points, accuracy gain and $\mathrm{PH}_{\text{true}}$ correlate at 0.990, with a 95% interval of $[0.980, 0.993]$ from resampling seeds; over the 120 individual runs the correlation is 0.962. Two predictions hold along the sweep, and Table 23 adds three configurations from a separate three-seed run. First, at fixed raw error, gain increases with true headroom and vanishes for the Bayes-optimal model even though its accuracy is no better. Second, the fitted $\widehat{\beta}$ falls from 0.75 to 0.03 as the gap disappears, although the fitting data do not explicitly mark the absence of headroom.

Table 23 makes the control explicit at matched base accuracy: the ambiguity-dominated configuration gains nothing on either scale, while the bias-dominated one gains $+0.11$ in accuracy and $+1.02$ in loss. This is the opposite of what the heuristic "correction helps weak models" would predict.

F Proofs for the Correction Reduction

F.1 The correction reduction

Loss conventions. The gradient in equation 1 is with respect to scores z, whereas blending and harm evaluation use predictions p. We distinguish $\ell(y, z)$ from $\bar{\ell}(y, p)$ throughout. For categorical cross-entropy,

$$\ell(y, z) = \log \sum_c e^{z_c} - z_y, \qquad m(z) = \mathrm{softmax}(z), \qquad \bar{\ell}(y, p) = -\log p_y.$$

For squared error, $\ell(y, z) = \frac{1}{2}\|z - y\|^2$ and $m(z) = z$. The implemented Brier score is $\bar{\ell}(y, p) = \frac{1}{2}\|p - e_y\|^2$. It is convex in p; composing it with softmax does not give the canonical score gradient $p - e_y$. The unhalved Brier convention used for $\mathrm{PH}_{\mathrm{true}}$ and the excess-risk formulas below is twice this loss.

For multi-label outputs, summed binary cross-entropy has gradient $\mathrm{sigmoid}(z) - y$. The implementation averages over the K labels, giving gradient $(\mathrm{sigmoid}(z) - y)/K$. This positive scale leaves the minimizer unchanged but sets the units of δ: the reported multi-label tolerance is per label. The Bayes targets are marginal label probabilities; no conditional independence of the labels is required. Its excess risk is the mean of the coordinatewise Bernoulli KL divergences, rather than a categorical KL divergence between multi-label vectors.

The canonical identities and the convex-risk comparison use these ideal losses. For numerical evaluation, categorical probabilities inside logarithms are lower-clipped at 10^{-12}, and Bernoulli probabilities are clipped to $[10^{-12}, 1 - 10^{-12}]$. The harm event and fit loss use the same implemented convention. CRC still applies to the resulting bounded indicator; the ideal gradient and convexity arguments are not claims about clipped losses at saturation.

Proof of Theorem 1. On $\mathbb{R}^K$, equation 1 implies a representation

$$\ell(y, u) = A(u) - \langle t(y), u\rangle + c(y), \qquad \nabla A(u) = m(u).$$

Indeed, for any fixed label y_0, take $A(u) = \ell(y_0, u) + \langle t(y_0), u\rangle$; the difference above has zero gradient and is constant in u. Integrability of $t(Y)$ and of the loss at one finite score ensures integrability of $c(Y)$. Thus the conditional risk has the explicit form

$$R_X(u) = A(u) - \langle \pi(X), u\rangle + \mathbb{E}[c(Y) \mid X], \qquad \nabla R_X(u) = m(u) - \pi(X).$$

This proves equation 3 without interchanging a derivative and an expectation. Every finite minimizer in the unconstrained score space has zero gradient. Convexity makes this condition sufficient. If $m(z(X)) = \pi(X)$ almost surely, $R_X(z'(X)) \geq R_X(z(X))$ pointwise for every measurable finite edit, and integration proves the risk claim whenever the risks are defined. Softmax reaches only the open simplex: an interior categorical target is attained by $z_c^\star = \log \pi_c + a$, with any scalar a; a boundary target is approached by a sequence of finite logits. Squared error attains its target at $z^\star = \pi(X)$. □

The geometry of headroom. For categorical log loss and half-squared loss respectively, the excess conditional risks are

$$R_X(p) - R_X(\pi) = \mathrm{KL}(\pi \| p), \qquad R_X(p) - R_X(\pi) = \tfrac{1}{2}\|p - \pi\|^2, \tag{13}$$

where R_X here denotes prediction risk under $\bar{\ell}$. The first identity follows by subtracting the entropy of π; the second follows by expanding the square and using $\mathbb{E}[t(Y) - \pi \mid X] = 0$. These are the classical Bregman geometries of a proper loss (Banerjee et al., 2005; Gneiting & Raftery, 2007), two different geometries of the same target discrepancy. What Theorem 1 adds to them is a statement about deployment: under a frozen predictor, the part of the divergence that any measurable edit can remove is fixed by the information the corrector observes, which is why degradation severity and correction headroom separate. A decrease in cross-entropy need not decrease the squared Bayes gap, and neither identity orders accuracy gains of fitted correctors. In Figure 3, $\mathrm{PH}_{\mathrm{true}} = \mathbb{E}\|d\|^2$ is unhalved Brier headroom; its association with accuracy gain is an empirical result.

Feature-space reduction. Consider a residual feature correction $h \mapsto h + P_S a(X)$ with $a(X) \in \mathbb{R}^r$, an X-measurable injection $P_S \in \mathbb{R}^{d \times r}$, and head Jacobian $J = \partial z / \partial h$. For a twice continuously differentiable convex canonical loss, form the damped Gauss–Newton surrogate

$$\begin{aligned} q_i(a) &= \tfrac{1}{2} a^\top B(X_i) a + g_i^\top a, \qquad g_i = P_S^\top J_i^\top \big(m(z_i) - t(y_i)\big), \\ B(X) &= P_S^\top J^\top \nabla^2 A(z) J P_S + \kappa R(X), \qquad \kappa > 0, \quad R(X) \succ 0. \end{aligned} \tag{14}$$

Here $R(X)$ is symmetric and X-measurable, so $B(X)$ is symmetric positive definite and independent of Y given X. Label independence of the score Hessian follows from the canonical representation, not just from the CE and squared-error examples. The surrogate omits second derivatives of a nonlinear head and adds damping; it is not the exact nonlinear prediction risk.

Proposition 3 (Feature-space reduction)**.** *Under the conditions of equation 14, assume the conditional moments below exist and $\mathbb{E}[g_i^\top B(X_i)^{-1} g_i] < \infty$. Then*

$$\bar{g}(X) := \mathbb{E}[g_i \mid X] = P_S^\top J^\top d(X), \tag{15}$$

and the unique minimizer, up to null sets, of the expected surrogate over measurable corrections of finite B-energy is

$$a_B(X) = -B(X)^{-1} P_S^\top J^\top \big(m(z) - \pi(X)\big). \tag{16}$$

Strict surrogate headroom requires $P_S^\top J^\top d(X) \neq 0$ on a set of positive probability. The identity in equation 15 is exact; the approximation is the surrogate in equation 14.

Lemma 4 (Completion of squares)**.** *If B is symmetric positive definite and $q(a) = \frac{1}{2} a^\top B a + g^\top a$, its unique minimizer is $a^\star = -B^{-1} g$, and*

$$q(a) = \tfrac{1}{2}\|a - a^\star\|_B^2 - \tfrac{1}{2} g^\top B^{-1} g, \qquad \|v\|_B^2 = v^\top B v. \tag{17}$$

Proof. Expanding $\frac{1}{2}(a + B^{-1}g)^\top B (a + B^{-1}g)$ gives $q(a) + \frac{1}{2} g^\top B^{-1} g$. Positive definiteness gives uniqueness. □

Proof of Proposition 3. Conditioning on X fixes P_S, J, z, B and replaces $t(Y)$ by $\pi(X)$, proving equation 15. For any admissible measurable a,

$$\mathbb{E}[q_i(a) \mid X] = \tfrac{1}{2} a^\top B(X) a + \bar{g}(X)^\top a.$$

Completion of squares therefore gives

$$\mathbb{E}[q_i(a) - q_i(a_B) \mid X] = \tfrac{1}{2}\|a - a_B\|_{B(X)}^2.$$

The moment assumption and conditional Jensen ensure that a_B has finite B-energy; Cauchy–Schwarz makes the linear terms integrable for every correction in this class. Integrating the nonnegative difference proves optimality and uniqueness up to null sets. □

F.2 Output-space dominance

Proposition 5 (Output-space dominance)**.** *Let $h(X)$ and the head f_ψ be measurable. Let $\mathcal{A}$ contain measurable residual feature corrections $h \mapsto h + P_S a(X)$, and let $\mathcal{D}$ contain all measurable score edits $z \mapsto z + \omega(X)$. Every $a \in \mathcal{A}$ induces*

$$\omega_a(X) = f_\psi(h(X) + P_S a(X)) - f_\psi(h(X)).$$

Thus $\inf_{\omega \in \mathcal{D}} R(\omega) \leq \inf_{a \in \mathcal{A}} R(a)$ for a common prediction risk whenever these risks are defined. For an affine head $f_\psi(h) = Gh + c$, $\omega_a = G P_S a$. Under Theorem 1, any finite Bayes-optimal edit satisfies $m(z + \omega^\star) = \pi(X)$, and any edit attaining this target is optimal.

Proof. The constructed edit is measurable and gives exactly the same scores as the feature correction. Inclusion of predictor classes proves the risk inequality; this comparison does not include different parameter penalties or computational constraints. For a canonical convex loss, the gradient of the conditional edit risk is $m(z + \omega) - \pi(X)$, so the optimum follows from Theorem 1. For squared error, $\omega^\star = \pi - z$; for categorical CE, boundary targets may only be approached in the limit.
If f_ψ is differentiable at h, its local expansion is $f_\psi(h + P_S a) = f_\psi(h) + J P_S a + o(\|a\|)$ for fixed P_S. The remainder is $O(\|a\|^2)$ if the Jacobian is locally Lipschitz, for example under a locally bounded continuous second derivative. Exact measurable containment above does not require either local expansion. □

F.3 Headroom and ambiguity

Under the moment assumptions of Proposition 3, completion of squares yields

$$\mathbb{E}[q_i(a(X)) \mid X] = \tfrac{1}{2}\|a(X) - a_B(X)\|_{B(X)}^2 + \mathbb{E}[q_i(a_B(X)) \mid X]. \tag{18}$$

If $m(z) = \pi(X)$ almost surely, then $\bar{g} = 0$ and $a_B = 0$: every nonzero correction on a positive-probability set strictly increases the expected damped surrogate. This strictness concerns q, not necessarily the undamped nonlinear prediction risk. Conversely, optimal surrogate improvement is

$$\mathbb{E}[q_i(0)] - \mathbb{E}[q_i(a_B)] = \tfrac{1}{2}\mathbb{E}[\bar{g}^\top B^{-1}\bar{g}], \tag{19}$$

strictly positive exactly when $\bar{g} \neq 0$ on a set of positive probability. It does not guarantee that taking the full step a_B decreases the actual nonlinear loss.
Write $g = \bar{g} + (g - \bar{g})$. Since $B^{-1}\bar{g}$ is X-measurable and $\mathbb{E}[g - \bar{g} \mid X] = 0$, the cross term vanishes, giving the oracle-energy decomposition

$$\mathbb{E}[g^\top B^{-1} g \mid X] = \bar{g}^\top B^{-1}\bar{g} + \mathbb{E}[(g - \bar{g})^\top B^{-1}(g - \bar{g}) \mid X]. \tag{20}$$

For square-integrable $t(Y)$, let X^{full} include the full observation and the masking context, so that $\sigma(X) \subseteq \sigma(X^{\text{full}})$. The law of total covariance gives

$$\begin{aligned}\operatorname{Cov}(t(Y) \mid X) = \mathbb{E}[\operatorname{Cov}(t(Y) \mid X^{\text{full}}) \mid X] \\ + \operatorname{Cov}\big(\mathbb{E}[t(Y) \mid X^{\text{full}}] \mid X\big).\end{aligned} \tag{21}$$

The first term is ambiguity remaining with that full information; the second comes from full states that the poorer observation cannot distinguish. For random informative masks, including the mask in X^{full} is necessary for the asserted nesting.

F.4 Cross-mask inheritance and transfer

Throughout, $\mathcal{G}_S = \sigma(X_S) \subseteq \mathcal{G}_T = \sigma(X_T)$ under one joint law, and $t(Y)$ is integrable. Fixed nested raw modality sets satisfy this condition. For random masks, the richer observation must also contain the poorer mask context, or the claim must be applied conditionally on a fixed context. Set inclusion $S \subseteq T$ alone does not establish sigma-field inclusion when the mask carries information about Y. Separately computed frozen representations also need not be nested.
Write $\pi_S = \mathbb{E}[t(Y) \mid X_S]$, $\pi_T = \mathbb{E}[t(Y) \mid X_T]$, $d_S = m_S - \pi_S$, $d_T = m_T - \pi_T$, and $b = \mathbb{E}[d_T \mid X_S]$. The results below give an identity, contraction bounds, a population recovery condition and a comparison of retrieval targets.

Lemma 6 (Inheritance identity)**.** *Under the nesting above, $\pi_S = \mathbb{E}[\pi_T \mid X_S]$ and*

$$d_S = \big(m_S - \mathbb{E}[m_T \mid X_S]\big) + \mathbb{E}[d_T \mid X_S].$$

Proof. The tower property gives $\mathbb{E}[\mathbb{E}[t(Y) \mid X_T] \mid X_S] = \pi_S$. Hence $\mathbb{E}[d_T \mid X_S] = \mathbb{E}[m_T \mid X_S] - \pi_S$; rearrangement proves equation 4. □

Lemma 7 (The inherited term contracts)**.** *If $d_T \in L^2$, then $\mathbb{E}\|b\|^2 \leq \mathbb{E}\|d_T\|^2$. For categorical probability vectors and $\tau_S = \mathbb{E}[m_T \mid X_S]$,*

$$\operatorname{KL}(\pi_S \| \tau_S) \leq \mathbb{E}[\operatorname{KL}(\pi_T \| m_T) \mid X_S] \tag{22}$$

almost surely, and the same inequality holds after taking expectations on both sides. KL is interpreted in the extended nonnegative sense at boundary probabilities.

Proof. Conditional Jensen gives $\|\mathbb{E}[d_T \mid X_S]\|^2 \leq \mathbb{E}[\|d_T\|^2 \mid X_S]$; integration proves the first claim. Joint convexity of KL applied to the pair (π_T, m_T), whose conditional mean is (π_S, τ_S), proves equation 22. The same argument applies coordinatewise to Bernoulli KL for multi-label outputs, followed by the required sum or mean. These bounds measure conditional prediction error, which is stronger than calibration aggregated over equal predicted probabilities. □

Proposition 8 (Recovery condition, and non-identifiability)**.** *The population transfer target $\tau_S = \mathbb{E}[m_T \mid X_S]$ equals π_S almost surely if and only if $b = 0$ almost surely. For a fixed frozen predictor and unlabeled paired law, π_S is not identifiable over unrestricted conditional label laws with at least two distinct attainable label statistics.*

Proof. Inheritance gives $\tau_S = \pi_S + b$, proving the equivalence. For non-identifiability, keep the paired feature law and predictor fixed. Choosing $Y = y_0$ almost surely in one law and $Y = y_1$ almost surely in another, with $t(y_0) \neq t(y_1)$, leaves every unlabeled observation unchanged but changes π_S. A condition linking predictions to labels is therefore required. This is a population statement: a finite-neighbor average need not equal τ_S even when $b = 0$. □

Proposition 9 (Transfer guarantees). *Let $p_\beta(q) = (1-\beta)m_S + \beta q$, with $\beta \in [0,1]$. In both parts, the query label has its deployment conditional law independently of the retrieval sample given X_S.*

(i) Finite-sample dominance. *For a source pool of size n, let $\mathcal{F}$ fix the query, base prediction, selected richer-view neighbor features and nonnegative weights summing to one. Suppose $\mathbb{E}[t(Y_j) \mid \mathcal{F}] = \pi_T(X_T^{(j)})$ for every selected neighbor, and $m_T = \pi_T$. For the same fixed β, any admissible $1 \le k \le n$, and convex conditional prediction risk R_{X_S} (including ideal CE and Brier),*

$$\mathbb{E}R_{X_S}\big(p_\beta(\widehat{\pi}^{(T)})\big) \le \mathbb{E}R_{X_S}\big(p_\beta(\widehat{\pi}^{(Y)})\big).$$

The expectation is over queries and retrieval samples.

(ii) Oracle blend and estimation error. *For squared risk, assume $t(Y)$, d_S and $\varepsilon_n = \widehat{\pi}_n - \pi_S$ are square-integrable. Set*

$$D = \mathbb{E}\|d_S\|^2, \qquad V_n = \mathbb{E}\|\varepsilon_n\|^2, \qquad C_n = \mathbb{E}\langle d_S, \varepsilon_n\rangle.$$

The unhalved excess risk of a scalar blend is

$$F_n(\beta) = (1-\beta)^2 D + 2\beta(1-\beta)C_n + \beta^2 V_n. \tag{23}$$

For $A_n = D + V_n - 2C_n > 0$, its minimizer is

$$\beta_n^\star = \left[\frac{D - C_n}{D + V_n - 2C_n}\right]_{[0,1]}. \tag{24}$$

If $A_n = 0$, every $\beta \in [0,1]$ attains $F_n(\beta) = D$. If the unconstrained minimizer is in $[0,1]$, the minimum is $(DV_n - C_n^2)/A_n$; otherwise it is the appropriate endpoint value D or V_n.

Let n_U and n_L be the unlabeled and labeled pool sizes. If $\widehat{\pi}_{n_U}^{(T)} \to \tau_S = \pi_S + b$ in L^2 as $n_U \to \infty$, then

$$\min_{\beta\in[0,1]} F_{n_U}^{(T)}(\beta) \longrightarrow \min_{\beta\in[0,1]} \mathbb{E}\|(1-\beta)d_S + \beta b\|^2. \tag{25}$$

For hard-label retrieval, $\min_\beta F_{n_L}^{(Y)} = DV_{n_L}/(D + V_{n_L})$ when $C_{n_L} = 0$ and $D + V_{n_L} > 0$. More generally, if $D > 0$, $V_{n_L} > 0$ tends to zero and $C_{n_L}^2 = o(V_{n_L})$, then $\min_\beta F_{n_L}^{(Y)} = V_{n_L}(1 + o(1))$. If $D = 0$, choosing $\beta = 0$ already gives zero excess risk.

Proof. Part (i). The conditional-mean assumption and the $\mathcal{F}$-measurable weights give

$$\mathbb{E}[\widehat{\pi}^{(Y)} \mid \mathcal{F}] = \sum_{j\in\mathcal{N}_k} w_j \pi_T(X_T^{(j)}) = \widehat{\pi}^{(T)}.$$

The map $q \mapsto R_{X_S}((1-\beta)m_S + \beta q)$ is convex for $\beta \in [0,1]$. Conditional Jensen therefore gives

$$R_{X_S}\big(p_\beta(\widehat{\pi}^{(T)})\big) \le \mathbb{E}\left[R_{X_S}\big(p_\beta(\widehat{\pi}^{(Y)})\big) \mid \mathcal{F}\right].$$

Taking expectations proves the claim. Nonnegative CE may take the value $+\infty$ when a hard-label target has zeros, in which case Jensen is understood in the extended sense. Independence of neighbor labels is not needed for this step; preservation of their conditional means is. For i.i.d. source samples, a query independent of the pool, and label-independent fixed feature maps and neighbor selection, that condition holds. It also holds for $\mathcal{F}$-measurable blends if conditioning on their construction preserves the stated conditional means.

Choosing different blends, temperatures or neighbors by inspecting retrieval labels is not covered by this comparison. Even a separate fit split can select settings using predictions that depend on those labels, so disjointness alone does not justify conditioning on the selected settings. Tempering $m_T = \pi_T$ generally changes the conditional target and removes the hypothesis. The proposition compares the two fixed candidates in convex prediction risk, before fitting separate gates, and makes no claim about their accuracy ordering.

Part (ii). The residual is $p_\beta(\widehat{\pi}_n) - \pi_S = (1-\beta)d_S + \beta\varepsilon_n$. Expanding its squared norm gives equation 23. Equivalently,

$$F_n(\beta) = D - 2\beta(D - C_n) + \beta^2 A_n, \qquad A_n = \mathbb{E}\|d_S - \varepsilon_n\|^2 \ge 0.$$

For $A_n > 0$, completion of squares and projection onto $[0,1]$ prove equation 24. Substitution gives the interior minimum. If $A_n = 0$, $d_S = \varepsilon_n$ almost surely, so the residual and risk are constant in β.

The formula $DV_n/(D+V_n)$ is the special case $C_n = 0$. Conditional unbiasedness $\mathbb{E}[\varepsilon_n \mid X_S] = 0$ suffices for $C_n = 0$; unconditional mean zero alone does not.

For the transfer limit, put $e_n = \widehat{\pi}_{nU}^{(T)} - \tau_S$ and $F_\infty(\beta) = \mathbb{E}\|(1-\beta)d_S + \beta b\|^2$. Cauchy–Schwarz gives

$$\sup_{\beta\in[0,1]} |F_{nU}^{(T)}(\beta) - F_\infty(\beta)| \le 2(\|d_S\|_{L^2} + \|b\|_{L^2})\|e_n\|_{L^2} + \|e_n\|_{L^2}^2 \longrightarrow 0.$$

Uniform convergence implies convergence of minima, proving equation 25. For the hard-label asymptotic claim, $A_n \to D > 0$ and the unconstrained minimizer is eventually positive. If it is at most one, the minimum divided by V_n is $(D - C_n^2/V_n)/A_n \to 1$; if it exceeds one, the constrained minimum is exactly V_n. This covers both cases. □

Neighborhood bias and variance. The moment formula includes smoothing bias in both retrieval modes. To make this explicit, condition on the query and the selected poorer-view features, with weights determined by those features. Assume the selected source responses are conditionally independent given them. Write $\Sigma_Y(x) = \mathrm{Cov}(t(Y) \mid X_S = x)$. Then

$$a_n^{(Y)} = \sum_j w_j \pi_S(X_S^{(j)}) - \pi_S(X_S), \tag{26}$$

$$\mathrm{Cov}(\widehat{\pi}^{(Y)} \mid X_S, \{X_S^{(j)}\}) = \sum_j w_j^2 \Sigma_Y(X_S^{(j)}), \tag{27}$$

$$V_n^{(Y)} = \mathbb{E}\|a_n^{(Y)}\|^2 + \mathbb{E}\sum_j w_j^2 \,\mathrm{tr}\,\Sigma_Y(X_S^{(j)}), \qquad C_n^{(Y)} = \mathbb{E}\langle d_S, a_n^{(Y)}\rangle. \tag{28}$$

For transfer, replace the conditional mean π_S by $\tau_S = \pi_S + b$ and Σ_Y by $\Sigma_T(x) = \mathrm{Cov}(m_T(X_T) \mid X_S = x)$. Its conditional bias relative to π_S is $b(X_S) + \sum_j w_j \tau_S(X_S^{(j)}) - \tau_S(X_S)$, not just $b(X_S)$ at a finite bandwidth.

When $m_T = \pi_T$, total covariance gives

$$\Sigma_Y(X_S) = \mathbb{E}[\mathrm{Cov}(t(Y) \mid X_T) \mid X_S] + \mathrm{Cov}(\pi_T \mid X_S),$$

so $\Sigma_T(X_S) \preceq \Sigma_Y(X_S)$. This is a matrix inequality; their traces give the scalar noise contributions to squared risk. Under the fixed weighted comparison above, the two modes share the same smoothing bias and the transfer mode has no greater conditional covariance. For uniform weights in a neighborhood of radius h around x, continuity of the covariance gives $\Sigma(x)/k + o(1/k)$ as $h \to 0$, rather than an exact finite-pool formula. A Lipschitz conditional mean gives smoothing bias $O(h)$. In an integrated regime with bounded constants and nonvanishing integrated noise of order $1/k$, $kh^2 \to 0$ makes its squared contribution $o(1/k)$ and implies $C_n^2 = o(V_n)$. These are additional regularity and sampling conditions; they do not follow from the existence of a finite retrieval pool. Weighted neighbors instead contribute through $\sum_j w_j^2$.

More generally, if an estimator has the exact decomposition $\varepsilon = b_n(X_S) + \xi$ with $\mathbb{E}[\xi \mid X_S] = 0$, its blend denominator is

$$D + V - 2C = \mathbb{E}\|d_S - b_n\|^2 + \mathbb{E}\|\xi\|^2, \qquad C = \mathbb{E}\langle d_S, b_n\rangle.$$

The integrated noise $\mathbb{E}\|\xi\|^2$ must not be replaced by an unaveraged, query-dependent covariance. In the special unbiased case $b_n = 0$, lower integrated noise lowers the oracle risk $DV/(D+V)$ for $D > 0$.

Remark 10 (Two limits, and what they do not say). Under the L^2 consistency assumption in Proposition 9, the transfer target's MSE tends to $\mathbb{E}\|b\|^2$, while the oracle blend's excess risk tends to

$$\min_{\beta\in[0,1]} \mathbb{E}\|(1-\beta)d_S + \beta b\|^2 \le \min\{D, \mathbb{E}\|b\|^2\}.$$

The target limit follows from L^2 convergence, and the inequality follows by choosing either endpoint. Consistency requires estimation error to vanish; increasing the pool with a fixed k need not suffice. These statements concern an oracle scalar blend and retrieval labels only. GUARD still uses labels for fitting and calibration, selects $\widehat{\beta}$ on a finite grid, and applies a learned gate. No consistency or risk ordering for that fitted policy follows from the oracle formula.

F.5 FINITE-SAMPLE HARM CONTROL

Conditioning and the loss family. Let $\mathcal{H}$ contain all fitting data, fixed randomization and choices made before calibration: the retrieval pool and rule, reference p_{base}, candidate p_{corr}, feature transforms, temperature, blend, scorer, prediction-loss convention and tolerance δ. Conditional on $\mathcal{H}$, assume the n calibration samples and the next deployment sample are exchangeable. Disjoint data roles alone are not a substitute for this assumption. The fitted objects define a common measurable function for each sample; any query-dependent retrieval weighting must follow that same rule at calibration and deployment.

Take finite scores and thresholds in $\overline{\mathbb{R}}$, with $A_{-\infty} = 1$ and $A_{+\infty} = 0$. The functions $L_i(\lambda)$ in equation 9 are exchangeable conditional on $\mathcal{H}$, bounded by one, non-increasing and right-continuous, because the gate uses $r_\theta > \lambda$. At $+\infty$ the loss is zero. These are the conditions of conformal risk control (Angelopoulos et al., 2024). If $n = 0$ or the corrected empirical criterion is infeasible, set $\widehat{\lambda} = +\infty$.

Proof of Proposition 2. For $n \geq 1$, define the symmetric auxiliary threshold

$$\lambda^\circ = \inf\left\{\lambda : \frac{1}{n+1}\sum_{i=1}^{n+1} L_i(\lambda) \leq \alpha\right\}.$$

The finite step functions and the no-intervention endpoint ensure feasibility at the infimum. Since $L_{n+1} \leq 1$, every feasible CRC threshold is feasible here, so $\widehat{\lambda} \geq \lambda^\circ$, also under the empty-set convention. Monotonicity and conditional exchangeability give

$$\mathbb{E}[L_{n+1}(\widehat{\lambda}) \mid \mathcal{H}] \leq \mathbb{E}[L_{n+1}(\lambda^\circ) \mid \mathcal{H}] = \mathbb{E}\left[\frac{1}{n+1}\sum_i L_i(\lambda^\circ)\,\middle|\,\mathcal{H}\right] \leq \alpha.$$

For $n = 0$ the loss is zero by construction. Integrating over $\mathcal{H}$ proves equation 11. □

The probability is marginal over calibration and deployment, and is relative to the declared p_{base}, loss and tolerance. It bounds joint harmful intervention, not harm conditional on applying, mean loss increase, or a benchmark metric such as AUROC. An unbounded ideal CE loss poses no problem because the calibrated loss is its bounded harm indicator. At $\alpha < 1/(n+1)$, no threshold satisfies the corrected criterion; the implementation also chooses no intervention at equality, a conservative convention.

Tightening and what must remain fixed. For any measurable post-calibration choice $\lambda^\star \geq \widehat{\lambda}$, $L_{\lambda^\star} \leq L_{\widehat{\lambda}}$ for every calibration draw and deployment point. Thus tightening preserves the bound even if the chosen threshold depends on calibration. It must retain the same score, candidate, reference, loss and tolerance; changing any of these changes the loss family and is not justified by threshold monotonicity. GUARD uses D_{fit} for its utility-based tightening.

Mask-conditional calibration. Use a partition by masks fixed before calibration. For a group s, apply equation 10 to its n_s calibration observations and use no intervention when $n_s = 0$.

Corollary 11 (Mask-conditional harm control)**.** *Suppose, conditional on $\mathcal{H}$, $S_{\text{test}} = s$ and the group count n_s, the group calibration observations and the deployment observation are exchangeable under the group law. Then the group threshold satisfies*

$$\mathsf{Pr}(\Delta_Y(X) > \delta \ \wedge\ \text{apply} \mid S_{\text{test}} = s) \leq \alpha.$$

Proof. For each nonempty group and each count, apply the preceding proof under the stated conditional law; empty or infeasible groups have zero intervention loss. Average over the counts and fitting information conditional on $S_{\text{test}} = s$. □

Cost of a poor scorer. A poor score can reduce utility without invalidating this bound. A constant score offers only blanket intervention or no intervention. Blanket intervention is admitted only if its calibration harm satisfies $(n\widehat{R}_n + 1)/(n+1) \leq \alpha$, not merely $\widehat{R}_n \leq \alpha$. The agreement-only examples on NinaPro and AVE illustrate these two policies. Neither policy certifies the outcome of an individual query.

Learn-then-Test comparator. Learn-then-Test (Angelopoulos et al., 2025) instead provides a high-probability risk statement over the calibration draw. Its Hoeffding–Bentkus tests require i.i.d. calibration losses (or a separately justified valid concentration bound), a stronger condition than arbitrary exchangeability. Our comparator fixes its candidate threshold grid and order using D_{fit}, tests $H_\lambda : R(\lambda) > \alpha$ in sequence at level $\zeta = 0.05$, and stops at the first non-rejection. The first true null, if reached, is rejected with probability at most ζ, which controls erroneous certification along the sequence. The risk being tested is the same joint harm. As specified in App. D, PTB-XL additionally differs in the logistic scorer's entropy feature, so its reported LTT row is not an isolated calibration ablation.

F.6 Selecting the blend weight

The action score r_θ is a logistic model fitted on D_{fit} against $Z = \mathbf{1}\{\Delta_Y < 0\}$. Its five features are the largest reference probability, reference entropy, largest target probability, the agreement flag $\mathbf{1}\{K^{-1}\sum_c |p_{\mathrm{base},c} - \widehat{\pi}_c| < 0.1\}$, and the ℓ_1 discrepancy $\sum_c |p_{\mathrm{base},c} - \widehat{\pi}_c|$. Probabilities used in the reference features are clipped for numerical stability. Categorical entropy is $-\sum_c p_c \log p_c$; for multi-label predictions GUARD uses $-\sum_c [p_c \log p_c + (1-p_c)\log(1-p_c)]$, the sum of marginal Bernoulli entropies.

For two categorical simplex rows, $\sum_c |p_c - q_c| \le 2$, so the agreement flag is identically one when $C > 20$, as on AVE and NinaPro DB5; it carries no signal there. The coordinate clipping changes this distance by at most $C\,10^{-12}$ and does not alter that conclusion for these benchmarks. This simplex argument does not apply to multi-label vectors. The feature is retained as implemented; a different scaling would require evaluating a different scorer.

The blend is selected from 41 equally spaced values in $[0, 1]$ by minimizing mean prediction loss $\bar{\ell}$ on D_{fit}, before calibration. Retrieval settings are selected using the benchmark metric, except for the accuracy-based comparison in Table 3. This fit criterion is separate from the ideal oracle analysis of Proposition 9; no generalization or consistency bound for the grid selector is claimed.

Utility-based tightening. Let $g_i = -\Delta_{Y_i}(X_i)$ on D_{fit}. Among the calibrated threshold and distinct fit scores above it, evaluate

$$G(\lambda) = \sum_{i : r_\theta(X_i) > \lambda} g_i, \qquad s(\lambda) = \sqrt{N_\lambda}\,\widehat{\mathrm{sd}}\{g_i : r_\theta(X_i) > \lambda\},$$

where N_λ is the number admitted and $s = 0$ when fewer than two predictions are admitted. If the best gain is positive, retain the largest candidate threshold whose gain is within $s(\lambda_{\mathrm{best}})$ of that best gain. Otherwise move above all fit scores, without dropping below the CRC threshold. This is the implemented one-standard-error heuristic on accepted gains; it is not a confidence interval for total deployment utility. Every selected threshold is at least $\widehat{\lambda}_{\mathrm{CRC}}$, which is the property needed for harm control.

For fixed data and score, the CRC threshold alone is non-increasing in α. The subsequent utility search need not preserve that ordering: it changes both the best fit gain and the set of near-best thresholds as the admissible set expands. Figure 7 therefore describes an empirical intervention curve, not a monotonicity theorem for the tightened policy.

All fitted choices optimize a risk on the fit distribution Q. Matching deployment Q' requires more than reducing sampling error on Q; additional fit data do not remove a distribution mismatch. Our fitting splits are drawn from the evaluated deployment condition. The harm guarantee separately requires the conditional exchangeability stated in App. F.5.

F.7 The cross-mask accuracy screen

The population accuracy ordering gives a necessary condition for a conditionally correct richer predictor. For categorical labels and any poorer-view classifier $a_S(X_S)$,

$$\mathbb{E}\max_c \pi_{T,c} \;\ge\; \mathbb{E}\max_c \pi_{S,c} \;\ge\; \mathsf{Pr}(a_S(X_S) = Y). \tag{29}$$

The first inequality follows from conditional Jensen for the maximum and the tower identity; the second is the Bayes-optimality of $\arg\max_c \pi_{S,c}$. If $m_T = \pi_T$, its argmax attains the left side. Thus a strictly lower *population* accuracy rules out conditional correctness. Passing this ordering is not sufficient: accurate class decisions need not estimate conditional probabilities correctly.

Finite-sample accuracy reversals are descriptive, not statistical rejections without an appropriate uncertainty assessment. Among the twenty-eight dataset–condition–seed cells of Figure 8 with

hard-label gated gain at least 0.02, the ordering flags two, both retaining essentially none of that gain. Ten of the twenty-six passing cells also retain none. Here, richer-mask accuracy separates outcomes more clearly than its margin over the poorer predictor.

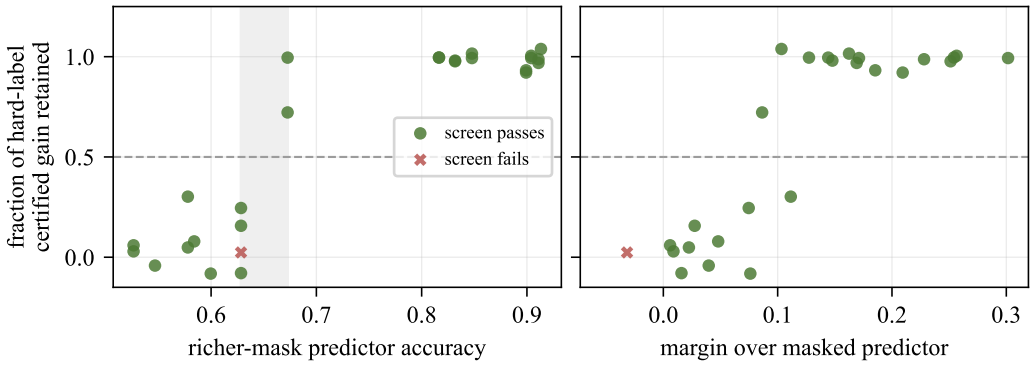

Figure 8: **Retained cross-mask gain against two screening statistics.** Twenty-eight dataset–condition–seed cells from binary DrugBAN and five-class OPPORTUNITY have hard-label gated gain at least 0.02. *Left:* richer-mask accuracy separates the observed modes, with no cell between 0.628 and 0.673. *Right:* the margin over the masked predictor overlaps between 0.086 and 0.111. The separation persists for headroom cuts 0.01–0.03 and fails at 0.005. This is a descriptive comparison, not a universal accuracy threshold or a test of conditional correctness.

Absolute accuracy depends on class cardinality and balance: uniform guessing scores $1/2$ on binary DrugBAN and $1/5$ on five-class OPPORTUNITY, while majority-class baselines depend on label frequencies. The observed separation does not transfer directly to 29-way AVE or many-class NinaPro.

Retained fractions here exceed 0.72 or fall below 0.31. Lemma 7 bounds inherited error by richer-view error in expected squared norm and KL. These population bounds do not imply continuity of accuracy or gated gain, which also depend on finite retrieval, tuning and discrete decisions.