WHAT FUSION ACCURACY HIDES ABOUT MISSING-MODALITY ROBUSTNESS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal models are usually evaluated with all modalities available, yet at deployment a modality is often missing. We show that fusion accuracy says little about this setting: models with nearly identical complete-input accuracy can differ sharply once one modality is removed, because complete-input training constrains only the joint output of the modality branches, not how each branch behaves alone. To locate such failures, we separate what the retained representation supports from what the model's own classifier extracts from it, using probes on frozen features and an independently trained unimodal model as reference. On audio–video, text–image, and clinical EHR–X-ray tasks, low missing-modality accuracy often hides a highly predictive representation; balanced training can even make the weaker modality's representation more predictive than unimodal training, while the classifier fails to use it. We propose Gated Residual Repair (GRR), a small affine correction applied only when a modality is missing, which recovers most of this gap without changing the encoders or complete-input predictions. Adapting the encoders helps only where the retained representation falls well below the unimodal reference, a case the diagnosis identifies in advance. Missing-modality robustness is thus a distinct property of multimodal models that fusion accuracy does not reveal.

Keywords: Multimodal learning; Missing-modality robustness; Balanced multimodal learning; Representation decodability; Classifier adaptation

1 Introduction

Multimodal learning improves performance on downstream tasks by combining complementary information from multiple modalities (Baltrušaitis et al., 2019). Joint training often faces modality imbalance, where one modality dominates optimization and suppresses the learning of the others (Wang et al., 2020; Wu et al., 2022). Balanced multimodal learning methods mitigate this imbalance through gradient modulation, unimodal supervision, classifier reallocation, or modality-specific optimization, and improve fusion performance (Peng et al., 2022; Kontras et al., 2024; Jiang et al., 2025; Chen et al., 2026). These gains are mainly measured with complete inputs, where every modality seen during training is also available at inference. However, in deployment, a modality is often missing because sensors fail, clinical measurements are unavailable, or acquisition is costly, and the model must predict from the remaining modality alone (Ma et al., 2022; Reza et al., 2025). Since balancing aims to improve how each modality is learned, one would expect a balanced model to remain useful in this setting. A question arises: does better fusion imply better use of each modality alone?

Figure 1 shows that it does not. On CREMA-D, MLB and MASAM reach nearly the same complete-input accuracy (71.15% and 70.92%), yet respond very differently when one modality is removed. With video alone, MASAM drops to 17.70%; with audio alone, MLB drops to 23.70%. Both fall far below independently trained unimodal models. A model can fuse two modalities well and still be worse than a unimodal model on the modality it retains. We observe this behavior across other methods and datasets as well.

Complete-input evaluation cannot detect this behavior, because it observes the modalities only through their joint prediction. Removing a modality exposes the failure but not its cause: the re-

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.

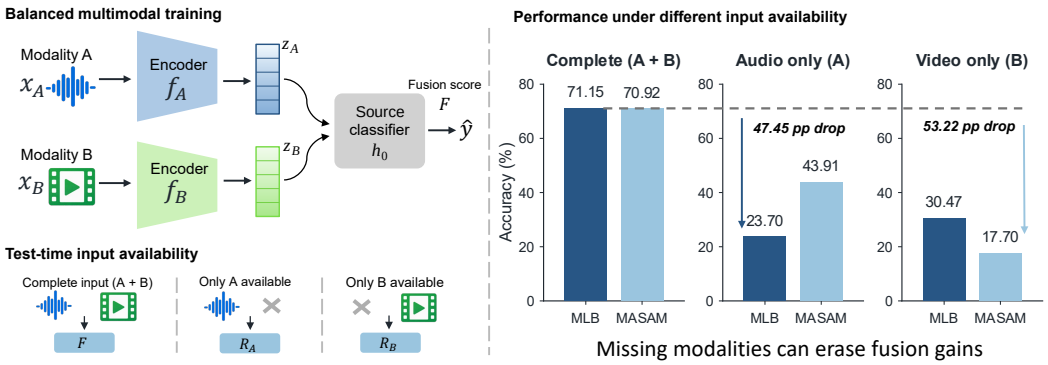

Figure 1: Similar fusion accuracy, very different single-modality accuracy. Left: a source model (encoders f_A, f_B, classifier h_0) trained on complete inputs is evaluated with both modalities (F), audio only (R_A), or video only (R_B). Right: on CREMA-D, MLB and MASAM have nearly the same F, but MLB fails with audio alone and MASAM with video alone. Dashed lines: complete-input accuracy; arrows: drop in points.

maining modality may carry too little task information, or the classifier may fail to use information that is still there. Both cases produce the same low accuracy, yet they reflect different effects of joint training: one changes what a modality encodes, the other changes how the encoded information is combined and used. A similar distinction appears elsewhere in representation learning. In work on class imbalance and spurious correlations, separating learned features from the classifier has changed how failures are understood and fixed (Kang et al., 2020; Kirichenko et al., 2023). Missing modalities pose a related problem, since the available inputs change at test time. For models trained with complete inputs, however, this separation has received little attention. The separation matters for interpreting what balancing methods improve when they raise fusion performance, for comparing multimodal models beyond fusion accuracy, and for deciding whether a model facing missing inputs needs a better representation or a new classifier.

To separate these two cases, we first analyze why complete inputs cannot tell them apart. We show that complete-input training leaves the standalone behavior of each modality branch undetermined: under complete inputs the branches can compensate for one another, and this compensation disappears once a modality is removed. We then decompose missing-modality failure into a representation component, which reflects what the retained representation supports, and a readout component, which reflects what the source classifier h_0 fails to use through its readout (formalized as the source readout t_m in Section 3). We test each component by changing only the readout or also adapting the representation. We evaluate multiple multimodal training methods on three datasets from different domains: audio–video emotion recognition, text–image classification, and clinical prediction from EHR and chest X-rays. The retained representation is often considerably more predictive than the model's own missing-input output, and retraining only the readout recovers much of the lost performance. Adapting the representation helps further in a smaller set of cases, mainly when the retained representation is weaker than that of an independently trained unimodal model. Missing-modality robustness thus depends on how representations and readouts interact under a given input availability, not on fusion accuracy alone.

Our main contributions are summarized as follows:

1. **Underdetermination under complete inputs.** We show that complete-input performance does not uniquely determine how individual modality branches behave in isolation. We formalize this ambiguity under affine fusion and show empirically that models with similar fusion performance can behave very differently when one modality is missing.
2. **Decomposing representation and readout.** We separate missing-modality failure into the predictive capability that remains in the retained representation and the capability the model realizes from it under the changed input state, and use this decomposition to characterize multimodal models beyond fusion accuracy.
3. **Recovery on fixed representations.** We show that many severe missing-modality failures can be recovered without modifying the learned representations, whereas representation adaptation brings additional gains mainly for particular models and missing modalities.

2 Related Work

Multimodal balancing. In supervised multimodal learning, modalities are often not learned or exploited equally. Prior work attributes this imbalance to differences in optimization and generalization across modality branches (Wang et al., 2020; Wu et al., 2022) and to competition between modalities during joint training (Huang et al., 2022). These analyses have motivated methods that modulate encoder gradients (Peng et al., 2022; Guo et al., 2024; Kontras et al., 2024), regulate information acquisition or unimodal adaptation (Fan et al., 2023; Zhang et al., 2024; Huang et al., 2025), diagnose each modality's learning state and re-initialize its encoder accordingly (Wei et al., 2024), alter classifier allocation (Jiang et al., 2025), or apply modality-specific sharpness-aware updates (Chen et al., 2026). These methods are developed and evaluated mainly with all modalities available, and improved balance is typically shown through higher fusion accuracy or through stronger unimodal encoders, for example under linear evaluation (Du et al., 2023). Neither shows whether the trained model can still predict from one modality when the other is missing. More generally, models with the same held-out performance can behave very differently under distribution shift, because the training pipeline and i.i.d. evaluation leave this behavior unconstrained (D'Amour et al., 2022). Complete-input underdetermination (Section 3.2) is a concrete case in multimodal fusion: the objective fixes the joint output of the branches but not their standalone behavior.

Learning with missing modalities. Missing modalities have been addressed through modality-subset training (Neverova et al., 2016), missing-feature reconstruction (Ma et al., 2021; Zhao et al., 2021), shared–specific representation learning (Wang et al., 2023), prompting for missing-modality cases (Lee et al., 2023), and parameter-efficient adaptation to particular missing-modality states (Reza et al., 2025). In clinical prediction, where some modalities are often unavailable for a patient, models are built to learn from partially paired records (Hayat et al., 2022; Wu et al., 2024). These methods anticipate missingness and train models to handle it. Closer to this work, Ma et al. (2022) show that multimodal transformers trained on complete inputs degrade sharply when a modality is missing at test time, and Hazarika et al. (2022) find that multimodal sentiment models are highly sensitive to missing or noisy modalities. The question here is different: why a model trained on complete inputs fails when a modality is missing, and whether the failure lies in the retained representation or in how the model uses it.

Representation and readout. Joint multimodal training can alter modality-specific representations relative to unimodal training (Du et al., 2023), and probing characterizes what information can be decoded from frozen representations (Alain & Bengio, 2017; Hewitt & Liang, 2019; Voita & Titov, 2020). What a representation makes decodable, however, need not coincide with what the model's own classifier extracts from it. In long-tailed recognition and under spurious correlations, separating the two has proven useful: retraining only the last layer on frozen features recovers much of the lost performance (Kang et al., 2020; Kirichenko et al., 2023), and features from standard training often already contain the information needed for robust prediction (Izmailov et al., 2022). Missing modalities pose a related but distinct problem. The failure arises from a change in which inputs are available rather than in the label or group distribution, and a repair must leave the model's behavior on complete inputs intact. GRR addresses this setting by correcting the readout only when a modality is missing.

3 Understanding Missing-Modality Failure

3.1 Setup

Consider a source model trained on modalities A and B, which consists of two encoders and a fusion classifier. The encoder f_m of modality $m \in \{A, B\}$ maps an input x_m to a representation $z_m = f_m(x_m) \in \mathbb{R}^{d_m}$, and the fusion classifier h_0, referred to as the source classifier, maps the concatenated representation $[z_A, z_B]$ to logits C. All additional classifiers are then trained on its frozen source representations.

At test time, either both modalities or only one of them may be available, yielding three input configurations: AB, A, and B. A missing modality is represented by zero-padding its feature vector prior to fusion:

$$z^{(AB)} = [z_A, z_B], \qquad z^{(A)} = [z_A, \mathbf{0}], \qquad z^{(B)} = [\mathbf{0}, z_B]. \tag{1}$$

For models that fuse branch logits rather than features, the logits of the missing branch are omitted instead.

When only modality m is available, the source model predicts through $t_m(z_m) = h_0(z^{(m)})$. Such a mapping from a single representation to logits is called a readout, and t_m is specifically termed the source readout of modality m. This mapping directly inherits the classifier parameters learned during multimodal training under the missing modality setting. Task performance is measured by a score $\mathcal{S}$, e.g., accuracy, where $F = \mathcal{S}(h_0(z^{(AB)}))$ and $R_m = \mathcal{S}(t_m(z_m))$ denote the scores with complete inputs and with modality m alone, respectively.

3.2 Complete-Input Underdetermination

The source classifier is trained and selected on complete inputs, so its training signal constrains the source readouts t_A and t_B only through their joint output. To see what this leaves undetermined, consider an affine fusion classifier

$$h(z^{(AB)}) = W_A z_A + W_B z_B + b, \tag{2}$$

with branch weight matrices W_A, W_B and bias b. Modifying the branch weights to $W_A + \Delta_A$ and $W_B + \Delta_B$ changes the complete-input logits by $\Delta_A z_A + \Delta_B z_B$, whereas the only-$m$ logits change by $\Delta_m z_m$ alone. Whenever $\Delta_A z_A + \Delta_B z_B = 0$ for almost all complete-input pairs, the modified classifier therefore has the same complete-input logits as the original one, while its only-A and only-B predictions can differ. Appendix B extends this to approximate cancellation, in which the change in complete-input risk is bounded by the residual of the cancellation.

For a concrete example, consider binary labels $Y \in \{-1, +1\}$, a single logit whose sign gives the prediction, and two representations that both encode the label perfectly, $z_A = z_B = Y$. The fusion classifiers with outputs $(z_A + z_B)/2$, $-z_A + 2z_B$, and $2z_A - z_B$ all output Y on complete inputs and are therefore perfectly accurate. With only A, their outputs become $Y/2$, $-Y$, and $2Y$; with only B, they become $Y/2$, $2Y$, and $-Y$. The first classifier thus remains perfectly accurate, the second fails whenever only A is present, and the third fails whenever only B is present, although all three share the same complete-input logits and the same perfectly informative representations.

Complete-input performance therefore does not determine R_m, and a low R_m need not indicate an uninformative representation.

3.3 Representation and Readout Failures

A low R_m has two possible causes. In a representation failure, the retained representation z_m carries too little label information, and only changing the encoder can help. In a readout failure, z_m remains predictive but the source readout t_m does not use it, and changing the classifier is enough. To separate the two, we train a new readout on top of the frozen z_m to see if it can predict accurately.

Let ℓ be the task loss and $\mathcal{H}_m$ a family of readouts on $\mathbb{R}^{d_m}$, taken as all affine maps in our experiments. For a readout g and a representation z, define

$$L(g; z) = \mathbb{E}\big[\ell(g(z), Y)\big], \qquad L_m^\star(z) = \inf_{g \in \mathcal{H}_m} L(g; z), \tag{3}$$

so that $L_m^\star(z)$ is the lowest loss any readout in $\mathcal{H}_m$ attains on z. Let z_m^{uni} be the representation of an independently trained unimodal encoder. Relative to this unimodal reference, the missing-input loss of the source model splits into two terms:

$$L(t_m; z_m) - L_m^\star(z_m^{\mathrm{uni}}) = \underbrace{L(t_m; z_m) - L_m^\star(z_m)}_{\Gamma_m^{\mathrm{read}}\ \text{(readout term)}} + \underbrace{L_m^\star(z_m) - L_m^\star(z_m^{\mathrm{uni}})}_{\Gamma_m^{\mathrm{repr}}\ \text{(representation term)}}. \tag{4}$$

The two terms vary the readout and the representation separately. Γ_m^{read} keeps the representation z_m and compares the source readout with the best readout on it; a large value means that z_m contains usable information that t_m leaves unused. It is nonnegative whenever $t_m \in \mathcal{H}_m$. Γ_m^{repr} keeps the readout family and compares z_m with z_m^{uni}; it is positive when joint training has made the retained representation less predictive than unimodal training, and negative when joint training has improved it.

Both terms are population quantities, and we estimate them with task scores. We fit an affine probe on the frozen z_m and denote its score by P_m. The readout term is estimated by the *readout gap*:

$$G_m = P_m - R_m, \tag{5}$$

which compares two readouts on the same representation and input state. Since a fitted probe cannot improve on the best readout in its family, the loss reduction it achieves is at most Γ_m^{read} (Appendix B); a large G_m therefore reflects information that a simple affine map can already extract.

Meanwhile, the representation term is estimated by comparing P_m with the score of the independently trained unimodal model (Table 1). A failure with a large G_m and a P_m close to this reference is mainly a readout failure, whereas a P_m that stays well below the reference indicates a limitation of the representation.

3.4 Readout Repair

The probe shows how much of the failure a new readout could avoid, but it is a separate single-modality model and does not tell how the source model should be changed. We therefore test whether the source model itself can recover missing-input performance while f_A and f_B stay frozen. Two readout interventions are considered, with additional feature adaptation methods as reference. In all cases, complete-input predictions use the original source model.

Classifier refitting. Does the source classifier fail because it never observed missing inputs during training? We train two new classifiers with the architecture of h_0 on the frozen representations. h_{full} is trained on complete inputs only, as h_0 was. h_{mix} differs only in its training inputs: it also sees only-A and only-B inputs and minimizes

$$L_{\mathrm{mix}}(h) = \pi_{AB} L_{AB}(h) + \pi_A L_A(h) + \pi_B L_B(h), \qquad \pi_{AB} + \pi_A + \pi_B = 1, \tag{6}$$

where $L_s(h) = \mathbb{E}[\ell(h(z^{(s)}), Y)]$ is the expected loss of h in input state s, and $(\pi_{AB}, \pi_A, \pi_B) = (1/2, 1/4, 1/4)$. The gap between their single-modality scores measures the effect of exposure to missing inputs (Appendix B).

Gated residual repair (GRR). Can the readout gap be reduced without changing how the source model handles complete inputs? We propose GRR for this purpose. GRR keeps h_0 unchanged and adds a learned affine correction to the source readout, applied only when a single modality is observed:

$$\hat{y} = \begin{cases} h_0(z^{(AB)}), & \text{both modalities observed,} \\ t_m(z_m) + V_m \bar{z}_m + c_m, & \text{only modality } m \text{ observed,} \end{cases} \tag{7}$$

where $\bar{z}_m = D_m^{-1}(z_m - \mu_m)$ is z_m standardized with the training-set mean μ_m and the diagonal matrix D_m of training-set standard deviations. Only $V_m \in \mathbb{R}^{C \times d_m}$ and $c_m \in \mathbb{R}^C$ are trained. Both start at zero, so GRR starts from the source readout, and complete-input predictions remain those of h_0. When t_m is affine, adding $V_m \bar{z}_m + c_m$ can reach any affine readout of z_m (Appendix B). The best GRR therefore matches the best affine probe and removes the readout term Γ_m^{read}; in other words, GRR turns the diagnostic probe into a repair of the source model. When t_m is nonlinear, GRR keeps it and adds an affine correction.

Feature adaptation. To see what remains after the readout is repaired, we also adapt the features themselves, at two levels of cost. PEA-SSF (Reza et al., 2025) keeps all source weights frozen and learns scale-and-shift parameters on intermediate features, separately for each missing direction. ModDrop (Neverova et al., 2016) fine-tunes a full copy of the source model, encoders included, under the same availability mixture as h_{mix}. Unlike the readout interventions, both methods change the retained representation z_m, so their gains beyond GRR point to a limitation of the representation rather than of the readout.

4 Experiments

4.1 Experimental Setup

The experiments cover three two-modality classification tasks from different domains: CREMA-D emotion recognition (Cao et al., 2014) with audio and video, UPMC-Food101 classification (Wang

Table 1: Datasets and unimodal references. Unimodal A/B is the score of a model trained on that modality alone (mean $\pm$ standard deviation over three runs).

Dataset	Modalities A/B	Metric	Train/val/test	Unimodal A	Unimodal B
CREMA-D	Audio / Video	Accuracy	6,028/670/744	56.05 $\pm$1.45	45.21 $\pm$0.79
Food101	Text / Image	Accuracy	60,734/8,675/17,359	84.80 $\pm$0.14	67.18 $\pm$0.29
MIMIC	EHR / CXR	AUPRC	4,995/723/1,436	47.78 $\pm$0.19	40.40 $\pm$0.21

Table 2: Fusion and single-modality performance of the source models. F: complete-input score. $R_m \rightarrow P_m$: source-model score with only modality m, followed by the affine-probe score on the same frozen representation. Metrics as in Table 1.

Dataset	**Source**	Fusion F	Only A: $R_A \rightarrow P_A$	Only B: $R_B \rightarrow P_B$
CREMA-D	Concat	58.24	49.42 $\rightarrow$ 56.00	18.86 $\rightarrow$ 22.76
A: Audio	Concat+Aux	63.40	56.18 $\rightarrow$ 57.93	27.02 $\rightarrow$ 46.06
B: Video	OGM-GE	65.82	42.34 $\rightarrow$ 56.59	39.74 $\rightarrow$ 46.42
	MLB	71.15	23.70 $\rightarrow$ 51.75	30.47 $\rightarrow$ 52.24
	AUG	70.74	51.75 $\rightarrow$ 52.73	54.53 $\rightarrow$ 57.17
	MASAM	70.92	43.91 $\rightarrow$ 57.84	17.70 $\rightarrow$ 36.56
Food101	Concat	90.35	83.64 $\rightarrow$ 84.54	46.80 $\rightarrow$ 54.77
A: Text	Concat+Aux	92.09	84.49 $\rightarrow$ 85.06	63.45 $\rightarrow$ 67.67
B: Image	OGM-GE	89.39	83.10 $\rightarrow$ 83.83	47.92 $\rightarrow$ 54.49
	MLB	92.20	84.02 $\rightarrow$ 84.80	63.11 $\rightarrow$ 67.44
	AUG	89.77	84.80 $\rightarrow$ 85.20	66.24 $\rightarrow$ 67.98
	MASAM	91.58	83.51 $\rightarrow$ 84.05	51.27 $\rightarrow$ 57.16
MIMIC	Concat	49.74	46.27 $\rightarrow$ 46.89	37.59 $\rightarrow$ 38.15
A: EHR	Concat+Aux	49.86	46.52 $\rightarrow$ 47.38	39.28 $\rightarrow$ 39.92
B: CXR	OGM-GE	49.39	48.07 $\rightarrow$ 47.60	33.58 $\rightarrow$ 34.87
	MLB	49.86	46.42 $\rightarrow$ 47.55	39.30 $\rightarrow$ 40.01
	AUG	50.47	47.71 $\rightarrow$ 47.52	40.39 $\rightarrow$ 40.32
	MASAM	50.62	46.39 $\rightarrow$ 46.81	38.78 $\rightarrow$ 38.98

et al., 2015) with text and images, and MIMIC phenotype prediction from CareBench (Yin et al., 2026) with EHR and chest X-rays (CXR). In each task, A denotes the modality with the stronger independently trained unimodal model (audio, text, and EHR, respectively) and B the other; the gap between the two varies across tasks (Table 1). Performance is measured by accuracy on CREMA-D and Food101 and by macro AUPRC on MIMIC.

Source models are trained with different strategies that address modality imbalance at different points of training: Concat (standard feature fusion), Concat+Aux (with additional unimodal losses), OGM-GE and MLB (modified encoder optimization; Peng et al., 2022; Kontras et al., 2024), AUG (adaptive classifier allocation; Jiang et al., 2025), and MASAM (modality-specific sharpness-aware updates; Chen et al., 2026). Recovery experiments use Concat+Aux, MLB, AUG, and MASAM, which cover distinct training mechanisms and missing-input behaviors. All results are means over three runs, with standard deviations in Appendix C.

Probes and interventions are fitted on training labels and selected on validation data; test labels are used only for evaluation. Except for the feature adaptation methods PEA-SSF and ModDrop, the source encoders stay frozen. An adapted model is used for a missing direction only if it improves on the source model's validation score in that direction; otherwise, the source model is kept. Further implementation details are given in Appendix A.

4.2 Fusion–Missing Mismatch

Recent balancing methods clearly improve fusion: on CREMA-D, MLB, AUG, and MASAM gain more than 12 points over Concat, while differences on Food101 and MIMIC are smaller (Table 2). These gains do not carry over to single-modality inputs. MLB, AUG, and MASAM reach almost

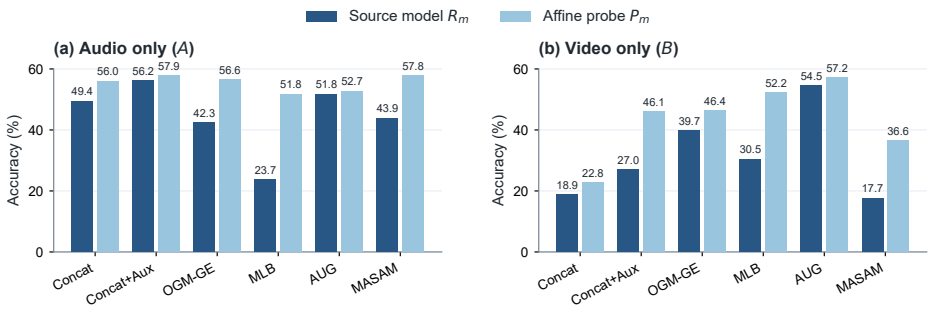

Figure 2: Readout gaps on CREMA-D with (a) audio or (b) video alone. Dark bars: source-model score R_m; light bars: affine-probe score P_m on the same frozen representation.

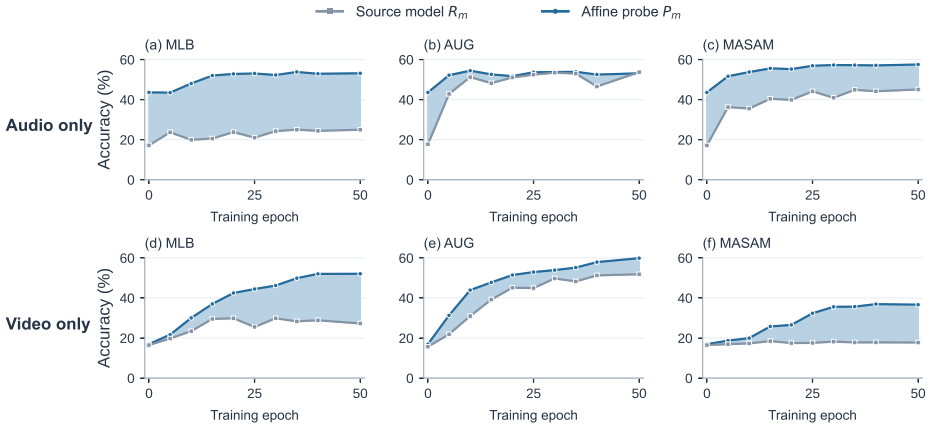

Figure 3: Readout gaps during CREMA-D training. Source-model score R_m and affine-probe score P_m at selected training checkpoints with audio only (top) or video only (bottom); shading marks the readout gap $P_m - R_m$.

the same fusion accuracy on CREMA-D, yet their video-only accuracies range from below 20% to above 50%. Compared with Concat+Aux, MLB fuses better but loses more than half of its audio-only accuracy, so the model with the highest fusion accuracy has the lowest audio-only accuracy.

Which branch fails depends on the training method rather than on the task. On CREMA-D, MLB fails mainly on audio, the stronger modality, whereas MASAM fails mainly on video. On Food101, methods that agree on fusion and text-only accuracy differ widely on images, while on MIMIC the differences are small in both directions. This is the behavior that complete-input underdetermination permits: the training objective fixes the joint output but not how each branch performs alone. Fusion accuracy therefore cannot serve as a proxy for missing-modality robustness.

4.3 Readout Gaps

For each source model and missing direction, an affine probe is fitted on the frozen retained representation, and its score P_m is compared with R_m and with the independently trained unimodal model (Table 2, Figure 2).

On CREMA-D, an affine probe on MLB's frozen audio features more than doubles its original audio-only accuracy, and its video branch shows a similar gap. MLB's video probe even exceeds the independently trained video model, as does AUG's. Joint training has thus improved the weaker modality's representation, as balancing methods intend, but the source classifier cannot use this improvement once the other modality is absent. Judged by R_m alone, the failure would be attributed to the representation rather than to the readout. MASAM shows a different case: its video probe roughly doubles the original video-only accuracy, yet remains well below the unimodal video model. Part of its failure can be recovered from the frozen features, and part is missing from them.

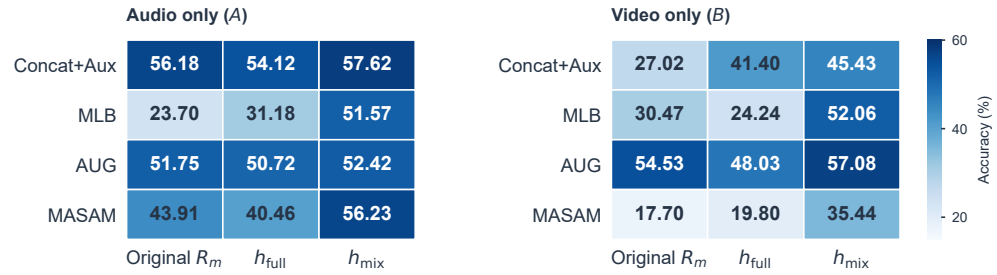

Figure 4: Classifier refitting on frozen CREMA-D representations. A classifier with the architecture of h_0 is refitted on complete inputs only (h_{full}) or on complete and single-modality inputs (h_{mix}); R_m is the original source-model score. For MLB and MASAM, both R_m and h_{full} fall well below h_{mix}.

Table 3: Recovery with readout repair (GRR, shaded) and feature adaptation (PEA-SSF, ModDrop). $R_m(P_m)$: source-model score and affine-probe score. Bold: GRR above both adaptation methods; underline: ModDrop more than 5 points above GRR. Entries equal to R_m fall back to the source model because adaptation did not improve validation performance. Metrics as in Table 1.

		(a) Only A available				(b) Only B available			
Dataset	**Model**	$R_A(P_A)$	GRR	PEA-SSF	ModDrop	$R_B(P_B)$	GRR	PEA-SSF	ModDrop
CREMA-D	Concat+Aux	56.18 (57.93)	**57.30**	56.68	57.12	27.02 (46.06)	46.37	41.53	48.66
	MLB	23.70 (51.75)	51.66	51.97	52.20	30.47 (52.24)	50.54	49.73	52.33
	AUG	51.75 (52.73)	52.96	53.54	52.78	54.53 (57.17)	59.27	60.57	60.35
	MASAM	43.91 (57.84)	56.32	58.38	59.01	17.70 (36.56)	36.65	34.86	<u>46.06</u>
Food101	Concat+Aux	84.49 (85.06)	**85.02**	84.93	84.50	63.45 (67.67)	67.45	67.50	63.45
	MLB	84.02 (84.80)	**84.79**	84.55	84.09	63.11 (67.44)	67.31	67.59	63.11
	AUG	84.80 (85.20)	**85.19**	85.02	84.80	66.24 (67.98)	67.98	68.56	66.24
	MASAM	83.51 (84.05)	84.05	84.17	84.64	51.27 (57.16)	56.91	59.79	<u>63.15</u>
MIMIC	Concat+Aux	46.52 (47.38)	**47.81**	46.52	47.21	39.28 (39.92)	**40.05**	39.86	39.55
	MLB	46.42 (47.55)	**47.78**	46.42	47.03	39.30 (40.01)	**40.09**	39.79	39.41
	AUG	47.71 (47.52)	47.86	47.88	47.71	40.39 (40.32)	40.39	40.75	40.39
	MASAM	46.39 (46.81)	**47.49**	46.39	46.74	38.78 (38.98)	39.35	39.44	39.29

Readout gaps are smaller on the other tasks. On Food101, readout gaps are small for text and moderate for images. On MIMIC, missing-input scores are mostly close to the unimodal references and readout gaps are near zero, so joint training leaves single-modality prediction largely intact. An MLP probe shows a similar trend (Table C2).

The readout gap also opens during training rather than appearing only upon completion of training (Figure 3). On CREMA-D, MASAM's video probe improves overall during training, while its original video-only prediction stays near chance level, and MLB's video readout gap widens markedly. Training generally improves what each branch supports on its own, while the source readout, shaped only by complete inputs, does not follow.

4.4 Repair and Adaptation

The diagnosis predicts that a large readout gap can be recovered on frozen features, whereas a representation well below the unimodal reference requires changing the encoders. Figure 4 and Table 3 test both predictions.

On frozen CREMA-D features, a classifier refitted on complete inputs only (h_{full}) stays well below one that also sees single-modality inputs (h_{mix}; Figure 4). For MLB, this exposure alone raises audio-only and video-only accuracy by more than 20 points each. For MLB and MASAM, both the original R_m and h_{full} fall well below h_{mix}, so the failure follows from training on complete inputs alone.

Table 4: GRR with limited repair data. CREMA-D video-only accuracy with GRR fitted on 5–100% of the labeled training set; the source model and validation set are fixed.

		Fraction of repair data			
Source	Original R_B	5%	10%	25%	100%
MLB	30.47	46.01	48.61	48.66	50.54
AUG	54.53	55.78	56.94	57.03	59.27
MASAM	17.70	33.33	34.50	35.80	36.65

GRR closely matches the affine probe in all settings (Table 3), so it realizes nearly all of the diagnosed readout gap while leaving h_0 untouched. For MLB on CREMA-D, it more than doubles audio-only accuracy. The repair is also cheap: it trains only an affine correction per direction, and for MLB and MASAM a small fraction of the training data already recovers most of the gain (Table 4). A trained model can therefore be repaired afterwards without retraining any of its components.

The two branches whose probes fall furthest below the unimodal reference, MASAM's video branch on CREMA-D and its image branch on Food101, are the only ones where ModDrop improves on GRR by more than 5 points (9.41 and 6.24 points). Elsewhere, its gains over GRR are small, and on MIMIC neither feature adaptation method improves noticeably on GRR. PEA-SSF, which adapts features only through scale-and-shift parameters, brings at most modest gains over GRR and sometimes falls below it, including on MASAM's CREMA-D video branch. Changing the representation thus helps only when the representation itself is limited, and only with an adaptation strong enough to change it. The readout gap and the probe's distance to the unimodal reference indicate in advance which case applies.

5 Conclusion

Balanced multimodal models are usually compared by their performance with all modalities available. We show that such comparisons say little about behavior when a modality is missing: complete-input training leaves the standalone behavior of each branch undetermined, and models with similar fusion accuracy can fail in different branches and to very different degrees. Separating the retained representation from the classifier's use of it shows that many of these failures lie in the readout. The information needed for prediction remains in the frozen features, and balanced training sometimes even improves it, but a classifier trained only on complete inputs does not use it. Repairing the readout, by refitting the classifier with missing inputs or with GRR, recovers most of the lost performance at little cost, whereas adapting the encoders pays off mainly where the retained representation itself is weak. Missing-modality robustness should therefore be evaluated for each missing direction and diagnosed before a remedy is chosen. This study is limited to two-modality classification, and the repair assumes a small labeled set for each missing direction. We leave extending the diagnosis to more modalities and large pretrained multimodal models, as well as using it during training to keep each branch usable on its own, for future work.

AI use statement

Generative AI tools assisted with code implementation, literature searches, and polishing the writing.

Ethics statement

This work uses existing research datasets and involves secondary analysis without new human-subject data. The MIMIC experiments follow the access and use conditions of the source datasets.

Reproducibility statement

Appendix A documents data splits, optimization, model selection, probe settings, and method-specific implementation details. Appendix B gives the formal analysis, and Appendix C reports complete diagnostics and recovery results.

References

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *International Conference on Learning Representations, Workshop Track*, 2017.

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

Houwei Cao, David G. Cooper, Michael K. Keutmann, Ruben C. Gur, Ani Nenkova, and Ragini Verma. CREMA-D: Crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on Affective Computing*, 5(4):377–390, 2014.

Zijie Chen, Kejing Yin, Wenfang Yao, William K. Cheung, and Jing Qin. MASAM: Multimodal adaptive sharpness-aware minimization for heterogeneous data fusion. In *International Conference on Learning Representations*, 2026.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D. Hoffman, Farhad Hormozdiari, Neil Houlsby, Shaobo Hou, Ghassen Jerfel, Alan Karthikesalingam, Mario Lucic, Yian Ma, Cory McLean, Diana Mincu, Akinori Mitani, Andrea Montanari, Zachary Nado, Vivek Natarajan, Christopher Nielson, Thomas F. Osborne, Rajiv Raman, Kim Ramasamy, Rory Sayres, Jessica Schrouff, Martin Seneviratne, Shannon Sequeira, Harini Suresh, Victor Veitch, Max Vladymyrov, Xuezhi Wang, Kellie Webster, Steve Yadlowsky, Taedong Yun, Xiaohua Zhai, and D. Sculley. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(226):1–61, 2022.

Chenzhuang Du, Jiaye Teng, Tingle Li, Yichen Liu, Tianyuan Yuan, Yue Wang, Yang Yuan, and Hang Zhao. On uni-modal feature learning in supervised multi-modal learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 8632–8656, 2023.

Yunfeng Fan, Wenchao Xu, Haozhao Wang, Junxiao Wang, and Song Guo. PMR: Prototypical modal rebalance for multimodal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20029–20038, 2023.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021.

Zirun Guo, Tao Jin, Jingyuan Chen, and Zhou Zhao. Classifier-guided gradient modulation for enhanced multimodal learning. In *Advances in Neural Information Processing Systems*, volume 37, pp. 133328–133344, 2024.

Nasir Hayat, Krzysztof J. Geras, and Farah E. Shamout. MedFuse: Multi-modal fusion with clinical time-series data and chest X-ray images. In *Proceedings of the 7th Machine Learning for Healthcare Conference*, volume 182 of *Proceedings of Machine Learning Research*, pp. 479–503, 2022.

Devamanyu Hazarika, Yingting Li, Bo Cheng, Shuai Zhao, Roger Zimmermann, and Soujanya Poria. Analyzing modality robustness in multimodal sentiment analysis. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 685–696. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.naacl-main.50.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 2733–2743. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1275.

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

Yu Huang, Junyang Lin, Chang Zhou, Hongxia Yang, and Longbo Huang. Modality competition: What makes joint training of multi-modal network fail in deep learning? (Provably). In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 9226–9259, 2022.

Pavel Izmailov, Polina Kirichenko, Nate Gruver, and Andrew Gordon Wilson. On feature learning in the presence of spurious correlations. In *Advances in Neural Information Processing Systems*, volume 35, pp. 38516–38532, 2022.

Qing-Yuan Jiang, Longfei Huang, and Yang Yang. Rethinking multimodal learning from the perspective of mitigating classification ability disproportion. In *Advances in Neural Information Processing Systems*, volume 38, pp. 143717–143744, 2025.

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *International Conference on Learning Representations*, 2020.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. In *International Conference on Learning Representations*, 2023.

Konstantinos Kontras, Christos Chatzichristos, Matthew B. Blaschko, and Maarten De Vos. Improving multimodal learning with multi-loss gradient modulation. In *British Machine Vision Conference*. BMVA, 2024.

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.

Paul Pu Liang, Yun Cheng, Xiang Fan, Chun Kai Ling, Suzanne Nie, Richard Chen, Zihao Deng, Nicholas Allen, Randy Auerbach, Faisal Mahmood, Ruslan Salakhutdinov, and Louis-Philippe Morency. Quantifying & modeling multimodal interactions: An information decomposition framework. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. SMIL: Multimodal learning with severely missing modality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 2302–2310, 2021. doi: 10.1609/aaai.v35i3.16330.

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*, pp. 18177–18186, 2022.

Natalia Neverova, Christian Wolf, Graham W. Taylor, and Florian Nebout. ModDrop: Adaptive multi-modal gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(8):1692–1706, 2016.

Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8238–8247, 2022.

Md Kaykobad Reza, Ashley Prater-Bennette, and M. Salman Asif. Robust multimodal learning with missing modalities via parameter-efficient adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(2):742–754, 2025. doi: 10.1109/TPAMI.2024.3476487.

Elena Voita and Ivan Titov. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 183–196, 2020. doi: 10.18653/v1/2020.emnlp-main.14.

Hu Wang, Yuanhong Chen, Congbo Ma, Jodie Avery, Louise Hull, and Gustavo Carneiro. Multimodal learning with missing modality via shared-specific feature modelling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15878–15887, 2023.

Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12695–12705, 2020.

Xin Wang, Devinder Kumar, Nicolas Thome, Matthieu Cord, and Frédéric Precioso. Recipe recognition with large multimodal food dataset. In *IEEE International Conference on Multimedia and Expo Workshops*, pp. 1–6, 2015.

Yake Wei, Siwei Li, Ruoxuan Feng, and Di Hu. Diagnosing and re-learning for balanced multimodal learning. In *European Conference on Computer Vision*, pp. 71–86, 2024. doi: 10.1007/978-3-031-73039-9_5.

Nan Wu, Stanislaw Jastrzebski, Kyunghyun Cho, and Krzysztof J. Geras. Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 24043–24055, 2022.

Zhenbang Wu, Anant Dadu, Nicholas Tustison, Brian Avants, Mike Nalls, Jimeng Sun, and Faraz Faghri. Multimodal patient representation learning with missing modalities and labels. In *International Conference on Learning Representations*, 2024.

Kejing Yin, Haizhou Xu, Wenfang Yao, Chen Liu, Zijie Chen, Yui Haang Cheung, William K. Cheung, and Jing Qin. When does multimodal learning help in healthcare? a benchmark on EHR and chest X-ray fusion. *arXiv preprint arXiv:2602.23614*, 2026.

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

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

A Experimental Details

Table 1 lists the data splits and the independently trained unimodal references. For each task, all source methods use the same splits. Modality A is the one with the higher unimodal reference score and B is the other.

CREMA-D. CREMA-D (Cao et al., 2014) contains $7{,}442$ audiovisual clips of 91 actors speaking 12 sentences. Each clip is labeled with one of six emotions: neutral, happy, sad, fear, disgust, or angry. Audio is modality A and video is B. We keep the training and test lists of OGM-GE and InfoReg (Peng et al., 2022; Huang et al., 2025) and hold out a stratified 10% of the training clips for validation, which gives $6{,}028/670/744$ clips for training, validation, and testing. Audio is resampled to 22.05 kHz, repeated or truncated to three seconds, and converted to a log-magnitude STFT spectrogram with a 512-sample Hann window and a 353-sample hop. Video is sampled at one frame per second; training uses two sampled frames and evaluation uses the first and last frames. Both encoders are ResNet-18, and video frames are 224×224. We train with cross-entropy and report accuracy.

Food101. UPMC-Food101 (Wang et al., 2015) pairs food images with text from the web pages they come from. It covers 101 categories, with one label per pair. Text is modality A and image is B. After preprocessing, $86{,}768$ pairs are usable. We pool the official training and test files and draw a stratified $70/10/20$ split, which gives $60{,}734/8{,}675/17{,}359$ pairs. Text is truncated or padded to 50 tokens and encoded offline with `bert-base-uncased`; we store the 50×768 last-layer features, and a one-layer Transformer (hidden size 512, four heads) learns the task representation from them. The image encoder is an ImageNet-pretrained ResNet-50, trained with random resized crops and horizontal flips and evaluated at 224×224 with ImageNet normalization. We train with cross-entropy and report accuracy.

MIMIC. We use the matched EHR–CXR cohort of CareBench (Yin et al., 2026), built from MIMIC-IV electronic health records and MIMIC-CXR chest radiographs. Each example is an ICU stay with structured clinical measurements and a paired chest X-ray, and the target is a vector of 25 binary phenotype labels. EHR is modality A and CXR is B. We use fold 1 of the processed cohort, with $4{,}995/723/1{,}436$ stays for training, validation, and testing. The EHR encoder is a one-layer Transformer (hidden size 256, four heads) over sequences with 49 input channels; the CXR encoder is an ImageNet-pretrained ResNet-50. We train with binary cross-entropy and report macro AUPRC, the unweighted mean of average precision over the 25 labels, in percent.

Source models. Concat is trained with the fusion loss only, and Concat+Aux adds the two unimodal losses with unit weight. OGM-GE modulates encoder gradients (Peng et al., 2022). MLB uses the Concat+Aux objective and rescales encoder updates according to the unimodal prediction scores (Kontras et al., 2024). MASAM selects an encoder block for a sharpness-aware update (Foret et al., 2021) based on convergence and gradient alignment (Chen et al., 2026). AUG uses separate modality classifiers with a shared output layer, updates the modalities independently, and allocates classifiers adaptively (Jiang et al., 2025). Table A1 lists all settings. Source models are selected by complete-input validation performance. On MIMIC, source models and interventions are selected by validation macro AUROC, and test scores are macro AUPRC.

Seeds. Data splits use seed 42. Each source method is trained with seeds 42, 123, and 1234, except CREMA-D Concat+Aux, which uses 2026 in place of 1234; the training-trajectory runs in Appendix C use 42, 123, and 1234. Affine probes, GRR, and the CREMA-D adaptation methods are fitted with seed 42, and the only-B PEA-SSF fits on Food101 and MIMIC use the seed of their source model. The fixed partner encoders in Figure C3 come from the Concat+Aux model with seed 42.

Probes and interventions. For each source model, we cache the training, validation, and test features and the original logits, with the encoders in evaluation mode. For feature-fusion sources, the absent feature block is set to zero before h_0 (Equation (1)); for AUG, only the weighted logits of the available branch are kept. Probe inputs are standardized with training-set statistics. The MLP

probe has one hidden layer of 256 ReLU units. Endpoint probes train for at most 50 epochs with validation patience 5.

h_{full} and h_{mix} start from the same initialization and use the same validation score, which weights complete, only-A, and only-B inputs by $1/2$, $1/4$, and $1/4$; missing blocks are set to zero before standardization. GRR fits a zero-initialized residual on the cached features and logits (Algorithm A1). PEA-SSF learns scale-and-shift parameters on intermediate features while the source weights and normalization statistics stay frozen. ModDrop fine-tunes one copy of the source model under the same $1/2, 1/4, 1/4$ availability mixture and keeps the best validation state for each missing direction separately; its complete-input predictions come from the original model.

Table A1: Training settings.

Component	Settings
Source training	Budget of 50 epochs; batch size 64 for CREMA-D and 16 for Food101/MIMIC. CREMA-D feature-fusion sources use learning rate 0.002, momentum 0.9, weight decay 10^{-4}, and a 0.1 learning-rate factor at epoch 30. Food101/MIMIC use learning rate 10^{-4}.
MLB	$b_{\max} = 2$; $\alpha = 5$ on CREMA-D/MIMIC and 3 on Food101.
AUG	Target ratio 1.015877; margin 0.031404; allocation interval 15 epochs; classifier hidden size 256. AdamW at 10^{-4}. Merge weight 0.4 on CREMA-D/Food101 and 0.648392 on MIMIC.
MASAM	$\rho = 0.3$; score weight 0.8; score momentum 0.91.
Endpoint probes	Adam at 10^{-3}; batch size 256; at most 50 epochs; validation patience 5. MLP probes use the same budget.
h_{full}, h_{mix}	Adam at 10^{-3}; batch size 256; at most 50 epochs; patience 5. Model selection uses the same availability mixture for both.
Trajectory probes	Adam at 10^{-3}; batch size 256; 2,000 updates; validation every 50 updates.
GRR	Adam at 10^{-3}; batch size 256; 2,000 updates; validation every 50 updates; zero initialization.
LogitCal	Identity-initialized affine score map; Adam at 10^{-3}; batch size 256; 2,000 updates; validation every 50 updates.
PEA-SSF	Adam at 10^{-4}; source weights and normalization statistics frozen.
ModDrop	AdamW at 10^{-4} with weight decay 0.01; complete/only-A/only-B probabilities $1/2$, $1/4$, $1/4$.
CREMA-D adaptation	Batch size 256; 2,000 updates; validation every 50 updates, including initialization.
Food101/MIMIC adaptation	Source batch sizes; at most 50 epochs; validation each epoch. PEA-SSF stops after 5 epochs without improvement in its target direction; ModDrop stops after 5 epochs without improvement in either direction.

Repair data. Table 4 fits GRR on stratified, nested subsets of the CREMA-D training set with 301, 603, 1,506, or 6,028 examples, while the source models are trained on all 6,028 clips. Each fit standardizes features with the statistics of its subset and otherwise follows the main GRR setting (Table A1). The best state is selected by accuracy on the same 670 validation clips, evaluated every 50 updates.

GRR. Algorithm A1 summarizes how GRR is fitted and used for prediction. The task loss is cross-entropy on CREMA-D and Food101 and binary cross-entropy on MIMIC. For a direction with d_m retained features and C outputs, GRR learns $C(d_m + 1)$ parameters. Table A2 reports the parameter counts and fitting times on cached features; one fit takes between 4.07 and 6.63 seconds on average.

Algorithm A1 Gated Residual Repair (GRR)

Input: source encoders f_A, f_B and classifier h_0; labeled training and validation sets for repair. The source readout t_m is defined in Section 3.1.

1. Freeze f_A, f_B, h_0 in evaluation mode and cache $(z_m, t_m(z_m), y)$ on the training and validation sets for $m \in \{A, B\}$.
2. **For each** retained modality $m \in \{A, B\}$:
3. Compute the training mean μ_m and standard deviation σ_m; set $D_m = \mathrm{diag}(\max(\sigma_m, 10^{-6}))$ and $\bar{z}_m = D_m^{-1}(z_m - \mu_m)$.
4. Initialize $V_m = 0$, $c_m = 0$ and record the validation score.
5. **For** $k = 1, \ldots, 2{,}000$:
6. On the next shuffled training minibatch (size 256), compute the task loss of $t_m(z_m) + V_m \bar{z}_m + c_m$.
7. Update only (V_m, c_m) with Adam, learning rate 10^{-3}.
8. Every 50 updates, evaluate on validation and keep the best state.
9. Restore the best (V_m, c_m); the zero initialization is also a candidate.

Prediction: with both modalities, return $h_0(z^{(AB)})$; with only modality m, return $t_m(z_m) + V_m \bar{z}_m + c_m$.

Table A2: GRR parameters and fitting time per missing direction.

			Fit time (s)	
Dataset	**Source**	Parameters per direction	Only A	Only B
CREMA-D	Concat+Aux	3,078	4.64 ±0.05	4.17 ±0.08
	MLB	3,078	4.66 ±0.05	4.07 ±0.03
	AUG	3,078	4.68 ±0.06	4.09 ±0.07
	MASAM	3,078	4.58 ±0.04	4.08 ±0.07
Food101	Concat+Aux	51,813	5.02 ±0.06	4.58 ±0.07
	MLB	51,813	5.04 ±0.11	4.72 ±0.17
	AUG	51,813	5.01 ±0.02	4.57 ±0.01
	MASAM	51,813	5.08 ±0.06	4.61 ±0.09
MIMIC	Concat+Aux	6,425	6.53 ±0.09	6.07 ±0.09
	MLB	6,425	6.63 ±0.05	6.13 ±0.07
	AUG	6,425	6.61 ±0.14	6.09 ±0.09
	MASAM	6,425	6.62 ±0.05	6.09 ±0.05

B Theory and Analysis

This appendix gives the formal statements behind Section 3. We use the input states $s \in \{AB, A, B\}$ and the state-wise losses $L_s(h) = \mathbb{E}[\ell(h(z^{(s)}), Y)]$ defined there.

Underdetermination. Consider the affine fusion classifier h in Equation (2) and a modified classifier h' with branch weights $W_A + \Delta_A$ and $W_B + \Delta_B$. On complete inputs, the two differ by $\delta(z_A, z_B) = \Delta_A z_A + \Delta_B z_B$; with only A or only B, they differ by $\Delta_A z_A$ or $\Delta_B z_B$. If $\delta = 0$ almost surely, h and h' have the same complete-input logits and risk, while their single-modality logits still differ by these two terms.

When the cancellation is only approximate, assume that $\ell(\cdot, Y)$ is ρ-Lipschitz in a fixed norm on the logits and that all risks are finite. The Lipschitz property applied pointwise gives

$$\begin{aligned} |L_{AB}(h') - L_{AB}(h)| &\le \mathbb{E}\big|\ell(h([z_A, z_B]) + \delta(z_A, z_B), Y) - \ell(h([z_A, z_B]), Y)\big| \\ &\le \rho\, \mathbb{E}\|\Delta_A z_A + \Delta_B z_B\|. \end{aligned} \tag{8}$$

To see how δ can be small while the single-modality changes are large, suppose the two representations are approximately linearly related, $z_B = T z_A + \xi$ with residual ξ. For any matrix U, choosing $(\Delta_A, \Delta_B) = (-UT, U)$ gives $\delta = U\xi$, and hence

$$|L_{AB}(h') - L_{AB}(h)| \le \rho\, \mathbb{E}\|U\xi\| \le \rho \|U\|_{\mathrm{op}}\, \mathbb{E}\|\xi\|. \tag{9}$$

For a fixed U and a small residual ξ, the complete-input risk barely changes, whereas the single-modality changes $-UTz_A$ and Uz_B can be large. If $\xi = 0$ almost surely, every such U leaves the complete-input logits unchanged.

Exact cancellation. In the example of Section 3.2, $z_A = z_B = Y$, so all three classifiers output Y on complete inputs. With only A, they output $Y/2$, $-Y$, and $2Y$; with only B, they output $Y/2$, $2Y$, and $-Y$. Taking signs, their only-A accuracies are 1, 0, and 1, and their only-B accuracies are 1, 1, and 0, while the representations are the same in all three cases.

Loss decomposition. Fix the data distribution, the loss ℓ, and a readout family $\mathcal{H}_m$ on the feature space of modality m. Let z_m be the retained representation of the source model and z_m^{uni} that of an independently trained unimodal model for the same task, and assume that all risks and infima below are finite. Adding and subtracting $L_m^\star(z_m)$ gives Equation (4), with readout term $\Gamma_m^{\mathrm{read}} = L(t_m; z_m) - L_m^\star(z_m)$ and representation term $\Gamma_m^{\mathrm{repr}} = L_m^\star(z_m) - L_m^\star(z_m^{\mathrm{uni}})$. If $t_m \in \mathcal{H}_m$, the definition of the infimum gives $L_m^\star(z_m) \le L(t_m; z_m)$, so $\Gamma_m^{\mathrm{read}} \ge 0$. If t_m is nonlinear and $\mathcal{H}_m$ is affine, as for AUG, this inclusion can fail and the readout term can be negative. The representation term can have either sign, because joint training can make the best risk in $\mathcal{H}_m$ lower or higher than that of the unimodal reference.

A fitted probe $\widehat{p}_m \in \mathcal{H}_m$ has population excess risk $\epsilon_m = L(\widehat{p}_m; z_m) - L_m^\star(z_m) \ge 0$, and

$$L(t_m; z_m) - L(\widehat{p}_m; z_m) = \Gamma_m^{\mathrm{read}} - \epsilon_m. \tag{10}$$

The loss reduction achieved by a fitted probe is therefore at most Γ_m^{read}, so a large reduction implies a large readout term.

Mixture risk. The mixture risk in Equation (6) can separate classifiers that agree on complete inputs, through their single-modality risks.

Proposition 1 (Availability-aware risk comparison) *If $h([z_A, z_B]) = h'([z_A, z_B])$ almost surely, then*

$$L_{\mathrm{mix}}(h') - L_{\mathrm{mix}}(h) = \pi_A \Delta L_A + \pi_B \Delta L_B, \tag{11}$$

where $\Delta L_m = L_m(h') - L_m(h)$.

Proof. Equal complete-input logits give equal complete-input risks, $L_{AB}(h') = L_{AB}(h)$. Subtracting Equation (6) for h from that for h' cancels the complete-input term and gives Equation (11).

The complete-input risk cannot distinguish h from h', whereas the mixture risk does whenever $\pi_A \Delta L_A + \pi_B \Delta L_B \neq 0$. With positive weights, lowering one single-modality risk without raising the other strictly lowers the mixture risk.

GRR expressivity. Let $\mathcal{H}_m^{\text{aff}}$ be the family of all affine readouts on z_m, and let $\mathcal{Q}_m^{\text{GRR}}$ be the family of single-modality GRR readouts $q_m(z_m) = t_m(z_m) + V_m \bar{z}_m + c_m$ in Equation (7).

Corollary 1 (Affine repairability) *If t_m is affine and $\mathcal{H}_m = \mathcal{H}_m^{\text{aff}}$, then $\mathcal{Q}_m^{\text{GRR}} = \mathcal{H}_m^{\text{aff}}$. Thus*

$$L(t_m; z_m) - \inf_{q \in \mathcal{Q}_m^{\text{GRR}}} L(q; z_m) = \Gamma_m^{\text{read}}. \tag{12}$$

Proof. Let $t_m(z) = W_m z + b_m$ and $\bar{z} = D_m^{-1}(z - \mu_m)$, with D_m positive diagonal. Expanding the GRR readout gives

$$q_m(z) = (W_m + V_m D_m^{-1}) z + b_m + c_m - V_m D_m^{-1} \mu_m, \tag{13}$$

which is affine. Conversely, any affine readout $Az + a$ is obtained with $V_m = (A - W_m) D_m$ and $c_m = a - b_m + (A - W_m)\mu_m$. The two families are therefore equal and have the same infimum risk $L_m^\star(z_m)$. Subtracting it from $L(t_m; z_m)$ gives Equation (12).

If t_m is nonlinear, GRR spans $t_m + \mathcal{H}_m^{\text{aff}}$: it keeps the nonlinear source readout and adds an affine correction.

Conditional gain. Let P_{AB} be the score of a joint affine probe on a representation pair, and define the conditional gains $\Delta_{B|A} = P_{AB} - P_A$ and $\Delta_{A|B} = P_{AB} - P_B$. They measure how much one representation adds to a given partner under a given probe protocol. Replacing z_A changes both the joint prediction problem and the single-modality reference P_A, even when z_B is unchanged, so $\Delta_{B|A}$ is a property of the pair rather than of the B encoder alone. This complements information-decomposition views of multimodal interactions (Liang et al., 2023). Figure C3 shows this dependence empirically.

C Additional Results

Diagnostics and probes. Table C1 reports all diagnostic scores with standard deviations. It also includes the score P_{AB} of a joint affine probe on $[z_A, z_B]$ and the conditional gain $\Delta_{B|A} = P_{AB} - P_A$ (Appendix B). Table C2 compares affine and MLP probes, which give a similar picture. On CREMA-D, MASAM's video probe stays below that of Concat+Aux with both probe families: 36.56% versus 46.06% for affine probes and 38.17% versus 45.56% for MLP probes. The same ordering holds for Food101 images, so the weaker branch of MASAM is less predictive even under the more expressive probe.

Table C1: Diagnostic scores of all source models.

Dataset	Source	F	R_A	P_A	R_B	P_B	P_{AB}	$\Delta_{B\|A}$
CREMA-D	Concat	58.24 ±0.87	49.42 ±2.64	56.00 ±1.51	18.86 ±0.21	22.76 ±1.09	59.90 ±1.55	3.90 ±0.48
	Concat+Aux	63.40 ±1.35	56.18 ±1.28	57.93 ±1.30	27.02 ±2.91	46.06 ±2.71	67.79 ±2.02	9.86 ±1.20
	OGM-GE	65.82 ±2.41	42.34 ±2.36	56.59 ±3.66	39.74 ±3.09	46.42 ±0.95	67.65 ±2.07	11.07 ±2.64
	MLB	71.15 ±2.44	23.70 ±1.53	51.75 ±1.32	30.47 ±1.35	52.24 ±0.82	70.34 ±2.44	18.59 ±1.35
	AUG	70.74 ±0.54	51.75 ±3.81	52.73 ±2.05	54.53 ±2.48	57.17 ±1.36	69.27 ±0.62	16.53 ±1.64
	MASAM	70.92 ±0.87	43.91 ±3.65	57.84 ±1.08	17.70 ±0.21	36.56 ±1.58	70.12 ±0.68	12.28 ±0.91
Food101	Concat	90.35 ±0.16	83.64 ±0.27	84.54 ±0.34	46.80 ±0.86	54.77 ±1.01	90.84 ±0.06	6.30 ±0.33
	Concat+Aux	92.09 ±0.22	84.49 ±0.22	85.06 ±0.14	63.45 ±0.79	67.67 ±0.36	92.54 ±0.07	7.48 ±0.20
	OGM-GE	89.39 ±0.21	83.10 ±0.18	83.83 ±0.24	47.92 ±0.93	54.49 ±0.81	90.07 ±0.15	6.24 ±0.09
	MLB	92.20 ±0.09	84.02 ±0.42	84.80 ±0.29	63.11 ±0.32	67.44 ±0.49	92.61 ±0.07	7.80 ±0.22
	AUG	89.77 ±0.30	84.80 ±0.02	85.20 ±0.06	66.24 ±0.29	67.98 ±0.30	92.18 ±0.25	6.97 ±0.22
	MASAM	91.58 ±0.08	83.51 ±0.16	84.05 ±0.13	51.27 ±0.52	57.16 ±0.49	91.63 ±0.05	7.58 ±0.12
MIMIC	Concat	49.74 ±0.05	46.27 ±0.10	46.89 ±0.36	37.59 ±0.46	38.15 ±0.36	49.13 ±0.26	2.23 ±0.40
	Concat+Aux	49.86 ±0.32	46.52 ±0.23	47.38 ±0.53	39.28 ±0.62	39.92 ±0.60	49.21 ±0.22	1.82 ±0.62
	OGM-GE	49.39 ±0.10	48.07 ±0.23	47.60 ±0.24	33.58 ±0.56	34.87 ±0.61	48.88 ±0.15	1.28 ±0.38
	MLB	49.86 ±0.02	46.42 ±0.29	47.55 ±0.07	39.30 ±0.13	40.01 ±0.18	49.34 ±0.12	1.79 ±0.16
	AUG	50.47 ±0.15	47.71 ±0.08	47.52 ±0.09	40.39 ±0.08	40.32 ±0.19	49.21 ±0.56	1.69 ±0.58
	MASAM	50.62 ±0.02	46.39 ±0.25	46.81 ±0.11	38.78 ±0.13	38.98 ±0.23	49.39 ±0.11	2.59 ±0.21

Table C2: Affine versus MLP probes.

		Affine				MLP			
Dataset	Source	P_A	P_B	P_{AB}	$\Delta_{B\|A}$	P_A	P_B	P_{AB}	$\Delta_{B\|A}$
CREMA-D	Concat	56.00 ±1.51	22.76 ±1.09	59.90 ±1.55	3.90 ±0.48	55.78 ±1.61	22.13 ±1.25	58.69 ±0.81	2.91 ±0.81
	Concat+Aux	57.93 ±1.30	46.06 ±2.71	67.79 ±2.02	9.86 ±1.20	57.35 ±2.02	45.56 ±1.10	67.07 ±2.20	9.72 ±0.68
	OGM-GE	56.59 ±3.66	46.42 ±0.95	67.65 ±2.07	11.07 ±2.64	56.81 ±2.67	45.56 ±0.93	65.95 ±2.41	9.14 ±1.76
	MLB	51.75 ±1.32	52.24 ±0.82	70.34 ±2.44	18.59 ±1.35	52.91 ±0.85	51.52 ±1.44	70.03 ±2.16	17.11 ±1.42
	AUG	52.73 ±2.05	57.17 ±1.36	69.27 ±0.62	16.53 ±1.64	53.14 ±1.79	57.57 ±0.86	69.62 ±1.55	16.49 ±0.54
	MASAM	57.84 ±1.08	36.56 ±1.58	70.12 ±0.68	12.28 ±0.91	55.73 ±1.48	38.17 ±0.82	69.94 ±1.90	14.20 ±2.89
Food101	Concat	84.54 ±0.34	54.77 ±1.01	90.84 ±0.06	6.30 ±0.33	84.74 ±0.09	54.48 ±0.56	90.34 ±0.15	5.60 ±0.05
	Concat+Aux	85.06 ±0.14	67.67 ±0.36	92.54 ±0.07	7.48 ±0.20	84.93 ±0.03	67.08 ±0.12	92.00 ±0.15	7.07 ±0.14
	OGM-GE	83.83 ±0.24	54.49 ±0.81	90.07 ±0.15	6.24 ±0.09	83.79 ±0.26	53.41 ±0.73	89.34 ±0.19	5.56 ±0.17
	MLB	84.80 ±0.29	67.44 ±0.49	92.61 ±0.07	7.80 ±0.22	84.76 ±0.16	66.62 ±0.17	92.09 ±0.04	7.33 ±0.20
	AUG	85.20 ±0.06	67.98 ±0.30	92.18 ±0.25	6.97 ±0.22	85.11 ±0.12	67.25 ±0.15	91.62 ±0.05	6.51 ±0.08
	MASAM	84.05 ±0.13	57.16 ±0.49	91.63 ±0.05	7.58 ±0.12	84.09 ±0.04	56.41 ±0.15	91.14 ±0.07	7.05 ±0.05
MIMIC	Concat	46.89 ±0.36	38.15 ±0.36	49.13 ±0.26	2.23 ±0.40	47.31 ±0.16	38.20 ±0.44	48.99 ±0.07	1.69 ±0.17
	Concat+Aux	47.38 ±0.53	39.92 ±0.60	49.21 ±0.22	1.82 ±0.62	47.41 ±0.18	39.94 ±0.30	48.94 ±0.36	1.54 ±0.53
	OGM-GE	47.60 ±0.24	34.87 ±0.61	48.88 ±0.15	1.28 ±0.38	47.85 ±0.18	34.78 ±0.76	49.06 ±0.44	1.21 ±0.54
	MLB	47.55 ±0.07	40.01 ±0.18	49.34 ±0.12	1.79 ±0.16	47.36 ±0.08	39.89 ±0.03	49.36 ±0.28	1.99 ±0.20
	AUG	47.52 ±0.09	40.32 ±0.19	49.21 ±0.56	1.69 ±0.58	47.43 ±0.14	40.25 ±0.11	49.28 ±0.68	1.86 ±0.60
	MASAM	46.81 ±0.11	38.98 ±0.23	49.39 ±0.11	2.59 ±0.21	47.19 ±0.39	39.12 ±0.22	49.55 ±0.34	2.36 ±0.53

Training trajectories. We train separate Concat+Aux, MLB, AUG, and MASAM models on CREMA-D. At epochs $\{0, 5, 10, 15, 20, 25, 30, 35, 40, 50\}$, a new affine probe is fitted on the frozen training features with 2,000 Adam updates, and validation accuracy is evaluated every 50 updates. Figure C1 tracks single-modality and joint probe scores. For MASAM, from epoch 10 to 50, P_{AB} rises from 54.70% to 70.25%, P_A from 53.76% to 57.53%, and P_B from 19.94% to 36.65%. Joint and single-modality predictive ability thus develop at different rates within the same model. Figure C2 repeats Figure 3 with Concat+Aux added.

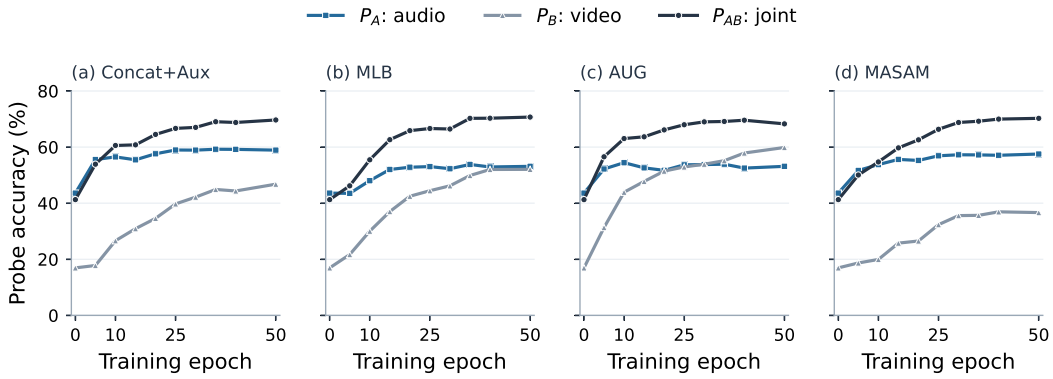

Figure C1: Single-modality and joint probe scores during CREMA-D training.

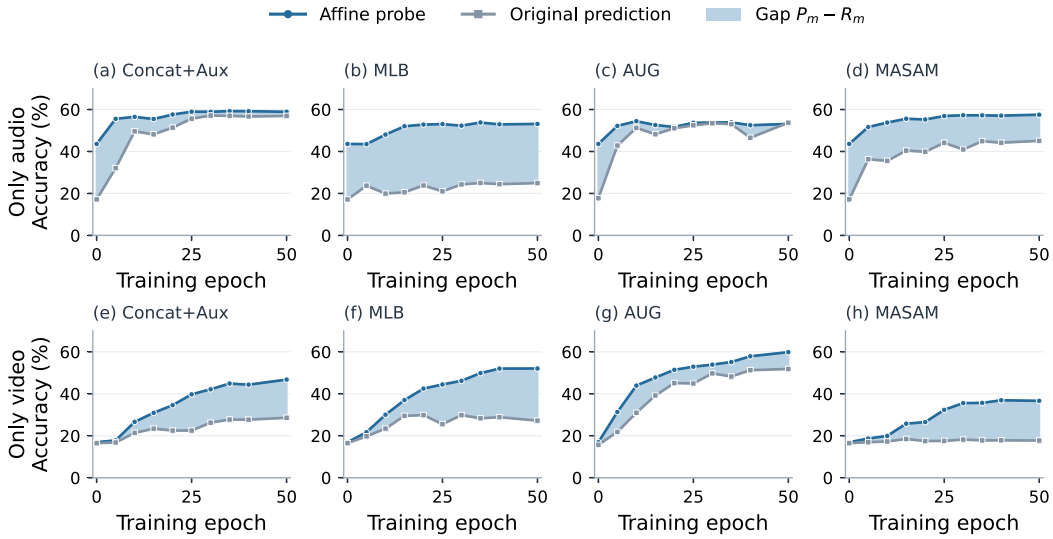

Figure C2: Readout gaps during CREMA-D training for all four source models.

Partner dependence. Figure C3 pairs each video (CREMA-D) or image (Food101) representation either with its native partner or with a fixed Concat+Aux audio or text encoder, and fits a new joint affine probe on each pair. On CREMA-D, MASAM's $\Delta_{B|A}$ exceeds that of Concat+Aux with native partners (12.28 versus 9.86 points), but the order reverses with the fixed audio encoder (8.69 versus 10.57). On Food101, the values are 7.58 versus 7.48 points with native partners and 6.31 versus 7.47 with the fixed text encoder. The conditional gain of a representation thus depends on its partner (Appendix B).

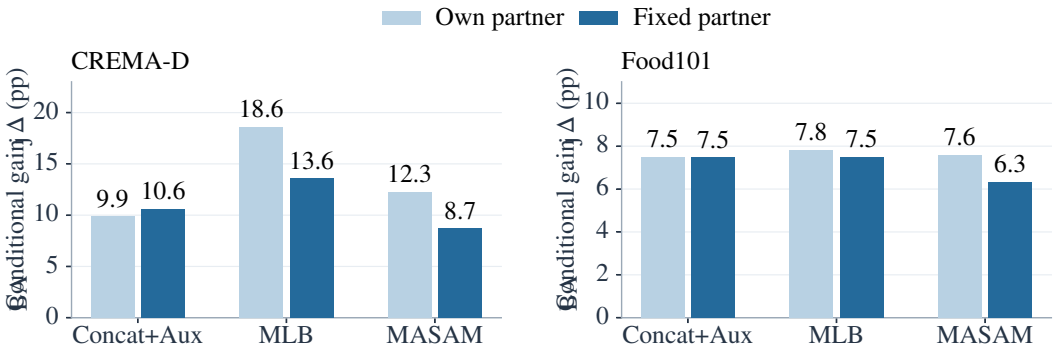

Figure C3: Conditional gain with native and fixed partners.

Prediction collapse. Figure C4 compares confusion matrices with complete and single-modality inputs on the same CREMA-D test clips. The audio failure of MLB and the video failure of MASAM come with predictions concentrated on a single class: MASAM assigns 99.15% of video-only clips to Disgust, and MLB assigns 77.82% of audio-only clips to Angry. By comparison, the most frequent video-only prediction of AUG accounts for 31.54% of clips.

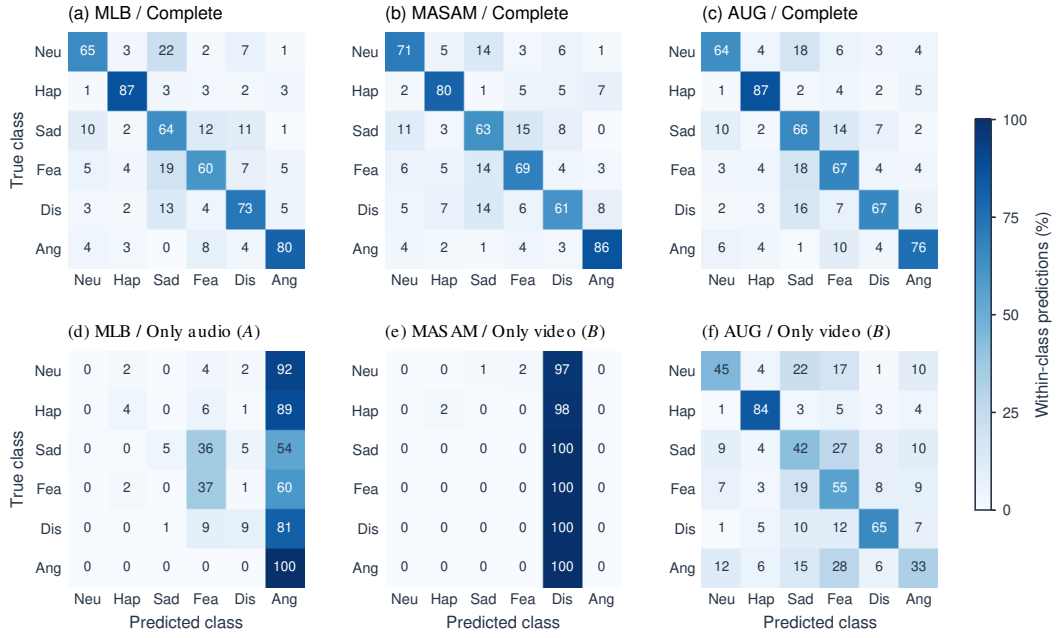

Figure C4: CREMA-D confusion matrices with complete inputs (top) and one modality (bottom).

Classifier refitting. Table C3 gives the full comparison of h_{full} and h_{mix} on CREMA-D and Food101. The two classifiers share the frozen features, architecture, initialization, and validation criterion, and differ only in whether single-modality inputs are used for training. On Food101, refitting with single-modality inputs helps more for images than for text.

Table C3: Classifier refitting on frozen representations.

		Only A		**Only** B	
Dataset	**Source**	h_{full}	h_{mix}	h_{full}	h_{mix}
CREMA-D	Concat+Aux	54.12 ±3.92	57.62 ±1.70	41.40 ±5.77	45.43 ±2.10
	MLB	31.18 ±2.33	51.57 ±0.66	24.24 ±5.28	52.06 ±1.60
	AUG	50.72 ±2.70	52.42 ±2.74	48.03 ±1.67	57.08 ±0.82
	MASAM	40.46 ±2.56	56.23 ±3.10	19.80 ±4.27	35.44 ±0.34
Food101	Concat+Aux	84.87 ±0.17	85.00 ±0.21	63.88 ±1.12	66.91 ±0.64
	MLB	84.53 ±0.43	84.59 ±0.41	63.63 ±1.57	66.58 ±1.01
	AUG	84.93 ±0.02	85.09 ±0.06	64.50 ±0.42	67.16 ±0.17
	MASAM	83.31 ±0.22	83.85 ±0.10	51.06 ±0.36	55.06 ±0.08

Recovery results. Table C4 extends Table 3 with standard deviations and with LogitCal, which fits an affine map of the original missing-input logits. On CREMA-D, LogitCal raises MLB's audio-only accuracy from 23.70% to 49.15% and its video-only accuracy from 30.47% to 49.64%. For MASAM on Food101 images, PEA-SSF reaches 59.79%, compared with 56.91% for GRR and 63.15% for ModDrop. On MIMIC, the largest gain over GRR is 0.36 AUPRC points, from PEA-SSF on AUG's CXR branch. As in the main results, recovery depends on the source model and the missing direction.

Table C4: Full recovery results.

Only A						
Dataset	**Source**	Original	LogitCal	GRR	PEA-SSF	ModDrop
CREMA-D	Concat+Aux	56.18 ±1.28	57.48 ±1.73	57.30 ±1.89	56.68 ±1.58	57.12 ±1.42
	MLB	23.70 ±1.53	49.15 ±1.88	51.66 ±2.32	51.97 ±1.35	52.20 ±1.55
	AUG	51.75 ±3.81	52.91 ±2.42	52.96 ±1.58	53.54 ±1.96	52.78 ±1.04
	MASAM	43.91 ±3.65	56.45 ±0.47	56.32 ±1.17	58.38 ±0.21	59.01 ±1.90
Food101	Concat+Aux	84.49 ±0.22	84.99 ±0.13	85.02 ±0.12	84.93 ±0.15	84.50 ±0.22
	MLB	84.02 ±0.42	84.64 ±0.35	84.79 ±0.26	84.55 ±0.31	84.09 ±0.36
	AUG	84.80 ±0.02	85.01 ±0.03	85.19 ±0.05	85.02 ±0.10	84.80 ±0.02
	MASAM	83.51 ±0.16	84.03 ±0.12	84.05 ±0.13	84.17 ±0.29	84.64 ±0.12
MIMIC	Concat+Aux	46.52 ±0.23	46.97 ±0.29	47.81 ±0.21	46.52 ±0.23	47.21 ±0.12
	MLB	46.42 ±0.29	46.73 ±0.30	47.78 ±0.06	46.42 ±0.29	47.03 ±0.03
	AUG	47.71 ±0.08	47.59 ±0.11	47.86 ±0.09	47.88 ±0.05	47.71 ±0.08
	MASAM	46.39 ±0.25	46.64 ±0.31	47.49 ±0.32	46.39 ±0.25	46.74 ±0.16
Only B						
Dataset	**Source**	Original	LogitCal	GRR	PEA-SSF	ModDrop
CREMA-D	Concat+Aux	27.02 ±2.91	40.99 ±2.97	46.37 ±1.40	41.53 ±2.87	48.66 ±3.62
	MLB	30.47 ±1.35	49.64 ±2.31	50.54 ±1.15	49.73 ±2.97	52.33 ±1.62
	AUG	54.53 ±2.48	57.62 ±1.25	59.27 ±1.76	60.57 ±2.02	60.35 ±1.40
	MASAM	17.70 ±0.21	36.02 ±0.48	36.65 ±0.43	34.86 ±1.69	46.06 ±1.45
Food101	Concat+Aux	63.45 ±0.79	67.21 ±0.50	67.45 ±0.22	67.50 ±0.07	63.45 ±0.79
	MLB	63.11 ±0.32	66.93 ±0.65	67.31 ±0.56	67.59 ±0.29	63.11 ±0.32
	AUG	66.24 ±0.29	67.03 ±0.16	67.98 ±0.21	68.56 ±0.11	66.24 ±0.29
	MASAM	51.27 ±0.52	55.82 ±0.36	56.91 ±0.29	59.79 ±0.66	63.15 ±0.32
MIMIC	Concat+Aux	39.28 ±0.62	39.88 ±0.42	40.05 ±0.33	39.86 ±0.16	39.55 ±0.33
	MLB	39.30 ±0.13	39.91 ±0.06	40.09 ±0.08	39.79 ±0.03	39.41 ±0.23
	AUG	40.39 ±0.08	40.45 ±0.09	40.39 ±0.08	40.75 ±0.09	40.39 ±0.08
	MASAM	38.78 ±0.13	39.39 ±0.05	39.35 ±0.29	39.44 ±0.05	39.29 ±0.16