# BEYOND PID SYNERGY: FUSION VALUE AND PAIRING VALUE FOR MULTIMODAL BAYES RISK

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Multimodal fusion is commonly motivated by partial information decomposition (PID) synergy, and a growing diagnostic literature treats this atom as the reason to combine two views. In this work, we show that this identification does not hold at Bayes optimality. The logarithmic loss gap between the better unimodal Bayes predictor and the joint Bayes predictor is not synergy alone, since it also carries the weaker unique atom. We call this gap fusion value. A second gap separates early fusion from the stacking of two calibrated unimodal posteriors, and this gap is not a linear function of any Williams and Beer quadruple. We call it pairing value. Pairing value lies between zero and fusion value, with AND attaining the lower bound under a strictly positive synergy atom and XOR attaining the upper bound with one bit of both. Adding a named view and deleting it produce the same Bayes gap, so the empirical missing modality drop reads the same quantity under two names. For trained models, the observed fusion gain decomposes into fusion value plus a signed excess risk remainder. An operational protocol built from held out logarithmic loss keeps this remainder near zero whenever a single view already saturates the label. On language saturated sentiment, the remainder is near zero, because the better singleton already captures the label. On caption matching, both views are necessary and late fusion stays at chance, since the label lives only in the pair. The two indices therefore separate the decision to retain a weak view from the decision to interact early.

## 1 INTRODUCTION

A recurring design choice in multimodal learning is whether to fuse two views or trust the stronger one. Image and caption pairs often describe the same fact, and a contrastive objective recovers this shared content from paired samples (Radford et al., 2021). Other signals live in only one view, and some labels emerge only from the combination, as in sarcasm where neither view is informative alone (Yang et al., 2025). Partial information decomposition (PID) was introduced to formalize this distinction, splitting the joint information into redundancy, unique information, and synergy (Williams & Beer, 2010; Bertschinger et al., 2014). It has since become the standard diagnostic vocabulary for multimodal interaction, carried from datasets to trained models, to sample level estimates, and to large vision language models (Liang et al., 2023a; Zhao et al., 2025; Xiu et al., 2026; Fang et al., 2026). Throughout this literature, synergy is read as the reason to fuse.

For Bayes risk under logarithmic loss, this reading is not an identity. The Bayes predictor incurs conditional entropy, so the gap from the better unimodal predictor to the joint predictor is a difference of mutual informations (Cover & Thomas, 2006). We call this gap fusion value, measuring the accuracy lost by discarding the weaker view. Every Williams and Beer decomposition writes fusion value as synergy plus the weaker unique atom, so identifying it with synergy alone holds only when the weaker atom vanishes. The cancellation is exact for a univariate Gaussian target under Barrett's collapse to minimum mutual information, and when one channel is a Blackwell degradation of the other (Barrett, 2015; Raginsky, 2011; Bertschinger et al., 2014). A weaker view that carries unique information of its own still contributes that atom to the joint gap, which the folklore overlooks.

Keeping the weaker view is a different question from letting the two views interact. Let $\eta_i$ denote the unimodal posterior $P(Y \mid X_i)$. Pairing value is the logarithmic loss gap between early Bayes fusion and stacking $(\eta_1, \eta_2)$, measuring the extra accuracy of early fusion over two calibrated experts. The



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.

definition does not depend on a choice of PID measure. On uniform independent bits, XOR has one bit of pairing value and one bit of synergy in every Williams and Beer decomposition, whereas AND has pairing value zero and about half a bit of BROJA synergy (Bertschinger et al., 2014). Calibrated stacking is Bayes optimal for AND and at chance for XOR. Multiplicative fusion was introduced for sentiment precisely because the unimodal scores need not be jointly sufficient for the label (Zadeh et al., 2017), so architecture choice follows pairing value rather than the synergy atom alone.

The same comparison fixes the cost of a missing view. For any loss, the Bayes gain from adding a view equals the drop from deleting it, so the missing modality drop reads one quantity under two names. Experiments relate this drop to unique information and then to synergy once uniqueness is small, both measurements of the same gap (Liang et al., 2023a). Redundancy lower bounds what remains when an unknown view is kept, and we call this bound the residual floor. The drop also contains synergy, which stays large while redundancy is large. Joint networks improve only modestly over the stronger unimodal network when one modality dominates training (Wang et al., 2020; Huang et al., 2022; Peng et al., 2022; Wei et al., 2024), and late fusion prolongs this phase in deep linear networks (Zhang et al., 2024). Fusion value asks whether the gap is present, and pairing value whether late fusion can close it. Contrastive learning sees only the pair of views and therefore collects redundancy while missing a synergistic bit (Tosh et al., 2021; Radford et al., 2021; Liang et al., 2023b; Dufumier et al., 2025).

The chain rule accounts for fusion value, the equality of adding and deleting a view, and the residual floor. Pairing value is not settled by that rule, nor a linear function of a Williams and Beer quadruple, since no coefficients match a redundant copy, AND, and XOR together. We name both gaps in Shannon language and translate them into PID atoms only afterward. For trained predictors, the observed fusion gain equals fusion value plus unimodal excess risk minus joint excess risk. On language saturated sentiment, a coarse sketch of the joint can inflate fusion value, whereas calibrated unimodal heads recover a near zero remainder because the better view already carries the label (Yang et al., 2025). On caption matching, each view alone is uninformative and late fusion stays at chance, because the label lives only in the pair. The two indices therefore separate keeping a weak view from interacting early.

The main contributions of this work are summarized as follows.

- Fusion gain is not synergy. Under Bayes logarithmic loss, the gap from the better singleton to the joint predictor is fusion value, $\mathrm{FVI} = S + \min(U_1, U_2)$. Adding a view and deleting it are the same gap, and redundancy lower bounds what remains.
- Architecture is not BROJA synergy. Pairing value $\mathrm{PVI}$ is the logarithmic loss gap that stacking calibrated unimodal posteriors cannot close. AND and XOR attain the ends of $0 \leq \mathrm{PVI} \leq \mathrm{FVI}$, and pairing value is not a linear function of $(R, U_1, U_2, S)$.
- Empirical gain is not the Bayes index. For trained unimodal and joint predictors, the observed fusion gain equals $\mathrm{FVI}$ plus unimodal excess risk minus joint excess risk. Held out logarithmic loss recovers this comparison, whereas a coarse sketch need not.

## 2 Related work

**Partial information decomposition.** Williams and Beer introduced the redundancy lattice and the minimum specific information $I_{\min}$, which splits the joint mutual information into redundancy, unique information, and synergy (Williams & Beer, 2010). Harder et al. then showed that $I_{\min}$ assigns spurious redundancy to independent sources, which motivated operational definitions of unique information through pairwise couplings and Blackwell orderings (Harder et al., 2013; Bertschinger et al., 2014; Banerjee et al., 2018). Barrett proved that every pairwise PID collapses to minimum mutual information for a univariate Gaussian target, so the weaker unique atom vanishes in that case (Barrett, 2015). The same decomposition is now estimated for datasets, trained models, and contrastive objectives (Liang et al., 2023a; Yang et al., 2025; Liang et al., 2023b; Dufumier et al., 2025). No community consensus has emerged on the correct redundancy measure (Liang et al., 2023a), and we therefore use the lattice only as a translation. Every numbered statement in the main text is stated first in Shannon language.

**Multimodal fusion and its optimization.** More modalities can improve latent estimation when the views carry complementary information (Huang et al., 2021), but joint training often loses to the

best unimodal network through differential overfitting and modality competition (Wang et al., 2020; Huang et al., 2022; Peng et al., 2022; Wei et al., 2024). Deep linear networks prolong a unimodal phase under late fusion, which separates the optimization dynamics from the information theoretic gap (Zhang et al., 2024). Multi view analyses instead assume near redundancy and bound the gain from a second view by a conditional mutual information $\Gamma_{X_2} = I(X_2; Y \mid X_1)$ (Sridharan & Kakade, 2008; Tosh et al., 2021), and contrastive learning collects exactly this redundant part while remaining blind to a synergistic bit (Tosh et al., 2021; Radford et al., 2021; Liang et al., 2023b; Dufumier et al., 2025). On the architecture side, score fusion, evidential fusion, and mixture of experts stacking assume that two unimodal posteriors are jointly sufficient for the label, whereas tensor fusion was introduced because they need not be (Zadeh et al., 2017; Baltrušaitis et al., 2018; Lou et al., 2026). Pairing value is the logarithmic loss gap that distinguishes these two design choices, and cross modal mutual information does not, since the copy world maximizes alignment with pairing value zero while XOR has no alignment and pairing value one bit. PID splits $\Gamma = U + S$, and pairing value is not $\Gamma$. These papers explain why optimization may fail to attain a Bayes gap, and this paper names the gap.

**PID as a multimodal diagnostic.** Liang et al. estimate BROJA at the dataset and model level and select architectures by PID agreement (Liang et al., 2023a). Subsequent work uses PID to schedule fusion (Tang et al., 2025), to profile large vision language models (Xiu et al., 2026; Fang et al., 2026), to attribute representations (Amit et al., 2026), to order a curriculum (Singh et al., 2026), to estimate sample level interactions (Yang et al., 2025), and to compute Gaussian PID through normalizing flows (Zhao et al., 2025). Their measurements motivate the identities here. The missing modality correlations in Liang et al. (2023a) match the expansion $U_{\text{miss}} + S$ of the add and delete dual, and a later literature on large vision language models reports that language often dominates the joint prediction (Xiu et al., 2026; Fang et al., 2026). That observation is the scope corollary in another vocabulary, since the better singleton already captures the label and fusion value is small. Those papers estimate atoms and select architectures, whereas we name the Bayes gaps the atoms are being asked to explain, isolate pairing value as the late fusion remainder, and separate AND from XOR on that remainder.

## 3 Two Shannon indices

Let $(X_1, X_2, Y) \sim P$ with finite alphabets unless stated otherwise. A predictor outputs a distribution on $Y$ under logarithmic loss, or a real value under square loss and zero one loss. Let $\mathcal{R}^\star_{\log}(X) = H(Y \mid X)$ denote the logarithmic loss Bayes risk (Cover & Thomas, 2006), and let $\mathcal{R}^\star_{\text{sq}}$ denote the square loss Bayes risk. All logarithms are natural, so numerical information is measured in nats unless bits are marked explicitly.

**Definition 1** (Williams and Beer PID)**.** *A tuple $(R, U_1, U_2, S) \in \mathbb{R}^4_{\geq 0}$ is a Williams and Beer decomposition of $P$ if $I(X_1; Y) = R + U_1$, $I(X_2; Y) = R + U_2$, and $I(X_1, X_2; Y) = R + U_1 + U_2 + S$.*

The left hand sides of these three identities are Shannon mutual informations, so every such decomposition agrees on the conditional mutual information $I(X_2; Y \mid X_1) = U_2 + S$. The only degree of freedom is the redundancy $R$, which is pinched between coinformation and $\min_i I(X_i; Y)$. Minimum mutual information (MMI) takes the upper end and forces $\min(U_1, U_2) = 0$ (Barrett, 2015). The Bertschinger, Rauh, Olbrich, and Jost (BROJA) measure minimizes unique information over pairwise couplings (Bertschinger et al., 2014).

**Definition 2** (Three Shannon indices)**.** *Fusion value is the extra joint information beyond the better singleton,*

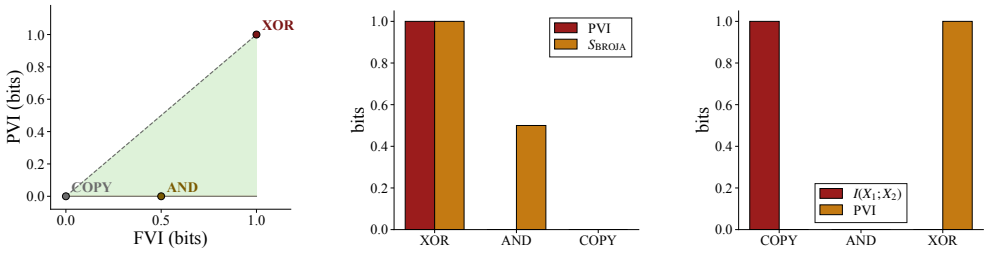
$$\text{FVI}(P) \;=\; I(X_1, X_2; Y) - \max_i I(X_i; Y). \tag{1}$$

*Pairing value is the extra joint information beyond the two unimodal posteriors $\eta_i = P(Y \mid X_i)$,*

$$\text{PVI}(P) \;=\; I(Y; X_1, X_2) - I(Y; \eta_1, \eta_2). \tag{2}$$

*The residual floor $\text{RI}(P) = \min_i I(X_i; Y)$ is the information that remains if an unknown singleton is kept.*

Three canonical worlds anchor the definitions. COPY gives two identical views of the same fact, so $\text{FVI} = \text{PVI} = 0$. AND gives two independent clues whose conjunction determines the label, so




(a) Feasible triangle. (b) PVI versus BROJA. (c) Alignment versus PVI.

Figure 1: COPY sits at the origin, AND on the fusion value axis, and XOR on the diagonal. BROJA synergy matches pairing value on XOR and splits by half a bit on AND. Cross modal mutual information ranks COPY above XOR, the opposite of pairing value.

$\mathrm{FVI} = 0.5$ bit and $\mathrm{PVI} = 0$. XOR gives an interaction visible only in the pair, so $\mathrm{FVI} = \mathrm{PVI} = 1$ bit. Figure 1 places the three worlds in the feasible triangle $0 \leq \mathrm{PVI} \leq \mathrm{FVI}$. Hard label voting is a coarser late fusion class than posterior stacking. If deployed unimodal models are miscalibrated, the measured pairing value is an upper bound on the calibrated gap, and the operational protocol of Section 4 calibrates before stacking.

### 3.1 Fusion value

**Theorem 1** (Logarithmic loss fusion gain)**.** $\mathcal{R}^\star_{\log}(X_i) - \mathcal{R}^\star_{\log}(X_1, X_2) = I(X_{3-i}; Y \mid X_i)$. *The gap to the better unimodal Bayes predictor equals* FVI. *For every Williams and Beer PID the first gap is* $U_{3-i} + S$ *and* $\mathrm{FVI} = S + \min(U_1, U_2)$.

The intuition behind Theorem 1 is the chain rule, not a new information measure. Conditioning on the stronger view already spends the redundancy and the stronger unique atom, leaving the weaker unique atom together with synergy. Folklore keeps only the second term and identifies the fusion gap with synergy alone. The identity shows that this identification is exact only when the weaker unique atom vanishes. A related simplification occurs on the pairing value side, where the weaker posterior is a garbling of the stronger posterior as under Blackwell nesting. Data processing then gives $I(Y; \eta_1, \eta_2) = I(Y; X_{\text{strong}})$ and therefore $\mathrm{PVI} = 0$.

**Corollary 1** (When folklore is exact)**.** *If* $\min(U_1, U_2) = 0$*, in particular under MMI, Barrett's univariate Gaussian PID, or BROJA uniqueness vanishing by Blackwell domination, then* $\mathrm{FVI} = S$.

**Theorem 2** (Gaussian square loss)**.** *If* $(Y, X_1, X_2)$ *is jointly Gaussian and* $Y$ *is scalar, with* $\sigma^2 = \mathrm{Var}(Y)$*, then*

$$\mathcal{R}^\star_{\text{sq}}(X_i) - \mathcal{R}^\star_{\text{sq}}(X_1, X_2) = \sigma^2 e^{-2I(X_i;Y)} \left(1 - e^{-2I(X_{3-i};Y|X_i)}\right). \tag{3}$$

*Under MMI this is a monotone function of* $S_{\text{MMI}}$ *alone, with first order coefficient* $2\sigma^2 e^{-2I_{\max}}$.

Non Gaussian square loss is residual variance of $\mathbb{E}[Y \mid X_1, X_2]$ given $X_i$, and this quantity is not a function of $S$ alone. Corollary 4 in the appendix shows that the Bayes error gap is a monotone function of the same conditional mutual information, and that this gap is not fixed by $S$ alone (Sason & Verdú, 2017).

The refinement matters for design. Huang et al. prove that a larger modality set estimates a latent more accurately (Huang et al., 2021), and Theorem 1 is that monotonicity refined to which atom pays the gap. A second view that is Blackwell weaker than the first pays nothing under uniqueness vanishing, and fusion value collapses to synergy. A second view that is an independent weak clue pays the unique atom even when pairing value is zero. Treating those two second views as the same synergistic resource is the first of the two errors named in the introduction.

### 3.2 The add and delete dual

**Theorem 3** (Add and delete dual and residual floor)**.** *For any loss, the Bayes gain of adding modality* $j$ *equals the Bayes drop of deleting* $j$ *and falling back to the remaining unimodal Bayes predictor. Under logarithmic loss both equal* $I(X_j; Y \mid X_{3-j}) = U_j + S$*. The worst case residual is* $\mathrm{RI} = R + \min(U_1, U_2)$*, hence* $\mathrm{RI} \geq R$*, with equality iff* $\min U = 0$.

Redundancy is insurance that both singletons still know something about $Y$, but it does not cap the drop. The drop is $U_{\text{miss}} + S$, which can be large while $R$ is large, provided $S$ is also large. The inequality that the drop is at most a function of $R$ therefore points the wrong way. The residual floor RI lower bounds what remains when an unknown view is kept, and it exceeds $R$ by the smaller unique atom. This $R$ is not Gács and Körner common information (Yu & Tan, 2020), which concerns the pair $(X_1, X_2)$ rather than $Y$. The two notions are distinct, and only the Shannon floor enters the identities below.

**Proposition 1** (Excess drop and imputation)**.** *On inputs $(X_1, T(X_1))$ any fusion model is at least as bad as unimodal Bayes on $X_1$. Data processing gives $I(Y; X_1, T(X_1)) \leq I(Y; X_1)$. Imputation cannot restore synergy.*

A model that represents only the synergistic circuit and discards unimodal features pays excess drop. The optimal missing modality logarithmic loss is $P(Y \mid X_1)$, not an impute then reevaluate map, unless the fused head internally marginalizes over the missing view. Mean fill and zero fill are common fallbacks, and both implement a kernel $T$ of the remaining view. Proposition 1 then says that the fallback cannot beat unimodal Bayes. On XOR that statement is sharp, because no function of one bit recovers the parity. On COPY the statement is vacuous, because there was never a synergistic remainder to restore. Empirical missingness tables that mix these two worlds therefore mix a hard information theoretic drop with a modeling remainder, and the mixture is not a property of redundancy alone.

### 3.3 Pairing value

Each unimodal posterior $\eta_i$ is sufficient for $Y$ given $X_i$. A pair of unimodal sufficient statistics need not be jointly sufficient for $Y$ given the pair of views, and that distinction is exactly pairing value.

**Theorem 4** (Pairing value)**.** $\text{PVI} = H(Y \mid \eta_1, \eta_2) - H(Y \mid X_1, X_2)$ *is the logarithmic loss gap between posterior late fusion and early fusion. It is nonnegative, and* $\text{PVI} = 0$ *iff $(\eta_1, \eta_2)$ is sufficient for $Y$ given $(X_1, X_2)$. Moreover* $0 \leq \text{PVI} \leq \text{FVI} = S + \min(U_1, U_2)$ *for every Williams and Beer PID.*

Both comparisons in Theorem 4 are data processing inequalities. Each posterior $\eta_i$ is a function of its own view, so the pair $(\eta_1, \eta_2)$ is a function of $(X_1, X_2)$ and $I(Y; \eta_1, \eta_2) \leq I(Y; X_1, X_2)$. Equality holds precisely when $(\eta_1, \eta_2)$ is sufficient for $Y$ given $(X_1, X_2)$. Because each posterior is already sufficient on its own view, $I(Y; \eta_1, \eta_2) \geq \max_i I(X_i; Y)$, and therefore $\text{PVI} \leq \text{FVI}$. Constant posteriors meet the upper bound, and so does the Blackwell garbling above. AND meets the lower bound while FVI stays strictly positive, so the sandwich is not an identity.

**Theorem 5** (AND and XOR separation)**.** *Let $X_1, X_2$ be independent and identically distributed (i.i.d.) uniform bits. If $Y = X_1 \oplus X_2$ then $\eta_i$ is constant,* $\text{PVI} = 1$ *bit, and every Williams and Beer PID has $R = U_1 = U_2 = 0$ and $S = 1$ bit. If $Y = X_1 \wedge X_2$ then $\eta_i$ encodes $X_i$ invertibly, hence* $\text{PVI} = 0$*, while BROJA and $I_{\min}$ and MMI assign $S \approx 0.5$ bit and $R \approx 0.311$ bit.*

**Lemma 1** (Pairing value is not a linear atom)**.** *There is no linear function of a Williams and Beer quadruple $(R, U_1, U_2, S)$ that equals* PVI *on COPY, AND, and XOR simultaneously. In particular* $\text{PVI} = S$ *and* $\text{PVI} = \text{FVI}$ *are both excluded.*

Suppose $\psi = aR + b_1U_1 + b_2U_2 + cS$ agrees with PVI on COPY, AND, and XOR. On COPY the quadruple is $(1, 0, 0, 0)$ bit and the pairing value is zero, which forces $a = 0$. On XOR the quadruple is $(0, 0, 0, 1)$ bit and the pairing value is one bit, which forces $c = 1$. On AND, MMI, $I_{\min}$, and BROJA all agree that $S \approx 0.5$ bit, $U_1 = U_2 = 0$, and $R \approx 0.311$ bit, while the pairing value is zero, so these coefficients return $\psi \approx 0.5$ bit. The three quadruples are distinct, so no linear function of $(R, U_1, U_2, S)$ can equal pairing value on all three joints. Figure 1b records the separation. On AND the calibrated scores determine the label, whereas on XOR the label is available only from the pair.

**Corollary 2** (Architecture)**.** *If* $\text{PVI} = 0$*, no early interaction layer can beat stacking two calibrated unimodal posteriors in logarithmic loss. If* $\text{PVI} > 0$*, early fusion has a strict information theoretic gap.*

Liang et al. observe that tensor fusion and transformers capture more synergy than additive heads (Liang et al., 2023a). That observation is the $\text{PVI} > 0$ regime, not a generic property of BROJA synergy.

The two indices answer two different product questions, and the data those questions require differ accordingly. Unpaired samples $\{(x_i, y)\}$ identify RI and a Shannon upper bound on FVI, but they do not identify PVI, which requires the same $y$ to be observed with both posteriors. Three product questions therefore use three numbers. Discarding the weaker view costs FVI. Collecting paired $(x_1, x_2, y)$ and training an early fusion head, relative to stacking two calibrated unimodal experts, is worth PVI. Keeping a second unpaired unimodal expert and stacking posteriors captures everything except PVI. The three questions are not interchangeable. A dataset with large fusion value and vanishing pairing value is a world in which a second view is worth keeping and an interaction layer is not. A dataset with pairing value equal to fusion value is a world in which stacking calibrated scores wastes the entire joint gap. Diagnostic tables that report a single synergy number cannot tell those two worlds apart.

Information theoretic availability is not statistical query learnability (Feldman, 2017; Blum et al., 2003). Cross modal $k$ parity embeds the classical orthogonal parity family into a split alphabet. On every member, $R = U_i = 0$ and $S = \mathrm{PVI} = 1$ bit, so Bayes early fusion has zero error while any unimodal or posterior late predictor is at chance. We do not claim a new statistical query bound. The only statement we use is that any objective which is a functional of $P_{X_1,X_2}$ alone is independent of the label. Contrastive pairing therefore cannot see a pure synergistic bit (Tosh et al., 2021; Radford et al., 2021). AND and XOR are indistinguishable from $P_{X_1,X_2}$ and yet opposite for pairing value (Figure 1c).

## 4 From Bayes identities to trained models

Theorems 1 to 4 are identities for Bayes predictors. A deployed system uses finite function classes, and the identities survive only up to the excess risks that those classes carry. The next three lemmas form the layer that connects Bayes theory to trained models.

**Assumption 1** (Bayes identities versus empirical risk minimization (ERM) remainders)**.** *A function class of Rademacher complexity $\mathfrak{R}_n$ and approximation error $\varepsilon$ adds $O(\mathfrak{R}_n + \varepsilon + \sqrt{\log(1/\delta)/n})$ to each risk (Bartlett & Mendelson, 2002; Xu & Raginsky, 2017; Bartlett et al., 2021).*

**Lemma 2** (Excess risk transfer)**.** *Let $f_i$ be any predictor of $Y$ from $X_i$ and let $f_{12}$ be any predictor from $(X_1, X_2)$. Let $G = \min_i \mathcal{R}_{\log}(f_i) - \mathcal{R}_{\log}(f_{12})$ denote the empirical fusion gain. Then*

$$G \;=\; \mathrm{FVI}(P) + \big(\min_i \mathcal{R}_{\log}(f_i) - \min_i \mathcal{R}^\star_{\log}(X_i)\big) - \big(\mathcal{R}_{\log}(f_{12}) - \mathcal{R}^\star_{\log}(X_1, X_2)\big). \tag{4}$$

*In particular $|G - \mathrm{FVI}(P)|$ is at most the sum of the two excess risks.*

The identity adds and subtracts the two Bayes risks $\min_i \mathcal{R}^\star_{\log}(X_i)$ and $\mathcal{R}^\star_{\log}(X_1, X_2)$. The first difference is $\mathrm{FVI}(P)$ by Theorem 1. The other two differences are the unimodal excess and the joint excess, and both are nonnegative. Lemma 2 is the reason a small fusion gain does not refute the Bayes identities. If the better unimodal head is already near Bayes, then $G$ cannot exceed FVI by more than the joint excess. If the joint head is underfit, then $G$ underestimates FVI. The same arithmetic yields a scope statement.

**Corollary 3** (Scope)**.** *If $\mathrm{FVI}(P) \le \varepsilon$ then $G \le \varepsilon$ plus unimodal excess risk. Saturated worlds, including copy worlds and language dominated sentiment, have no ranking object for any predictor of architecture gain.*

**Lemma 3** (Coarsening upper bound)**.** *If each $g_i$ is a function of $\eta_i$, then $H(Y \mid g_1, g_2) - H(Y \mid X_1, X_2) \ge \mathrm{PVI}(P)$. Temperature scaling, binning, and Platt calibration are of this form.*

An uncalibrated unimodal head need not be a function of $\eta_i$, and in that case the gap can move in either direction. The operational protocol therefore calibrates on held out data before stacking.

**Lemma 4** (Three gaps)**.** *Let $\mathrm{PVI}_{\hat\eta} = H(Y \mid \hat\eta_1, \hat\eta_2) - H(Y \mid X_1, X_2)$ and let $\mathrm{PVI}_{\mathrm{alg}} = \mathcal{R}_{\log}(\mathrm{late}) - \mathcal{R}_{\log}(f_{12})$. Then $\mathrm{PVI}_{\mathrm{alg}} = \mathrm{PVI}_{\hat\eta}$ plus the late combiner excess minus the early excess. A coarse sketch of $(X_1, X_2, Y)$ is a fourth quantity and is not equal to any of the three.*

The operational protocol follows the definition rather than a low dimensional histogram. One trains two unimodal heads and a joint head. On held out data, $\widehat{\mathrm{FVI}}_{\mathrm{op}}$ is the logarithmic loss gap between the better unimodal head and the joint head, and $\widehat{\mathrm{PVI}}_{\mathrm{op}}$ is the gap between a calibrated stack of

the two unimodal probabilities and the joint head. On a finite alphabet the plug in deviation of FVI follows the usual entropy rate in the joint alphabet $|\mathcal{X}_1||\mathcal{X}_2||\mathcal{Y}|$ (Paninski, 2003; Antos & Kontoyiannis, 2001). In high dimension that alphabet is large, and the operational remainder is the quantity that Lemma 2 licenses. A principal component sketch with a handful of bins is a convenient diagnostic, but it is not a consistent estimator of $\mathrm{FVI}(P)$ once the views are BERT or CLIP vectors. The same warning applies to pairing value. A quantized pair can hide the interaction that a trained early head extracts, and it can invent a remainder that calibrated unimodal heads already explain. Section 5 therefore reports both instruments and treats agreement with held out risk as the test of the instrument.

## 5 Empirical evidence

The empirical study follows a single chain rather than a catalogue of tasks. Enumerated joints test the identities. Pixels and frozen features test whether the same two numbers separate copy, unique, AND, and XOR once the alphabet grows beyond a pair of bits. Trained heads test the excess risk lemma. Caption matching tests pairing value on a label that is not a Boolean of two constructed bits. Sentiment tests the scope corollary on the benchmarks that a diagnostic paper is expected to report. Atomic worlds use a fixed seed, and gaps on trained heads are held out. Proofs and protocol details occupy the appendix.

### 5.1 Atomic worlds

Figure 2 overlays Bayes logarithmic loss gain on $S + \min U$ for seven canonical worlds. MMI, $I_{\min}$, and BROJA coincide on the diagonal because the translation $\mathrm{FVI} = S + \min U$ is a Shannon identity, and the maximum absolute error is $4.4 \times 10^{-16}$ nats. A sixty six point mixture grid has MMI translation error identically zero. On Dirichlet random joints with alphabets up to $5 \times 3 \times 2$, MMI error is identically zero and $0 \le \mathrm{PVI} \le \mathrm{FVI}$ in every draw. As shown in Figure 1b, an additive logistic head has test error $0.50$ on XOR and $0$ on AND, while the same model with a multiplicative coordinate reaches error $0$ on both. Label flip noise on XOR keeps $\mathrm{PVI} = \mathrm{FVI} = S_{\mathrm{BROJA}}$ down to chance. On AND, pairing value stays $0$ while BROJA synergy tracks fusion value. The two indices therefore separate the two canonical interactions that a single atom cannot. Bayes accuracy drop on deleting $X_2$ is $0.497$ on XOR, $0.254$ on AND, and $0$ on COPY. Hard imputation of $X_2$ from $P(X_2 \mid X_1)$ returns XOR to chance and leaves COPY at $1.00$, as Proposition 1 requires.

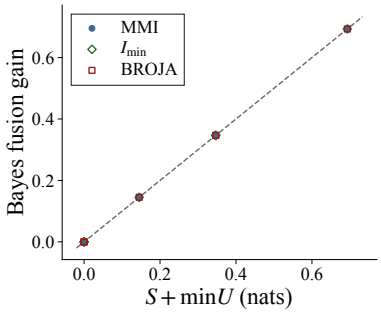



Figure 2: Bayes fusion gain versus $S + \min U$ on seven atomic worlds. MMI, $I_{\min}$, and BROJA coincide on the diagonal.

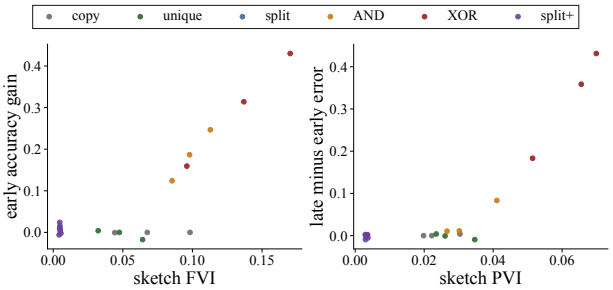
On Barrett's scalar construction every pairwise PID collapses to MMI, so $\min U = 0$ and $\mathrm{FVI} = S_{\mathrm{MMI}}$ on a six by six SNR grid. Sweeping observation noise correlation $\rho$ moves fusion value and the residual floor in opposite directions. Scalar asymmetric sources have linear MMSE pairing value $0$, while vector sources have linear MMSE pairing value $0.167$ across a noise grid because a direction that is marginally silent about $Y$ becomes useful together with the other view. Degree $2$ cross modal XOR is learned by a width $64$ MLP at $n = 200$, and linear models remain at one half. For $k \ge 3$ and $n \le 4000$ the MLP error stays in $[0.38, 0.53]$. Across mixtures, Spearman rank correlation of FVI with true Bayes gain is $1.00$, while the same correlation for $I(X_1; X_2)$ is $-0.12$ and for the first canonical correlation is $-0.14$. Pairing value ranks the late versus early logarithmic gap at $1.00$ on the enumerated table.

### 5.2 Trained fusion heads

A $14 \times 14$ MNIST canvas realizes the same trichotomy on pixels. Copy uses a noisy clone, unique uses an independent digit, and split uses left versus right half of one digit. AND pair and XOR pair take two independent digits and label the Boolean of their even or odd bits. A two dimensional PCA sketch with four quantile bins ranks early versus best unimodal accuracy gain at Spearman $0.77$,

while $I(X_1; X_2)$ anti ranks at $-0.14$. XOR pair is the regime in which both unimodal heads sit at chance while early fusion still gains, and the same coarse sketch reports pairing value near zero on that pair. Lemma 4 predicts the discrepancy, and refining the sketch restores a strictly positive pairing value that shares the sign of the observed gap. Frozen CLIP ViT-B/32 features of CIFAR 10 image and class prompt pairs recover the same trichotomy (Radford et al., 2021). Copy and unique gain 0.000 and 0.001, AND gains 0.216 which score fusion also reaches, and XOR gains 0.366 from unimodal heads at chance while score fusion stays at chance. Across thirty one versus rest tasks, a binned fusion value ranks gain at 0.74 while $I(X_1; X_2)$ anti ranks at $-0.39$.

Training the fusion heads instead of reading a frozen probe makes the excess risk visible. As shown in Figure 3, on twenty CLIP regimes that mix copy, unique, split, AND, and XOR with three label noise levels, Spearman correlation of a sketch FVI with early accuracy gain is 0.45 on the full grid and 0.95 on the nine interaction worlds. The corresponding pairing value correlations are 0.62 and 0.98, while the first canonical correlation anti ranks the same gain at $-0.26$ and $-0.73$. Copy and unique have operational fusion gain near zero because a single view already solves the label, and Corollary 3 says that those worlds should not be asked to produce a large rank correlation. The interaction subset is the regime in which the two questions have an answer, and the two indices separate it from CCA. The operational protocol of Section 4 then removes the sketch inflation. On the same CLIP bank, operational fusion value is $9 \times 10^{-5}$ nats on copy and $-0.005$ on unique, while the coarse sketch still reports 0.10 and 0.03. On XOR the operational pairing value is 0.65 nats against a sketch value of 0.07, which is Lemma 4 in numbers. The sketch underestimates the joint information, and the trained early head does not.



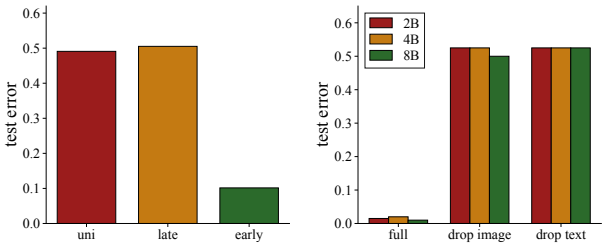

(a) FVI versus early gain (b) PVI versus late gap

Figure 3: Trained CLIP heads on twenty regimes. Sketch FVI and PVI rank the interaction worlds at Spearman 0.95 and 0.98.

### 5.3 Caption Matching

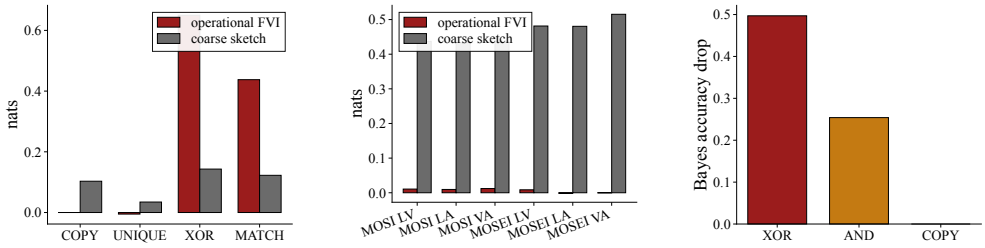
The interaction worlds above still use constructed Boolean labels. Caption matching uses the same CLIP image and class prompt features but sets the label to whether the prompt names the depicted class. By construction each view alone is uninformative, so both unimodal error and late fusion error should sit at one half while early fusion should not. The earlier animal versus object prompt places almost all of the label in the pixels and is therefore a unique vision world rather than a pairing value test, while caption matching is a pairing value test.

Across three seeds the mean unimodal error is 0.491, the mean late error is 0.505, and the mean early error is 0.102. Operational fusion value and pairing value both equal 0.43 nats, which confirms that the two views carry no individual signal and the pair does not. As shown in Figure 4, a LoRA tuned Qwen3-VL instruct family at 2B, 4B, and 8B repeats the test in token space. The prompt asks whether a caption matches the image, and full fine tuning reaches held out error 0.015, 0.020, and 0.010 at the three sizes. Removing the image or removing the caption returns error to chance at every size, between 0.50 and 0.525, with logarithmic loss gaps between 0.65 and 0.70 nats. The add and delete dual says that the two missingness gaps should be large and comparable, which is what the three sizes show. The same dual forbids reading a large




(a) CLIP caption match. (b) Qwen at three sizes.

Figure 4: Caption matching. Unimodal and late fusion sit at chance. Early fusion does not. Qwen repeats the pattern at 2B, 4B, and 8B.




(a) CLIP operational FVI versus a coarse sketch.

(b) MOSI and MOSEI operational FVI versus sketch.

(c) Bayes accuracy drop after deleting one view.

Figure 5: Operational fusion value recovers a near zero remainder on copy, unique, and language saturated sentiment. The coarse sketch does not. Missingness on atomic worlds follows the add and delete dual.

missingness drop as evidence of unique information alone. On this task both views are unique in the Williams and Beer sense and synergistic in the pairing value sense, because neither posterior is informative and the pair is. A single atom cannot carry both readings, and scale does not replace the pair.

### 5.4 Language saturated sentiment

CMU-MOSI and CMU-MOSEI are the natural affect benchmarks on which a diagnostic paper is expected to stand, and they are also the worlds in which Corollary 3 predicts that no architecture predictor has an object. Mean pooled BERT language already yields error $0.19$ on MOSI sign and $0.27$ on MOSEI sign, while vision and audio sit near $0.43$. Operational fusion value on the three MOSI pairs is $0.010$, $0.010$, and $0.012$ nats. On MOSEI, with $22856$ utterances, the three operational values are $0.009$, $-0.002$, and $0.000$. As shown in Figure 5b, the operational remainder agrees with the accuracy table while the sketch disagrees with both. Sentiment therefore confirms the scope corollary rather than serving as a ranking contest among fusion value, BROJA synergy, cross modal mutual information, and CCA, all of which fail to rank a half point residual when FVI is a hundredth of a nat.

On a pixel XOR pair the same operational protocol yields unimodal error $0.50$, early error $0.414$, and operational pairing value $0.022$ nats, of the same sign as a refined sketch. The gap is smaller than the CLIP XOR remainder of $0.65$ nats because the early head is farther from Bayes, which is the sign that Lemma 2 assigns to joint excess. The pixel world is therefore a check of the remainder rather than a second caption matching claim. Taken together, the enumerated identities, the trained CLIP heads, caption matching, and sentiment form one chain. Fusion value is the gap to the better singleton, and pairing value is the extra gap that stacking calibrated posteriors cannot close. The signs of the two numbers decide whether to keep a view and whether to interact early. The operational remainder is the quantity that a trained system is allowed to report, while the sketch is a diagnostic of the sketch. Canonical correlation and cross modal mutual information remain useful as descriptions of alignment, but they are the wrong rankers for either product question because COPY maximises them and XOR minimises them.

## 6 Conclusion

This work introduces two Shannon indices that separate the two design questions behind multimodal fusion. Fusion value is the logarithmic loss gap between the better unimodal Bayes predictor and the joint Bayes predictor, and every Williams and Beer decomposition translates it as synergy plus the weaker unique atom. Pairing value is the gap between early fusion and a stack of two calibrated unimodal posteriors, and it lies between zero and fusion value, with AND at the lower bound and XOR at the upper bound, and no linear function of the four atoms equals pairing value on all three worlds. For trained predictors, the observed fusion gain equals fusion value plus a signed excess risk remainder, and an operational protocol keeps this remainder near zero whenever a single view already saturates the label. The two numbers therefore separate the decision to retain a weak view from the decision to interact early, and they do so in Shannon language that is invariant to the choice of redundancy measure.

## AI USE STATEMENT

In this work, we used generative AI tools for providing feedback on research methodology or experiments, and assisting with translation. We have not used generative AI tools for formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses, implementing methods, supporting qualitative and thematic data analysis, interpreting results. Besides, generating synthetic data sets, helping develop theoretical models or conceptual frameworks, cleaning and reformatting dataset are not applicable to this work. Additionally, we used generative AI tools for modifying scientific figures or images, editing a research paper to improve readability. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

## REFERENCES

Padegal Amit, Omkar Mahesh Kashyap, Namitha Rayasam, Nidhi Shekhar, and Surabhi Narayan. Quantifying modality contributions via disentangling multimodal representations. In *International FLAIRS Conference Proceedings*, volume 39, 2026.

András Antos and Ioannis Kontoyiannis. Convergence properties of functional estimates for discrete distributions. *Random Structures & Algorithms*, 19(3-4):163–193, 2001.

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, 2018.

Pradeep Kr Banerjee, Eckehard Olbrich, Jürgen Jost, and Johannes Rauh. Unique informations and deficiencies. In *56th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 32–38. IEEE, 2018.

Adam B Barrett. Exploration of synergistic and redundant information sharing in static and dynamical gaussian systems. *Physical Review E*, 91(5):052802, 2015.

Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of machine learning research*, 3(Nov):463–482, 2002.

Peter L Bartlett, Andrea Montanari, and Alexander Rakhlin. Deep learning: a statistical viewpoint. *Acta numerica*, 30:87–201, 2021.

Nils Bertschinger, Johannes Rauh, Eckehard Olbrich, Jürgen Jost, and Nihat Ay. Quantifying unique information. *Entropy*, 16(4):2161–2183, 2014.

Avrim Blum, Adam Kalai, and Hal Wasserman. Noise-tolerant learning, the parity problem, and the statistical query model. *Journal of the ACM (JACM)*, 50(4):506–519, 2003.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2 edition, 2006. ISBN 978-0-471-24195-9.

Benoit Dufumier, Javiera Castillo Navarro, Devis Tuia, and Jean-Philippe Thiran. What to align in multimodal contrastive learning? In *International Conference on Learning Representations*, pp. 5408–5432, 2025.

Wanlong Fang, Tianle Zhang, Wen Tao, and Alvin Chan. Towards understanding modality interaction in multimodal language models via partial information decomposition. In *International Conference on Machine Learning*, 2026.

Vitaly Feldman. A general characterization of the statistical query complexity. In *Conference on learning theory*, pp. 785–830, 2017.

Vitaly Feldman, Elena Grigorescu, Lev Reyzin, Santosh S Vempala, and Ying Xiao. Statistical algorithms and a lower bound for detecting planted cliques. *Journal of the ACM (JACM)*, 64(2): 1–37, 2017.

Malte Harder, Christoph Salge, and Daniel Polani. Bivariate measure of redundant information. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 87(1):012130, 2013.

Yu Huang, Chenzhuang Du, Zihui Xue, Xuanyao Chen, Hang Zhao, and Longbo Huang. What makes multi-modal learning better than single (provably). In *Advances in Neural Information Processing Systems*, volume 34, pp. 10944–10956, 2021.

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

Paul Pu Liang, Yun Cheng, Xiang Fan, Chun Kai Ling, Suzanne Nie, Richard Chen, Zihao Deng, Nicholas Allen, Randy Auerbach, Faisal Mahmood, et al. Quantifying & modeling multimodal interactions: An information decomposition framework. In *Advances in Neural Information Processing Systems*, volume 36, pp. 27351–27393, 2023a.

Paul Pu Liang, Zihao Deng, Martin Q Ma, James Y Zou, Louis-Philippe Morency, and Ruslan Salakhutdinov. Factorized contrastive learning: Going beyond multi-view redundancy. In *Advances in Neural Information Processing Systems*, volume 36, pp. 32971–32998, 2023b.

Yiwei Lou, Yuanpeng He, Rongchao Zhang, Yongzhi Cao, Hanpin Wang, and Yu Huang. Multitasks-based deep evidential fusion network for blind image quality assessment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 7512–7520, 2026.

Liam Paninski. Estimation of entropy and mutual information. *Neural computation*, 15(6):1191–1253, 2003.

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 (CVPR)*, pp. 8228–8237. IEEE, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763, 2021.

Maxim Raginsky. Shannon meets blackwell and le cam: Channels, codes, and statistical experiments. In *2011 IEEE International Symposium on Information Theory Proceedings*, pp. 1220–1224. IEEE, 2011.

Igal Sason and Sergio Verdú. Arimoto–rényi conditional entropy and bayesian $m$-ary hypothesis testing. *IEEE Transactions on Information theory*, 64(1):4–25, 2017.

Itamar Shoshani and Ohad Shamir. Hardness of learning fixed parities with neural networks. *arXiv preprint arXiv:2501.00817*, 2025.

Ankush Pratap Singh, Houwei Cao, and Yong Liu. Spice: Synergy and partial information based curriculum evolution. *arXiv preprint arXiv:2606.16639*, 2026.

Karthik Sridharan and Sham M. Kakade. An information theoretic framework for multi-view learning. In *Annual Conference on Computational Learning Theory*, number 114, pp. 403–414, 2008.

Jiaqi Tang, Yinsong Xu, Yang Liu, and Qingchao Chen. Shaping initial state prevents modality competition in multi-modal fusion: A two-stage scheduling framework via fast partial information decomposition. *arXiv preprint arXiv:2509.20840*, 2025.

Christopher Tosh, Akshay Krishnamurthy, and Daniel Hsu. Contrastive learning, multi-view redundancy, and linear models. In *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, pp. 1179–1206, 2021.

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 (CVPR)*, pp. 12692–12702. IEEE, 2020.

Yake Wei, Di Hu, Henghui Du, and Ji-Rong Wen. On-the-fly modulation for balanced multimodal learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(1):469–485, 2024.

Paul L Williams and Randall D Beer. Nonnegative decomposition of multivariate information. *arXiv preprint arXiv:1004.2515*, 2010.

Lixin Xiu, Xufang Luo, and Hideki Nakayama. A comprehensive information-decomposition analysis of large vision-language models. In *International Conference on Learning Representations*, 2026.

Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Zequn Yang, Hongfa Wang, and Di Hu. Efficient quantification of multimodal interaction at sample level. In *International Conference on Machine Learning*, pp. 71302–71317, 2025.

Lei Yu and Vincent Y. F. Tan. On exact and $\infty$-Rényi common informations. *IEEE Transactions on Information Theory*, 66(6):3366–3406, 2020. doi: 10.1109/TIT.2020.2970396.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 1103–1114, 2017. doi: 10.18653/v1/D17-1115.

Yedi Zhang, Peter E. Latham, and Andrew M. Saxe. Understanding unimodal bias in multimodal deep linear networks. In *International Conference on Machine Learning*, volume 235, pp. 59100–59125, 2024.

Wenyuan Zhao, Adithya Balachandran, Chao Tian, and Paul Pu Liang. Partial information decomposition via normalizing flows in latent Gaussian distributions. In *Advances in Neural Information Processing Systems*, volume 38, pp. 4495–4530, 2025.

# APPENDIX

**A** Notation and PID measures . . . . . 14

**B** Shannon preliminaries . . . . . 15

**C** Proofs for fusion value and the dual . . . . . 16

**D** Proofs for pairing value . . . . . 17

**E** Worked examples . . . . . 18

**F** Proofs for finite class remainders . . . . . 20

**G** Contrastive blindness and a statistical query embedding . . . . . 22

**H** Sequential fusion of three modalities . . . . . 23

**I** Comparison with Huang and Liang . . . . . 23

**J** Experimental protocol . . . . . 23

**K** Additional numerical results . . . . . 24

# A NOTATION AND PID MEASURES

Logarithms are natural unless a displayed quantity is converted to bits by the factor $1/\ln 2$. Alphabets are finite except in the Gaussian calculations. A predictor maps an observation to a distribution on $Y$ under logarithmic loss, and to a real number under square loss or zero one loss. Table 1 lists the symbols that appear in the numbered statements.

Table 1: Symbols used in the numbered statements.

| Symbol | Meaning |
|---|---|
| $P$ | joint law of $(X_1, X_2, Y)$ |
| $R, U_i, S$ | Williams and Beer atoms |
| $\eta_i$ | unimodal posterior $P(Y \mid X_i)$ |
| FVI | fusion value $I(X_1, X_2; Y) - \max_i I(X_i; Y)$ |
| PVI | pairing value $I(Y; X_1, X_2) - I(Y; \eta_1, \eta_2)$ |
| $S_\Delta$ | synonym of PVI |
| RI | residual floor $\min_i I(X_i; Y)$ |
| $\Delta_P$ | BROJA couplings that match $P_{X_i,Y}$ |
| $\mathcal{C}_k$ | cross modal $k$ parity family |
| $\mathcal{R}^\star_{\log}(X)$ | logarithmic loss Bayes risk $H(Y \mid X)$ |
| $G$ | trained fusion gain $\min_i \mathcal{R}_{\log}(f_i) - \mathcal{R}_{\log}(f_{12})$ |
| $\widehat{\mathrm{FVI}}_{\mathrm{op}}$ | held out operational fusion value |
| $\widehat{\mathrm{PVI}}_{\mathrm{op}}$ | held out operational pairing value |

## A.1 PID MEASURES

A Williams and Beer decomposition is any nonnegative quadruple that satisfies the three Shannon identities in Section 3. Those identities force the conditional mutual information and the coinformation into the forms

$$I(X_2; Y \mid X_1) = U_2 + S, \tag{5}$$

$$I(X_1; X_2; Y) = R - S, \tag{6}$$

and they force the redundancy into the interval

$$\big[I(X_1; Y) + I(X_2; Y) - I(X_1, X_2; Y)\big]_+ \;\le\; R \;\le\; \min_i I(X_i; Y). \tag{7}$$

Different measures select different points of equation 7. An identity written in Shannon language is therefore unchanged when the redundancy measure changes, and only the numerical split of a reported $R$ moves.

Minimum mutual information sets $R_{\mathrm{MMI}} = \min_i I(X_i; Y)$, which forces $\min(U_1, U_2) = 0$ and $S_{\mathrm{MMI}} = \mathrm{FVI}$ (Barrett, 2015). Barrett proved that every decomposition whose redundancy and unique atoms depend only on the pairwise marginals $P_{X_i,Y}$ collapses to this choice when $Y$ is univariate Gaussian and $(Y, X_1, X_2)$ are jointly Gaussian. Specific information is $I(X; Y = y) = D_{\mathrm{KL}}(P_{X|y} \| P_X)$, where $D_{\mathrm{KL}}$ denotes KL divergence, and the original redundancy is the average $R_{I_{\min}} = \sum_y P(y) \min_i I(X_i; Y = y)$ (Williams & Beer, 2010). Harder et al. showed that this average can assign redundancy to independent sources (Harder et al., 2013). The tables below therefore report $I_{\min}$ only as a check that the Shannon identities survive that choice.

Let $\Delta_P = \{q \mid q(x_i, y) = P(x_i, y)\ \forall i\}$. BROJA unique information is $U_1 = \min_{q \in \Delta_P} I_q(X_1; Y \mid X_2)$ (Bertschinger et al., 2014). The same value is obtained by maximizing $H_q(Y \mid X_1, X_2)$ over $\Delta_P$, because the pairwise marginals pin $H_q(Y)$, and the resulting synergy is $S_{\mathrm{BROJA}} = I_P(X_1, X_2; Y) - I_{q^\star}(X_1, X_2; Y)$. Thus $R$ and $U$ are functions of the pairwise source to target channels, while $S$ is the only atom that sees the coupling $P_{X_1,X_2|Y}$. The paper treats BROJA as one admissible point of equation 7, since no single redundancy measure is accepted as canonical (Liang et al., 2023a). Table 2 records which numbered statements use the redundancy atom.

Table 2: Dependence of each numbered statement on the choice of redundancy.

| Claim | Shannon form | Depends on the choice of $R$ |
|---|---|---|
| Theorem 1 | yes | translation $S + \min U$ only |
| Corollary 4 | yes | no, envelope of the same conditional mutual information |
| Theorem 3 (dual) | yes | no |
| Theorem 3 (floor) | yes | translation $R + \min U$ |
| Proposition 1 | risk | no |
| Theorem 4 | yes | no |
| Theorem 5 | yes | AND uses BROJA numbers |
| Lemma 1 | yes | linear combinations of the quadruple |
| Lemma 2 | risk | no |
| Corollary 3 | risk | no |

# B SHANNON PRELIMINARIES

The proofs use four standard facts, recorded here in the notation of the main text. Equality cases are stated with the identities so that later arguments can cite them directly.

**Lemma 5** (Chain rule and data processing). $I(X_1, X_2; Y) = I(X_1; Y) + I(X_2; Y \mid X_1)$. *If $T = t(X)$ is measurable then $I(T; Y) \le I(X; Y)$, with equality if and only if $Y \perp X \mid T$.*

**Lemma 6** (Logarithmic loss Bayes). $\mathcal{R}_{\log}(q) = \mathbb{E}[-\log q(Y)]$. *The minimizer given $X = x$ is $q = P(Y \mid x)$, and the conditional risk is $H(Y \mid X = x)$. Hence $\mathcal{R}^\star_{\log}(X) = H(Y \mid X)$ and*

$$\mathcal{R}^\star_{\log}(A) - \mathcal{R}^\star_{\log}(A, B) = I(B; Y \mid A).$$

**Lemma 7** (Gaussian MMSE). *If $(Y, X)$ is jointly Gaussian and $Y$ is scalar, then $\mathbb{E}[Y \mid X]$ is linear and $\mathcal{R}^\star_{\mathrm{sq}}(X) = \mathrm{Var}(Y)\, e^{-2I(X;Y)}$ (Cover & Thomas, 2006).*

**Lemma 8** (Error probability and conditional entropy). *Let $P_e = P(\hat{Y} \neq Y)$ for $\hat{Y} = g(X)$. Then $H(Y \mid X) \le h_2(P_e) + P_e \log(|\mathcal{Y}| - 1)$, where $h_2$ denotes binary entropy. For $|\mathcal{Y}| = 2$ one has $H(Y \mid X) \le h_2(P_e)$ and, when entropy is measured in bits, $P^\star_e \le \frac{1}{2} H_2(Y \mid X)$ (Sason & Verdú, 2017). Consequently $P^\star_e(X)$ lies between two functions of $H(Y \mid X)$ that vanish at $0$ and increase in the argument.*

**Lemma 9** (Redundancy interval). *Nonnegativity of $U_i = I(X_i; Y) - R$ gives the upper bound of equation 7. Nonnegativity of $S = I(X_1, X_2; Y) - I(X_1; Y) - I(X_2; Y) + R$ gives the lower bound, or $0$ when the coinformation is negative.*

## B.1 FURTHER LEMMAS

The next statements are used in more than one proof, so they are collected before the arguments that cite them. Each one is either a rewriting of a Shannon identity or a data processing bound.

**Lemma 10** (Coinformation). $I(X_1; X_2; Y) = I(X_1; X_2) - I(X_1; X_2 \mid Y) = R - S$ *for every Williams and Beer decomposition. XOR has coinformation $-1$ bit, COPY has coinformation $+1$ bit, and AND has $R - S \approx -0.189$ bit under BROJA.*

**Lemma 11** (Blackwell nesting). *If there exists a channel $\kappa$ with $P_{X_2|Y} = \kappa \circ P_{X_1|Y}$, then $I(X_2; Y) \le I(X_1; Y)$. Whenever $X_2$ is a possibly randomized function of $X_1$ given $Y$ and that function adds no further information about $Y$, one also has $I(X_1, X_2; Y) = I(X_1; Y)$. BROJA unique information of $X_2$ then vanishes (Bertschinger et al., 2014; Banerjee et al., 2018). In that case $\min(U_1, U_2) = 0$ and $\mathrm{FVI} = S$, and $\mathrm{PVI} = 0$ whenever $\eta_2$ is a garbling of $\eta_1$.*

**Lemma 12** (Hard label late fusion). $I(Y; \arg\max \eta_1, \arg\max \eta_2) \le I(Y; \eta_1, \eta_2) \le I(Y; X_1, X_2)$. *The gap on the left is the cost of discarding posterior confidence, and the gap on the right is* PVI.

**Lemma 13** (Fano constants, binary $Y$). *When entropy is measured in bits, $\frac{1}{2} H_2(Y \mid X) \ge P^\star_e(X) \ge 0$ and $H_2(Y \mid X) \le h_2(P^\star_e)$ (Sason & Verdú, 2017). The zero one fusion gain therefore satisfies*

$$0 \;\le\; P^\star_e(X_i) - P^\star_e(X_1, X_2) \;\le\; \tfrac{1}{2} H_2(Y \mid X_i),$$

*which is $O(\delta)$ with $\delta = I(X_{3-i}; Y \mid X_i)$ when the residual entropy vanishes.*

**Lemma 14** (Plug in union bound). *Let $\mathcal{A}_0 = \{Y\}$, $\mathcal{A}_i = \{(X_i, Y)\}$, and $\mathcal{A}_{12} = \{(X_1, X_2, Y)\}$. Four entropy estimates, each controlled by the plug in deviation bound of Paninski at failure probability $\delta/4$, yield*

$$|\widehat{\mathrm{FVI}} - \mathrm{FVI}| \le 4C\sqrt{\frac{|\mathcal{X}_1||\mathcal{X}_2||\mathcal{Y}|\log(8/\delta)}{n}} + \frac{4|\mathcal{X}_1||\mathcal{X}_2||\mathcal{Y}|\log n}{n}$$

*with probability $1-\delta$ (Paninski, 2003). On binary alphabets the bound is $O(\sqrt{\log(1/\delta)/n})$.*

**Lemma 15** (Contrastive invariance). *Let $\mathcal{L}_{\mathrm{ctr}}(s; P_{X_1,X_2})$ be any contrastive loss that depends on a scoring function $s$ and on the pairing measure alone. On $\mathcal{C}_k$, the law $P_{X_1,X_2}$ is a product of uniforms for every support, so $\mathcal{L}_{\mathrm{ctr}}$ and every minimizer $s^\star$ are independent of $Y$.*

**Lemma 16** (Linear models miss $\mathcal{C}_k$). *For $Y = \chi_S(Z)$ with $|S| \ge 2$ and $Z \in \{\pm 1\}^d$ uniform, $\mathbb{E}[YZ_j] = 0$ for every coordinate $j$. The population logistic gradient at $w = 0$ vanishes, and the Bayes optimal linear classifier has error $1/2$.*

**Lemma 17** (Imputation data processing). *For any Markov kernel $T$ from $X_1$ to the alphabet of $X_2$, $I(Y; X_1, T(X_1)) \le I(Y; X_1) \le I(Y; X_1, X_2)$. The synergistic remainder equals $I(X_2; Y \mid X_1)$ and vanishes under $T$.*

**Proposition 2** (MMI translation is an identity). $S_{\mathrm{MMI}} + \min(U_1^{\mathrm{MMI}}, U_2^{\mathrm{MMI}}) = \mathrm{FVI}$ *holds for every $P$, and not only for Gaussians, because minimum mutual information sets $\min(U_1, U_2) = 0$ by definition. The mixture grid therefore has identically zero MMI translation error.*

**Proposition 3** (Three source chain). $I(X_1, X_2, X_3; Y) = I(X_1; Y) + I(X_2; Y \mid X_1) + I(X_3; Y \mid X_1, X_2)$. *Each summand is a Bayes logarithmic loss gain by Lemma 6. Higher order atoms of the eighteen atom lattice are included in the last two increments together with pairwise synergy.*

## C Proofs for fusion value and the dual

### C.1 Proof of Theorem 1

Lemma 6 gives $\mathcal{R}^\star_{\log}(X_i) - \mathcal{R}^\star_{\log}(X_1, X_2) = I(X_{3-i}; Y \mid X_i)$. The chain rule rewrites the conditional mutual information as $I(X_1, X_2; Y) - I(X_i; Y)$, and a Williams and Beer expansion then gives $(R + U_1 + U_2 + S) - (R + U_i) = U_{3-i} + S$.

For the gap to the better unimodal predictor, $\min_i \mathcal{R}^\star_{\log}(X_i) = H(Y) - \max_i I(X_i; Y)$. Subtracting $\mathcal{R}^\star_{\log}(X_1, X_2) = H(Y) - I(X_1, X_2; Y)$ yields FVI. Williams and Beer expands $\max_i I(X_i; Y) = R + \max(U_1, U_2)$, and therefore $\mathrm{FVI} = S + \min(U_1, U_2)$. If the ordering of the two views is reversed, the same algebra applies after the indices are swapped.

### C.2 Proof of Corollary 1

Minimum mutual information sets $R = \min_i I(X_i; Y)$, so $\min(U_1, U_2) = 0$ and $\mathrm{FVI} = S_{\mathrm{MMI}}$. Barrett's theorem identifies every pairwise decomposition with this choice on univariate Gaussians (Barrett, 2015). If $X_1$ Blackwell dominates $X_2$ for predicting $Y$, BROJA unique information of the weaker source vanishes as well (Bertschinger et al., 2014; Raginsky, 2011; Banerjee et al., 2018), and again $\min(U_1, U_2) = 0$.

### C.3 Proof of Theorem 2

Lemma 7 gives $\mathcal{R}^\star_{\mathrm{sq}}(X) = \sigma^2 e^{-2I(X;Y)}$. The gap between one view and the pair is therefore

$$\sigma^2\big(e^{-2I(X_i;Y)} - e^{-2I(X_1,X_2;Y)}\big) = \sigma^2 e^{-2I(X_i;Y)}\big(1 - e^{-2I(X_{3-i};Y|X_i)}\big).$$

The better unimodal form uses $i = \arg\max I(X_i; Y)$. Under minimum mutual information the exponent equals $S_{\mathrm{MMI}}$, and the first order expansion is $1 - e^{-2S} = 2S + O(S^2)$.

For a general square loss the same gap is the residual variance

$$\mathcal{R}^\star_{\mathrm{sq}}(X_i) - \mathcal{R}^\star_{\mathrm{sq}}(X_1, X_2) = \mathbb{E}\big[\mathrm{Var}(\mathbb{E}[Y \mid X_1, X_2] \mid X_i)\big].$$

This quantity is not a function of $S$ alone. The main text therefore states an equality only for logarithmic loss and for Gaussian square loss.

### C.4 Zero one sandwich

**Corollary 4** (Zero one envelope). *Let $\delta = I(X_{3-i}; Y \mid X_i)$, let $H_i = H(Y \mid X_i)$, and let $H_{12} = H_i - \delta$. There exist monotone maps $\alpha, \beta$, depending on $|\mathcal{Y}|$, such that $\alpha(H) \le P_e^\star \le \beta(H)$ and $\alpha(0) = \beta(0) = 0$. The Bayes error gap then satisfies*

$$\alpha(H_i) - \beta(H_{12}) \;\le\; P_e^\star(X_i) - P_e^\star(X_1, X_2) \;\le\; \beta(H_i) - \alpha(H_{12}).$$

Lemma 8 supplies the two maps. For binary $Y$, and when entropy is measured in bits, the binary upper bound yields a gap of at most $\delta/(2 \ln 2)$ after nats are converted to bits. The corollary bounds the Bayes error gap by the same conditional mutual information that appears in Theorem 1, and it does not identify that gap with $S$.

### C.5 Proof of Theorem 3

The Bayes gain from adding modality $j$ is $\mathcal{R}^\star(X_{3-j}) - \mathcal{R}^\star(X_1, X_2)$ by definition. The Bayes drop from deleting $j$, when the fallback is optimal, is the same difference, because that fallback is the unimodal Bayes predictor on $X_{3-j}$. Under logarithmic loss, Lemma 6 converts both expressions into $I(X_j; Y \mid X_{3-j})$, which equals $U_j + S$ in every Williams and Beer decomposition. Any empirical claim that the gain is controlled by synergy while the drop is controlled by uniqueness therefore fails for Bayes predictors, since both quantities are this one Shannon number. The correlations reported by Liang et al. match the expansion $U_{\text{miss}} + S$, with one term or the other dominating a given sweep (Liang et al., 2023a).

After modality $i$ is deleted, the residual information is $I(X_{3-i}; Y) = R + U_{3-i}$. The smallest residual is $\min_i I(X_i; Y) = R + \min(U_1, U_2) = \text{RI}$, so $\text{RI} \ge R$, with equality if and only if $\min(U_1, U_2) = 0$. The drop and the residual sum to $I(X_1, X_2; Y)$. Raising $R$ by moving mass from the unique atoms into redundancy, at fixed $S$, increases the residual and reduces the unique contribution to the drop, while the synergy term in the drop stays in place. Gács and Körner common information $C(X_1 \wedge X_2)$ is a property of the pair $(X_1, X_2)$ and can be independent of $Y$ (Yu & Tan, 2020). The identity uses only the Shannon floor $\min_i I(X_i; Y)$.

### C.6 Proof of Proposition 1

The map $x_1 \mapsto f_{12}(x_1, T(x_1))$ is a function of $X_1$ alone, so its risk is at least the unimodal Bayes risk of $X_1$ under logarithmic loss, square loss, and zero one loss. The excess is nonnegative, and it vanishes when the fallback attains that Bayes predictor. Data processing gives $I(Y; X_1, T(X_1)) \le I(Y; X_1)$, and the synergistic remainder $I(X_2; Y \mid X_1)$ disappears once $X_2$ is replaced by a function of $X_1$. On XOR one has $Y \perp X_1$, so no kernel $T$ recovers $Y$.

Under logarithmic loss with $X_2$ missing, the optimum is the posterior $P(Y \mid X_1)$. An impute and evaluate map attains this posterior only when $f_{12}$ computes the mixture $\sum_{x_2} P(x_2 \mid x_1)\, f_{12}(x_1, x_2)$ internally. Mean fill and zero fill skip that average, and both therefore incur excess drop. Appendix K records the corresponding numerical check.

## D Proofs for pairing value

### D.1 Proof of Theorem 4

Each posterior $\eta_i$ is a function of $X_i$, so the pair $(\eta_1, \eta_2)$ is a function of $(X_1, X_2)$. Lemma 5 yields $I(Y; \eta_1, \eta_2) \le I(Y; X_1, X_2)$. The logarithmic loss Bayes gap is $H(Y \mid \eta_1, \eta_2) - H(Y \mid X_1, X_2)$, which equals PVI. The chain rule also writes the same gap as the conditional mutual information $I(X_1, X_2; Y \mid \eta_1, \eta_2)$, which is nonnegative. Equality to zero holds if and only if $Y \perp (X_1, X_2) \mid (\eta_1, \eta_2)$. Each $\eta_i$ is sufficient for $Y$ given $X_i$, because $I(Y; \eta_i) = I(Y; X_i)$, but a pair of unimodal sufficient statistics need not be jointly sufficient for $Y$ given $(X_1, X_2)$.

For the upper bound, $I(Y; \eta_1, \eta_2) \ge I(Y; \eta_i) = I(Y; X_i)$ for each $i$. Hence $I(Y; \eta_1, \eta_2) \ge \max_i I(X_i; Y)$, and $\text{PVI} \le \text{FVI} = S + \min(U_1, U_2)$. The upper bound is attained when combining the two posteriors does not improve on the better unimodal channel. XOR is one such case, because both posteriors are constant, and Blackwell nesting is another, because the weaker posterior

is a garbling of the stronger one. AND attains the lower bound $\mathrm{PVI} = 0$ while FVI stays strictly positive, so the sandwich is not an identity. Lemma 12 applies the same data processing step to hard labels. The late versus early gap of a vote is therefore at least PVI, and voting is a coarser late fusion class than posterior stacking.

### D.2 Proof of Theorem 5

Let $X_1, X_2$ be i.i.d. uniform bits and $Y = X_1 \oplus X_2$. Then $P(Y = 1 \mid X_i = x) = 1/2$ for every $x$, so $\eta_i$ is constant and $I(Y; \eta_1, \eta_2) = 0$. Since $Y$ is a fair bit, $H(Y) = 1$ bit equals $I(Y; X_1, X_2)$, while $I(X_i; Y) = 0$. Every Williams and Beer decomposition is therefore forced to $R = U_1 = U_2 = 0$ and $S = 1$ bit, and $\mathrm{PVI} = S$.

Now let $Y = X_1 \wedge X_2$ with the same source. Then $P(Y = 1 \mid X_i = 0) = 0$ and $P(Y = 1 \mid X_i = 1) = 1/2$, so the vector $\eta_i$ encodes $X_i$ invertibly by the vanishing of the $Y = 1$ coordinate. Thus $\sigma(\eta_1, \eta_2) = \sigma(X_1, X_2)$ and $\mathrm{PVI} = 0$. The Shannon numbers are classical. Since $P(Y = 1) = 1/4$, one has $H(Y) = h_2(1/4) \approx 0.811$ bit. Also $H(Y \mid X_1 = 0) = 0$ and $H(Y \mid X_1 = 1) = 1$, so $H(Y \mid X_1) = 1/2$ and $I(X_1; Y) \approx 0.311$ bit. Jointly $I(X_1, X_2; Y) = H(Y) \approx 0.811$ bit, and therefore $\mathrm{FVI} \approx 0.500$ bit. On this world, minimum mutual information, $I_{\min}$, and BROJA all assign the whole of FVI to $S$ because $U_1 = U_2 = 0$ (Bertschinger et al., 2014; Liang et al., 2023a).

BROJA synergy equals $I_P(X_1, X_2; Y) - \min_{q \in \Delta_P} I_q(X_1, X_2; Y)$. The minimizer varies the coupling at fixed pairwise channels $(X_i, Y)$, whereas pairing value evaluates $I(Y; \eta_1, \eta_2)$ under the true $P$ and discards coordinates of $X_i$ that are absent from $\eta_i$. On AND the two quantities differ by half a bit. OR is AND with the label flipped, and COPY with $X_1 = X_2 = Y$ has $\mathrm{PVI} = S = 0$ and $R = H(Y)$.

### D.3 Proof of Lemma 1

Suppose a linear map $\psi(R, U_1, U_2, S) = aR + b_1U_1 + b_2U_2 + cS$ equals PVI on COPY, AND, and XOR. On COPY the quadruple is $(1, 0, 0, 0)$ bit and $\mathrm{PVI} = 0$, so $a = 0$. On XOR the quadruple is $(0, 0, 0, 1)$ bit and $\mathrm{PVI} = 1$ bit, so $c = 1$. On AND, minimum mutual information, $I_{\min}$, and BROJA all assign $S \approx 0.5$ bit with $U_1 = U_2 = 0$ and $R \approx 0.311$ bit, while $\mathrm{PVI} = 0$. These coefficients then return $\psi = cS \approx 0.5$ bit. The same three worlds exclude both $\mathrm{PVI} = S$ and $\mathrm{PVI} = \mathrm{FVI}$.

The lemma does not assert that two joints can share one quadruple and still differ in pairing value. A search over $3 \times 3 \times 2$ tables that share the same pairwise channels produced a largest pairing value gap of $0.021$ nats. That gap is recorded only as a numerical remark on joints with a common quadruple, and the separation used in the lemma is the one among COPY, AND, and XOR.

## E Worked examples

This section computes the Shannon numbers behind Theorem 5 from the joint table. The arithmetic does not call a partial information decomposition solver, so the separation between pairing value and synergy can be checked by hand.

### E.1 XOR

Let $X_1, X_2$ be i.i.d. $\mathrm{Bernoulli}(1/2)$ and $Y = X_1 \oplus X_2$. The joint places mass $1/8$ on each of the four even parity atoms and mass $0$ on the odd parity atoms, or equivalently mass $1/4$ on each of $(0, 0, 0)$, $(0, 1, 1)$, $(1, 0, 1)$, and $(1, 1, 0)$. Since $Y$ is unbiased, $H(Y) = \ln 2$. Also $P(Y = 1 \mid X_1 = 0) = P(X_2 = 1) = 1/2$, and the same conditional probability holds for every other conditioning on one bit, so $H(Y \mid X_i) = \ln 2$ and $I(X_i; Y) = 0$. Given $(X_1, X_2)$, the label is deterministic, hence $H(Y \mid X_1, X_2) = 0$ and $I(X_1, X_2; Y) = \ln 2$. It follows that $\mathrm{FVI} = \ln 2$ and $\mathrm{RI} = 0$. The unimodal posteriors are the constant pair $(1/2, 1/2)$, so $I(Y; \eta_1, \eta_2) = 0$ and $\mathrm{PVI} = \ln 2$. Williams and Beer nonnegativity then forces $R = U_1 = U_2 = 0$ and $S = \ln 2$.

### E.2 AND

Now let $Y = X_1 \wedge X_2$. Since $P(Y = 1) = 1/4$,

$$H(Y) = -\tfrac{1}{4}\ln\tfrac{1}{4} - \tfrac{3}{4}\ln\tfrac{3}{4} = h_2(1/4)\ln 2 \approx 0.562335 \text{ nats} = 0.811278 \text{ bits}.$$

Conditioning gives $P(Y = 1 \mid X_1 = 0) = 0$, so $H(Y \mid X_1 = 0) = 0$, and $P(Y = 1 \mid X_1 = 1) = 1/2$, so $H(Y \mid X_1 = 1) = \ln 2$. Hence $H(Y \mid X_1) = \frac{1}{2}\ln 2$ and

$$I(X_1; Y) = H(Y) - \tfrac{1}{2}\ln 2 \approx 0.215762 \text{ nats} = 0.311278 \text{ bits}.$$

The same values hold for $X_2$. Jointly $Y$ is a function of $(X_1, X_2)$, so $I(X_1, X_2; Y) = H(Y)$ and

$$\mathrm{FVI} = H(Y) - \big(H(Y) - \tfrac{1}{2}\ln 2\big) = \tfrac{1}{2}\ln 2 = 0.5 \text{ bits}.$$

The lower end of equation 7 equals $I(X_1; Y) + I(X_2; Y) - I(X_1, X_2; Y) = 2I(X_1; Y) - H(Y) \approx -0.1308$ nats, and therefore $R \geq 0$. The upper end is $\min_i I(X_i; Y) \approx 0.215762$ nats. Minimum mutual information takes that upper end, so $U_1 = U_2 = 0$ and $S_{\mathrm{MMI}} = \mathrm{FVI} = 0.5$ bit. The measure $I_{\min}$ agrees on this world, because the specific information $I(X_i; Y = 1)$ is the same for both sources and $I(X_i; Y = 0)$ is likewise the same.

The BROJA coupling that minimizes $I_q(X_1; Y \mid X_2)$ subject to the pairwise margins $(X_i, Y)$ is known in closed form for AND (Bertschinger et al., 2014). It yields $R \approx 0.311$ bit, $U_i = 0$, and $S \approx 0.500$ bit. A constrained concave maximization of $H_q(Y \mid X_1, X_2)$ over the transportation polytope recovers $S_{\mathrm{BROJA}} = 0.49999$ bit. The posterior $\eta_i$ is the pair $(P(Y = 0 \mid X_i), P(Y = 1 \mid X_i))$, equal to $(1, 0)$ when $X_i = 0$ and to $(1/2, 1/2)$ when $X_i = 1$. The map $X_i \mapsto \eta_i$ is invertible, so $I(Y; \eta_1, \eta_2) = I(Y; X_1, X_2)$ and $\mathrm{PVI} = 0$. AND is therefore a world on which $\mathrm{FVI} = S_{\mathrm{BROJA}} = 0.5$ bit, while stacking calibrated unimodal posteriors is already Bayes optimal.

### E.3 COPY AND THE UNIQUE WORLD

The copy $X_1 = X_2 = Y$ of a uniform bit gives $I(X_i; Y) = I(X_1, X_2; Y) = \ln 2$, together with $\mathrm{FVI} = \mathrm{PVI} = 0$, $R = \ln 2$, and $U_i = S = 0$. One also has $I(X_1; X_2) = \ln 2$. A contrastive pairing objective therefore sees a perfect alignment problem, while the fusion gain is zero. If instead $Y = X_1$ is independent of $X_2$, then $\mathrm{FVI} = \mathrm{PVI} = 0$, $I(X_1; Y) = \ln 2$, $I(X_2; Y) = 0$, and $\mathrm{RI} = 0$. Minimum mutual information, BROJA, and $I_{\min}$ all assign $U_1 = \ln 2$ and set the remaining atoms to zero. Deleting $X_2$ then drops nothing, and deleting $X_1$ drops the whole mutual information. OR is AND with the labels flipped, so it has the same Shannon numbers and $\mathrm{PVI} = 0$.

A noisy copy sends a common bit $Y$ through two independent binary symmetric channels of crossover $0.15$. That construction reports $\mathrm{FVI} \approx 0.145$ nats, $\mathrm{PVI} = 0$, $\mathrm{RI} \approx 0.270$ nats, and $I(X_1; X_2) \approx 0.125$ nats. Late fusion of the posteriors is Bayes optimal because the channels are nested in the Blackwell order, which matches $\mathrm{PVI} = 0$.

### E.4 GAUSSIANS

**Proposition 4** (Barrett). *If $Y$ is univariate Gaussian and $(Y, X_1, X_2)$ are jointly Gaussian, then every decomposition whose redundancy and unique atoms depend only on $P_{X_i,Y}$ equals minimum mutual information (Barrett, 2015).*

The pairwise marginals pin $I(X_i; Y)$. The weaker channel is a linear garbling of the one dimensional sufficient statistic $\mathbb{E}[Y \mid X_{\mathrm{strong}}]$, so its unique information vanishes in the Blackwell order. When each $X_i$ is scalar, $\mathbb{E}[Y \mid X_i]$ is an invertible function of $X_i$ under a nondegenerate covariance, and the early minimum mean squared error estimator is a linear combination of $(X_1, X_2)$, equivalently of $(\mathbb{E}[Y \mid X_1], \mathbb{E}[Y \mid X_2])$. Late linear fusion then equals early fusion. On the scalar asymmetric construction below one has $S_\Delta^{\mathrm{mmse}} = 0$ and $\mathrm{FVI} = S_{\mathrm{MMI}} = 0.091$ nats.

For a vector source, $\mathbb{E}[Y \mid X_i]$ is a scalar projection $c_i^\top X_i$, while early minimum mean squared error uses a coefficient vector $a$ on $X_1$ and a coefficient vector $b$ on $X_2$. If $a$ is not parallel to $c_1$, a direction in $X_1$ that is marginally silent about $Y$ becomes useful together with $X_2$. Then $\mathbb{E}[Y \mid X_1, X_2]$ is not a function of the two scalar posteriors, and $S_\Delta^{\mathrm{mmse}} > 0$. The vector construction $Y = A + B + C$, $X_1 = (A, B) + \text{noise}$, and $X_2 = (A, C) + \text{noise}$, with noise variance $0.25$ on each coordinate, has

linear minimum mean squared error pairing value $0.167$ nats. Degree $k$ parity is a non Gaussian interaction, so the numerical check in Appendix K tests the embedding of Appendix G.

Let $Y \sim \mathcal{N}(0,1)$ and $X_i = a_i Y + N_i$ with $a_i = \sqrt{\mathrm{SNR}_i}$, $N_i \sim \mathcal{N}(0,1)$ independent, $\mathrm{SNR}_1 = 4$, and $\mathrm{SNR}_2 = 1$. Then $I(X_i;Y) = \frac{1}{2}\log(1+\mathrm{SNR}_i)$, so $I(X_1;Y) = \frac{1}{2}\log 5 \approx 0.805$ nats and $I(X_2;Y) = \frac{1}{2}\log 2 \approx 0.347$ nats. The minimum mean squared error formula of Lemma 7 produces $I(X_1, X_2; Y) = 0.896$ nats and $\mathrm{FVI} = 0.091$ nats. Minimum mutual information synergy equals FVI, one has $\min(U_1, U_2) = 0$, and $\mathrm{RI} = I(X_2;Y) = 0.347$ nats. The late linear estimator $\mathbb{E}[Y \mid \mathbb{E}[Y \mid X_1], \mathbb{E}[Y \mid X_2]]$ coincides with $\mathbb{E}[Y \mid X_1, X_2]$, because each scalar $X_i$ is an invertible function of $\mathbb{E}[Y \mid X_i]$. Hence $S_\Delta^{\mathrm{mmse}} = 0$.

Let $A, B, C$ be i.i.d. standard Gaussians, and set $Y = A + B + C$, $X_1 = (A + N_{10},\, B + N_{11})$, and $X_2 = (A + N_{20},\, C + N_{21})$, with independent observation noise of variance $0.25$. Each unimodal projection $\mathbb{E}[Y \mid X_i]$ is a scalar, while the early estimator uses both coordinates of each $X_i$. The resulting numbers are $I(X_i;Y) = 0.381$, $I(X_1, X_2; Y) = 0.885$, $\mathrm{FVI} = 0.504$, and $I(Y; \mathbb{E}[Y \mid X_1], \mathbb{E}[Y \mid X_2]) = 0.718$, so $S_\Delta^{\mathrm{mmse}} = 0.167$ nats.

### E.5 BROJA optimization

Bertschinger et al. define the unique information of $X_1$ by

$$U_1(P) = \min_{q \in \Delta_P} I_q(X_1; Y \mid X_2) = H_P(Y \mid X_2) - \max_{q \in \Delta_P} H_q(Y \mid X_1, X_2), \tag{8}$$

where $\Delta_P$ is the polytope of joints with the same pairwise margins $P_{X_i,Y}$ (Bertschinger et al., 2014). The map $q \mapsto H_q(Y \mid X_1, X_2)$ is concave, so the problem is a concave maximization over a polytope. On binary alphabets $|\mathcal{X}_i| = |\mathcal{Y}| = 2$ the polytope has low dimension, and a constrained nonlinear solver recovers the table of Bertschinger et al. to $10^{-5}$ bits.

Write $q(x_1, x_2, y) = r(x_1, x_2 \mid y)P(y)$ and enforce $q(x_i, y) = P(x_i, y)$ by iterated proportional fitting on each slice of $y$. The free variables are the conditional couplings $r(\cdot, \cdot \mid y)$ in the transportation polytope with margins $P(x_i \mid y)$. Local optimality is checked by comparing $S + U_1 + U_2 + R$ with $I(X_1, X_2; Y)$. On the canonical worlds, that residual is below $10^{-12}$ nats. The minimizer $q^\star$ varies the source coupling at fixed channels $(X_i, Y)$. Pairing value does not vary a coupling, because it evaluates $I(Y; \eta_1, \eta_2)$ under the true $P$. Equations equation 8 and equation 2 therefore have no algebraic reason to agree, and Theorem 5 shows that they do not.

## F Proofs for finite class remainders

### F.1 Proof of Lemma 2

Let $G = \min_i \mathcal{R}_{\log}(f_i) - \mathcal{R}_{\log}(f_{12})$ denote the empirical fusion gain. Adding and subtracting the two Bayes risks gives

$$\begin{aligned} G = &\big(\min_i \mathcal{R}^\star_{\log}(X_i) - \mathcal{R}^\star_{\log}(X_1, X_2)\big) \\ &+ \big(\min_i \mathcal{R}_{\log}(f_i) - \min_i \mathcal{R}^\star_{\log}(X_i)\big) \\ &- \big(\mathcal{R}_{\log}(f_{12}) - \mathcal{R}^\star_{\log}(X_1, X_2)\big). \end{aligned}$$

The first parenthesis equals $\mathrm{FVI}(P)$ by Theorem 1. The second parenthesis is the unimodal excess risk, and the third is the joint excess risk. Both excesses are nonnegative. The displayed identity is equation 4. The triangle inequality then bounds $|G - \mathrm{FVI}(P)|$ by the sum of the two excesses.

### F.2 Proof of Corollary 3

If $\mathrm{FVI}(P) \le \varepsilon$, then equation 4 gives $G \le \varepsilon$ plus the unimodal excess. On a copy world, and on language saturated sentiment, the better singleton already accounts for almost all of $I(X_1, X_2; Y)$. In that case $\varepsilon$ is on the order of a hundredth of a nat, so a ranking of architecture gains is a ranking of optimization error.

### F.3 Proof of Lemma 3

If $g_i = t_i \circ \eta_i$, then $(g_1, g_2)$ is a function of $(\eta_1, \eta_2)$. Data processing yields $I(Y; g_1, g_2) \leq I(Y; \eta_1, \eta_2)$, and therefore $H(Y \mid g_1, g_2) - H(Y \mid X_1, X_2) \geq \mathrm{PVI}(P)$. Temperature scaling, quantile binning, and Platt calibration are all of this form. An uncalibrated unimodal head that is not a function of $\eta_i$ can move the gap in either direction. The operational protocol therefore calibrates on held out data before the two probabilities are stacked.

### F.4 Proof of Lemma 4

Let $\mathrm{PVI}_{\hat{\eta}} = H(Y \mid \hat{\eta}_1, \hat{\eta}_2) - H(Y \mid X_1, X_2)$, and let $\mathrm{PVI}_{\mathrm{alg}} = \mathcal{R}_{\log}(\mathrm{late}) - \mathcal{R}_{\log}(f_{12})$. Adding and subtracting the two conditional entropies gives

$$\begin{aligned}\mathrm{PVI}_{\mathrm{alg}} = \mathrm{PVI}_{\hat{\eta}} &+ \big(\mathcal{R}_{\log}(\mathrm{late}) - H(Y \mid \hat{\eta}_1, \hat{\eta}_2)\big) \\ &- \big(\mathcal{R}_{\log}(f_{12}) - H(Y \mid X_1, X_2)\big).\end{aligned}$$

A coarse sketch of $(X_1, X_2, Y)$ is a function of a quantized pair, and it is a fourth quantity. It need not equal any of the three gaps. On CLIP XOR the sketch pairing value is $0.07$ nats, while the operational pairing value is $0.65$ nats. On MOSI the sketch fusion value is near $0.43$ nats, while the operational remainder is $0.01$ nats. Both discrepancies are instances of this lemma, and they are consistent with Theorem 4.

### F.5 Plug in rates and unpaired samples

On a finite alphabet the plug in entropy satisfies, with probability $1 - \delta$,

$$|\hat{H}(A) - H(A)| \leq C\sqrt{\frac{|\mathcal{A}| \log(2/\delta)}{n}} + \frac{|\mathcal{A}| \log n}{n}$$

for an absolute constant $C$ (Antos & Kontoyiannis, 2001; Paninski, 2003). Mutual information is a linear combination of entropies. Fusion value uses $H(Y)$, $H(Y \mid X_i)$, and $H(Y \mid X_1, X_2)$, and the largest alphabet among these is $|\mathcal{X}_1||\mathcal{X}_2||\mathcal{Y}|$. The map $(a, b) \mapsto a - \max(b, \cdot)$ is 1 Lipschitz, so $|\widehat{\mathrm{FVI}} - \mathrm{FVI}|$ inherits the same rate. If $|\mathrm{FVI} - \tau| \geq 2\varepsilon_n$ for this deviation $\varepsilon_n$, the sign of $\widehat{\mathrm{FVI}} - \tau$ matches the population sign. Binning a continuous variable yields $I(X^{(\mathrm{bin})}; Y) \leq I(X; Y)$ by data processing, so a common binning ranks two datasets conservatively.

**Proposition 5** (Unpaired samples)**.** *Unpaired samples $\{(x_i, y)\}$ identify $P_{X_i,Y}$, and therefore identify $I(X_i; Y)$, the residual floor* RI*, and Shannon upper bounds on* FVI *such as* $\min\{I(X_2; Y), H(Y \mid X_1)\}$*. They do not identify $I(X_1, X_2; Y)$ or $I(Y; \eta_1, \eta_2)$, because the latter requires the same label to be paired with both posteriors. Hence* PVI *is the information theoretic value of collecting pairs, relative to training two unimodal experts.*

The proposition separates two design questions that use different samples. When the decision is whether to discard the weaker modality, the relevant index is FVI. When the decision is whether to collect paired triples and train an early fusion head, the relevant index is PVI.

### F.6 Rademacher and VC remainders

Assumption 1 uses standard Rademacher bounds (Bartlett & Mendelson, 2002) and the mutual information generalization bound of Xu and Raginsky (Xu & Raginsky, 2017). For a finite class $\mathcal{F}$ of maps into $\Delta_{\mathcal{Y}}$, Hoeffding's inequality and a union bound give, with probability $1 - \delta$,

$$\sup_{f \in \mathcal{F}} \big|\mathcal{R}_{\log}(f) - \hat{\mathcal{R}}_{\log}(f)\big| \leq \sqrt{\frac{\log(2|\mathcal{F}|/\delta)}{2n}}$$

whenever $-\log f(Y)$ lies in $[0, 1]$ after clipping. The empirical fusion gain of empirical risk minimization therefore tracks FVI up to $O(\sqrt{\log(|\mathcal{F}|/\delta)/n})$ plus approximation error. For infinite classes, the covering number of a Lipschitz network contributes a standard factor $L^{O(D)}$. The same comparison can be written with Rademacher complexity. If $\mathfrak{R}_n(\ell \circ \mathcal{F})$ denotes the complexity of a loss composed class, then with probability $1-\delta$ one has $|\mathcal{R}(\hat{f}) - \mathcal{R}(f^\star_{\mathcal{F}})| \leq 2\mathfrak{R}_n + O(\sqrt{\log(1/\delta)/n})$

(Bartlett & Mendelson, 2002). When $f_{\mathcal{F}}^{\star}$ is within $\varepsilon$ of the true Bayes predictor, the empirical fusion gain differs from Theorem 1 by $O(\mathfrak{R}_n(\mathcal{F}_i) + \mathfrak{R}_n(\mathcal{F}_{12}) + \varepsilon + \sqrt{\log(1/\delta)/n})$. Xu and Raginsky replace the Rademacher term by $\sqrt{I(W;S_n)/n}$ (Xu & Raginsky, 2017).

A VC bound on the zero one gap likewise adds a remainder of order $\sqrt{(V \log n + \log(1/\delta))/n}$, where $V$ is the VC dimension (Bartlett et al., 2021). For linear models on $d$ bits one has $V = O(d)$. Linear models still have correlation zero with $\mathcal{C}_k$ by Lemma 16, so the approximation error on that family is $1/2$ and already accounts for the chance level error.

## G Contrastive blindness and a statistical query embedding

Fix nonempty blocks $A$ and $B$, and set $Y = \chi_A(X_1)\chi_B(X_2)$ on the uniform hypercube. Given $X_1$, the factor $\chi_B(X_2)$ is an independent Rademacher random variable whenever $B$ is nonempty, so $Y \perp X_1$. The symmetric argument gives $Y \perp X_2$. The label is a nonconstant deterministic function of $(X_1, X_2)$ and is unbiased, hence $I(Y; X_1, X_2) = H(Y) = 1$ bit. Every Williams and Beer decomposition assigns $R = U_i = 0$ and $S = 1$ bit. The unimodal posterior is $P(Y = \pm 1 \mid X_i) = 1/2$, so $\eta_i$ is constant and $\mathrm{PVI} = 1$ bit. Early Bayes outputs $Y$ and has zero one error equal to zero, while any unimodal predictor or posterior late predictor is at chance.

Characters $\chi_S(z) = \prod_{j \in S} z_j$ on $\{\pm 1\}^d$ are orthonormal in $L^2$ of the uniform measure. Cross modal parities are ordinary characters on $d = d_1 + d_2$ whose support meets both blocks. The family $\mathcal{C}_k$ is therefore pairwise orthogonal, and its cardinality is

$$M = \sum_{a=1}^{k-1} \binom{d_1}{a} \binom{d_2}{k-a}.$$

For $d_1 = d_2 = d/2$ and fixed $k$ one has $M = d^{\Theta(k)}$. If $k$ scales linearly with $d$, then $M$ is exponential in $d$.

**Lemma 18** (Orthogonal statistical query bound)**.** *Let $\mathcal{F} \subset \{\pm 1\}^{\mathcal{X}}$ be pairwise orthogonal with $|\mathcal{F}| = M$. Any statistical query algorithm that uses tolerance $\tau$ and outputs a hypothesis with correlation at least $\gamma$ with an unknown $f \in \mathcal{F}$ requires $\Omega(\gamma^2 \tau^{-2})$ queries in the worst case over $f$, and it cannot succeed with $o(\min\{M, \tau^{-2}\})$ queries (Feldman, 2017; Blum et al., 2003; Feldman et al., 2017).*

The lemma applies the orthogonal statistical query bound to this embedding, and the paper does not add a new statistical query theorem. A query $\phi(x, y)$ has expectation $\mathbb{E}[\phi(X, f(X))] = \mathbb{E}[\phi_+] + \mathbb{E}[\phi_- f]$ under the label $f$. The vector of Fourier coefficients of a bounded function $\psi$ has $\ell_2$ mass $O(1)$, so a typical coordinate is $O(M^{-1/2})$. Tolerance $\tau$ masks every coordinate smaller than $\tau$. Adaptive query trees are controlled by the statistical query dimension of an orthogonal family, which is $\Theta(M)$. Embedding $\mathcal{C}_k$ into parities on $\{\pm 1\}^{d_1+d_2}$ adds no further combinatorial room. The bound constrains statistical query algorithms. Stochastic gradient descent with unbiased gradient estimates can fall outside that model (Shoshani & Shamir, 2025), and the claim used in the main text is the contrastive one.

On $\mathcal{C}_k$ the coordinates $X_1$ and $X_2$ are uniform and independent, and $P_{X_1,X_2}$ does not depend on the blocks $(A, B)$. Any statistic of paired or unpaired samples $\{(x_1, x_2)\}$ that does not see $Y$ is therefore independent of the label function. A contrastive loss such as $\mathbb{E}[\log(e^{s(x_1,x_2)} / \sum e^{s(x_1,x_2')})]$ is a functional of $P_{X_1,X_2}$ alone (Radford et al., 2021; Tosh et al., 2021). If the pairing distribution itself carries task relevant redundancy, contrastive learning can succeed on that redundant channel. The hard instance above has none of that redundancy.

Let $Z = [X_1; X_2] \in \{\pm 1\}^d$, $Y = \chi_S(Z)$, $|S| = k$, and $\sigma(w^\top Z)$ with $w \sim \mathcal{N}(0, I/d)$. Expand $\sigma$ in probabilists' Hermite polynomials. The inner product $\mathbb{E}[Y\sigma(w^\top Z) \mid w]$ receives contributions only from degree $k$. A monomial $\prod_{i \in S} w_i$ has Gaussian moment 0 for odd $k$ when the coordinates are not paired, and moment $O(d^{-k/2})$ after accounting for $\mathrm{Var}(w_i) = 1/d$. Typical feature to label correlations are then $\Theta(d^{-k/2})$. A linear score $w^\top Z$ satisfies $\mathbb{E}[Y Z_j] = 0$ for every coordinate when $k \geq 2$, so the population gradient at the origin vanishes. The lowest degree that correlates with $\mathcal{C}_k$ is $k$. Multiplicative interactions are therefore a natural inductive bias for synergistic parity

(Zadeh et al., 2017), and they are also the reason additive fusion records the lowest synergy in the comparison of Liang et al. (Liang et al., 2023a).

## H SEQUENTIAL FUSION OF THREE MODALITIES

The three source Williams and Beer lattice has eighteen atoms. Past two views, this paper uses the chain rule increment in the definition below.

**Definition 3** (Sequential fusion value)**.** *For an ordering $X_1, \dots, X_m$, let $\Delta_{j|<j}$ denote $I(X_j; Y \mid X_{<j})$, with $\Delta_{1|\emptyset} = I(X_1; Y)$. The chain rule gives $I(X_{1:m}; Y) = \sum_j \Delta_{j|<j}$. Each increment is the Bayes logarithmic loss gain from adding modality $j$, by Theorem 1.*

A conditional two source decomposition of $\Delta_{2|1}$ given $X_1$ has no standard redundancy measure, so the paper reports the Shannon increment only. Fang et al. condition on language $T$ and decompose $I(Y; V, A \mid T)$ (Fang et al., 2026). The chain rule writes this conditional mutual information as $I(Y; V \mid T) + I(Y; A \mid T, V)$, which is sequential fusion value for the two sensory views given language. When the first summand is a low order correlation and the second is a higher order interaction, the random feature calculation of Appendix G indicates that gradient methods fit the visual term first. That calculation is an account of learnability for their sensory bottleneck. Appendix K checks the same distinction on a three bit toy with $Y = V \oplus A$ and with $T$ independent of $Y$. The pair $(V, A)$ then carries pure synergy, while the pair $(V, T)$ has fusion value zero.

## I COMPARISON WITH HUANG AND LIANG

Huang et al. bound population risk by the distance between a learned latent $\hat{g}$ and a true latent $g^\star$, and they show that a larger modality set estimates $g^\star$ more accurately (Huang et al., 2021). If $g^\star$ is sufficient for $Y$ given $(X_1, X_2)$, then $I(g^\star; Y) = I(X_1, X_2; Y)$, and a subset $N$ satisfies $I(X_N; Y) \le I(X_{1:2}; Y)$. The resulting risk gap is Theorem 1. Their bound treats the modality set as a whole. The split into unique information and synergy, and the late fusion condition $\mathrm{PVI} = 0$, are separate from that comparison. Modality competition and unimodal phases explain why gradient descent may fail to attain the Bayes gap (Huang et al., 2022; Zhang et al., 2024; Wang et al., 2020). Theorem 1 names that gap, and Appendix G names a further gap that remains even when the hypothesis class can represent the target. Together the two layers account for the modest empirical fusion gains reported in that literature (Wang et al., 2020; Huang et al., 2022).

Liang et al. estimate BROJA at the dataset level and at the model level. Agreement of the decomposition with the trained predictor correlates with accuracy at $\rho = 0.81$. The missing modality drop correlates with $U_i$ at $\rho = 0.8$ and, when $U_i$ is small, with $S$ at $\rho = 0.73$ (Liang et al., 2023a). Those figures are correlations for trained models, whereas Theorems 1 to 3 are equalities for Bayes predictors. Their missingness comparison is the empirical counterpart of $U_{\mathrm{miss}} + S$. Pairing value is a further gap beyond that correlation. The same conditional mutual information already appears in multi view analysis as the complementarity $\Gamma_{X_2} = I(X_2; Y \mid X_1) = U_2 + S$ (Sridharan & Kakade, 2008). A Fano bound on the missing modality drop depends on $\Gamma$. That bound is the drop in Theorem 3, and it is coarser than the residual floor RI.

## J EXPERIMENTAL PROTOCOL

Finite alphabet joints are enumerated, Gaussian worlds use the closed form mutual information of a scalar $Y$, and parity worlds are sampled i.i.d. Atomic worlds use one fixed seed. Gaps on MNIST, CLIP, and caption matching are averages over three seeds, and sentiment remainders use the official MMSA train and test splits. COPY sets $X_1 = X_2 = Y$. The unique world sets $Y = X_1$ with $X_2$ independent. XOR sets $Y = X_1 \oplus X_2$, AND sets $Y = X_1 \wedge X_2$, and OR sets $Y = X_1 \vee X_2$. Noisy COPY observes a common bit through two independent binary symmetric channels of crossover $0.15$. Sources are uniform bits unless a construction says otherwise. A convex combination of COPY, the unique world, and XOR, with Dirichlet weights on sixty six points, produces a continuum of $(\mathrm{FVI}, \mathrm{PVI}, I(X_1; X_2))$. The atoms are not linear in the mixture weights. The checks are the Shannon identity between Bayes logarithmic loss gain and FVI, together with the translation $\mathrm{FVI} = S + \min(U_1, U_2)$ for minimum mutual information, $I_{\min}$, and BROJA.

Scalar Gaussians use $Y \sim \mathcal{N}(0,1)$ and $X_i = \sqrt{\mathrm{SNR}_i}\, Y + N_i$ with signal to noise ratios $(4,1)$. Vector Gaussians use $Y = A + B + C$ as in Appendix E.4. Parity uses $X_1, X_2 \in \{\pm 1\}^8$, hence concatenated dimension 16, and $Y = \chi_A \chi_B$ with $|A| + |B| = k \in \{2,3,4,6,8\}$ split across the two modalities. Training sizes lie in $\{200, 500, 1000, 2000, 4000\}$, and the test size is $1500$. The three source toy uses bits $V, A, T$, once with $Y = V \oplus A$ independent of $T$, and once with $Y$ aligned with $T$ and independent of $(V, A)$. Shannon quantities are computed as $-\sum p \log p$ on the joint table, with the convention $0 \log 0 = 0$. Minimum mutual information and $I_{\min}$ are closed form. BROJA maximizes $H_q(Y \mid X_1, X_2)$ over $\Delta_P$. Canonical correlation on bits is the Pearson correlation of the $\{0,1\}$ encodings. The plug in estimator of FVI uses empirical joints from $n$ i.i.d. samples.

Table Bayes takes the argmax of $P(y \mid x)$ on the true or empirical table. Posterior late fusion forms $\eta_i$ from $P(Y \mid X_i)$ and then builds the Bayes table on $(\eta_1, \eta_2, Y)$. Logistic regression uses four hundred steps, an $\ell_2$ penalty of $10^{-4}$, and learning rate $0.2$ with two halvings. The parity network is one hidden rectified linear layer of width $64$, trained for nine hundred steps of stochastic gradient descent with batch size $128$, learning rate $0.08$, and $\ell_2$ penalty $10^{-5}$. Networks receive concatenated $\pm 1$ coordinates and do not receive explicit monomial features. Two view MNIST downsamples digits to $14 \times 14$ by averaging $2 \times 2$ blocks, trains on $8000$ examples, and tests on $2000$. The coarse Shannon sketch is a two dimensional principal component analysis with four quantile bins, while the learners see a sixteen dimensional principal component projection of the same views. Copy uses a noisy clone, the unique world uses an independent digit, and the split world uses the left half and the right half of one digit. AND pair and XOR pair take two independent digits and label the Boolean of their even or odd bits.

Frozen CLIP uses ViT-B/32 image embeddings and the ten class prompts of CIFAR $10$. Named regimes use an animal versus vehicle label, except when the caption matching label is named. Fusion heads on CLIP features are two layer GELU networks, trained with early stopping on held out logarithmic loss. The operational protocol trains two unimodal heads and one joint head, and it reports held out logarithmic loss gaps. Late fusion is a polynomial stack of the two unimodal probabilities after held out calibration. Caption matching uses the same CLIP features and sets the label to whether the prompt names the depicted class. The vision language check fine tunes Qwen3-VL instruct models at 2B, 4B, and 8B with LoRA rank $8$ on the same matching prompt, and the ablations drop the image tokens or the caption tokens while keeping the rest of the prompt. Sentiment uses mean pooled MMSA features of CMU-MOSI and CMU-MOSEI. Language is BERT, and vision and audio are the official aligned streams. The operational remainder uses the same head family as the CLIP bank. A three dimensional principal component sketch with four quantile bins is reported only as a contrast with the operational gap, and it is not treated as $\mathrm{FVI}(P)$.

# K Additional numerical results

## K.1 Canonical identities

Seven worlds and three measures give a maximum of $|\mathrm{FVI} - (S + \min(U_1, U_2))|$ equal to $4.4 \times 10^{-16}$ nats, attained by BROJA on noisy COPY. The gap between FVI and the Bayes logarithmic loss gain is identically zero, because the table risks are computed from the same joint. Table 3 records the Shannon numbers. Sixty six Dirichlet mixtures have identically zero translation error under minimum mutual information, by Proposition 2. On Dirichlet random joints with alphabets up to $5 \times 3 \times 2$, that error remains identically zero, and $0 \le \mathrm{PVI} \le \mathrm{FVI}$ holds in every draw. Table 4 lists the five alphabet shapes. The BROJA remainder is below $10^{-6}$ nats except on one $3 \times 2 \times 3$ table, where it equals $0.012$ nats. That entry is a local solver remainder, and it is consistent with Theorem 1.

## K.2 AND versus XOR and architecture

Table 5 matches the classical values of Bertschinger et al. (Bertschinger et al., 2014). Fusion value equals $S + \min(U_1, U_2)$ in every row, while pairing value vanishes on AND and equals one bit on XOR. An additive logistic head has test error $0.50$ on XOR and error $0$ on AND. The same model with one multiplicative coordinate has error $0$ on both worlds, as Table 6 records. Posterior late fusion is at chance on XOR and Bayes optimal on AND. COPY is solved by every head in the

Table 3: Canonical worlds, in nats. The last column is $|\mathrm{FVI} - (S + \min(U_1, U_2))|$ under BROJA.

| World | FVI | PVI | $S_{\mathrm{BROJA}}$ | RI | $I(X_1; X_2)$ | BROJA error |
|---|---|---|---|---|---|---|
| COPY | 0 | 0 | 0 | 0.693 | 0.693 | 0 |
| $\mathrm{UNIQUE}_1$ | 0 | 0 | 0 | 0 | 0 | 0 |
| XOR | 0.693 | 0.693 | 0.693 | 0 | 0 | 0 |
| AND | 0.347 | 0 | 0.347 | 0.216 | 0 | $< 10^{-15}$ |
| OR | 0.347 | 0 | 0.347 | 0.216 | 0 | $< 10^{-15}$ |
| noisy COPY | 0.145 | 0 | 0.145 | 0.270 | 0.125 | $4.4 \times 10^{-16}$ |

Table 4: Random joints. Entries are $\max |\mathrm{FVI} - (S + \min(U_1, U_2))|$ in nats.

| $\lvert\mathcal{X}_1\rvert \times \lvert\mathcal{X}_2\rvert \times \lvert\mathcal{Y}\rvert$ | MMI | $I_{\min}$ | BROJA | PVI $\leq$ FVI |
|---|---|---|---|---|
| $2 \times 2 \times 2$ | 0 | $10^{-18}$ | $10^{-15}$ | 12/12 |
| $3 \times 3 \times 2$ | 0 | $10^{-15}$ | $5 \times 10^{-7}$ | 12/12 |
| $4 \times 4 \times 2$ | 0 | $10^{-17}$ | $10^{-10}$ | 12/12 |
| $3 \times 2 \times 3$ | 0 | $10^{-15}$ | 0.012 | 12/12 |
| $5 \times 3 \times 2$ | 0 | $10^{-15}$ | $10^{-7}$ | 12/12 |

table, including a head that sees only one view. A multiplicative interaction is required for Bayes optimality precisely when $\mathrm{PVI} > 0$.

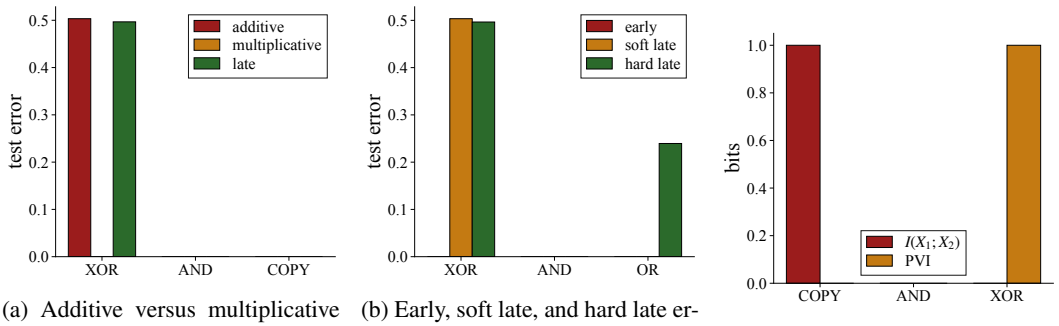
Table 5: Canonical worlds in bits. Fusion value equals $S + \min(U_1, U_2)$ in every row.

| World | PVI | $S_{\mathrm{BROJA}}$ | FVI | $I(X_1; Y)$ |
|---|---|---|---|---|
| XOR | 1.000 | 1.000 | 1.000 | 0 |
| AND | 0.000 | 0.500 | 0.500 | 0.311 |
| COPY | 0 | 0 | 0 | 1.000 |

Table 6: Test error on the canonical worlds. Late fusion stacks posteriors, and early fusion is table Bayes.

| World | Additive | Multiplicative | MLP | Late | Early |
|---|---|---|---|---|---|
| XOR | 0.50 | 0.00 | 0.00 | 0.50 | 0.00 |
| AND | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| COPY | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| OR | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |

Figure 6a compares the additive and multiplicative heads, and Figure 6b shows that hard labels are a coarser late class than posterior stacking. Figure 6c ranks COPY above XOR by cross modal mutual information, which is the opposite order from pairing value. Figure 6 places the three comparisons in one row.



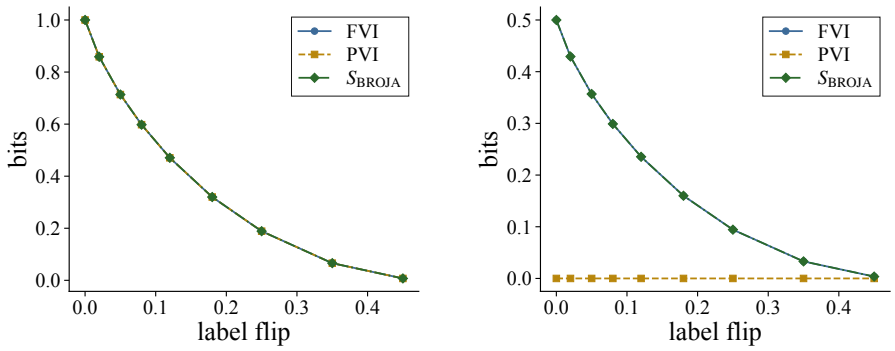

(a) Additive versus multiplicative test error.

(b) Early, soft late, and hard late error.

(c) Alignment versus pairing value.

Figure 6: Architecture follows pairing value, and hard labels are a coarser late class than posterior stacking. Cross modal mutual information ranks COPY above XOR, whereas pairing value ranks the two worlds in the opposite order.

The separation survives label flips. Figure 7a keeps $\mathrm{PVI} = \mathrm{FVI} = S_{\mathrm{BROJA}}$ at every flip rate on XOR, and Figure 7b keeps $\mathrm{PVI} = 0$ on AND while BROJA synergy tracks fusion value downward. Figure 7 collects the two sweeps.




(a) Label flips on XOR.

(b) Label flips on AND.

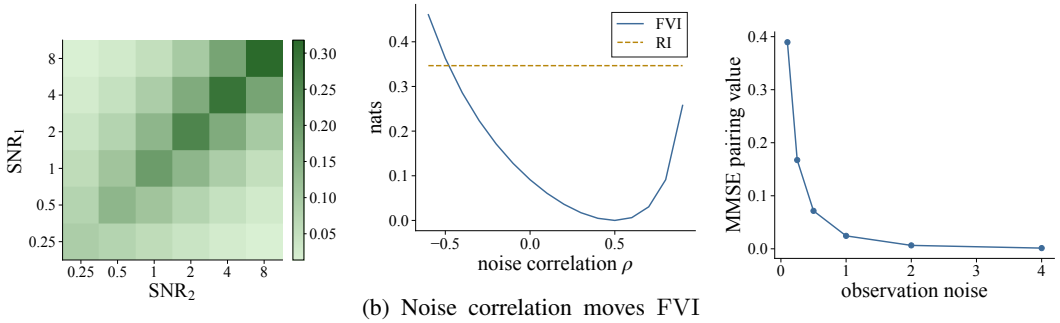
Figure 7: Label flips preserve the separation between AND and XOR.

### K.3 Missingness, Gaussians, and ranking

The Bayes zero one drop from deleting $X_2$ is $0.497$ on XOR, $0.254$ on AND, and $0$ on COPY. The logarithmic loss drop equals $I(X_2; Y \mid X_1)$ to machine precision. Hard samples of $X_2 \sim P(X_2 \mid X_1)$ return XOR to chance and leave COPY at accuracy $1.00$, as Proposition 1 requires. Table 7 records the Gaussian arithmetic of Appendix E.4. A six by six grid of signal to noise ratios on Barrett's scalar construction has $\min(U_1, U_2) = 0$ and $\text{FVI} = S_{\text{MMI}}$ in every cell, which is Figure 8a. Sweeping the observation noise correlation $\rho$ at signal to noise ratios $(4, 1)$ moves FVI and RI in opposite directions, as shown in Figure 8b. On the vector construction of Figure 8c, the linear pairing value $S_{\Delta}^{\text{mmse}}$ stays strictly positive across observation noise variances in $\{0.1, 0.25, 0.5, 1, 2, 4\}$. Figure 8 collects the three panels.

Table 7: Gaussian worlds, in nats. The late column is $I(Y; \mathbb{E}[Y \mid X_1], \mathbb{E}[Y \mid X_2])$.

| World | $I(X_1; Y)$ | $I(X_{12}; Y)$ | FVI | $I_{\text{late}}$ | $S_{\Delta}^{\text{mmse}}$ |
|---|---|---|---|---|---|
| scalar asymmetric | 0.805 | 0.896 | 0.091 | 0.896 | 0 |
| vector sources | 0.381 | 0.885 | 0.504 | 0.718 | 0.167 |



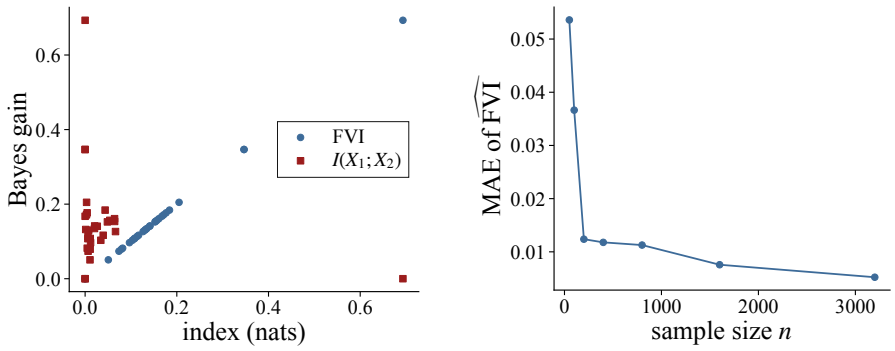

(a) Scalar grid of fusion value.

(b) Noise correlation moves FVI against RI.

(c) Vector pairing value.

Figure 8: Gaussian calculations. Scalar sources have vanishing linear pairing value, while vector sources keep a strictly positive value.

Cross modal mutual information expands as $I(X_1; X_2) = I(X_1; X_2 \mid Y) + I(X_1; X_2; Y)$. The first term is correlation between the sources that does not involve $Y$, and the second term equals $R - S$. COPY makes both $R$ and $I(X_1; X_2)$ large while $\text{FVI} = 0$, and XOR makes $I(X_1; X_2)$ small while $\text{FVI} = 1$ bit. Any monotone function of $I(X_1; X_2)$ therefore ranks fusion value in the opposite direction on the segment from COPY to XOR. Across the mixtures, the Spearman correlation of FVI with the true Bayes gain is $1.00$, while the same correlation for $I(X_1; X_2)$ is $-0.12$ and for the first canonical correlation is $-0.14$. Figure 9a shows the fusion value comparison. Against empirical gains of a multilayer perceptron, fusion value has correlation $0.63$ and canonical correlation analysis has correlation $-0.33$. Pairing value ranks the late versus early logarithmic gap at $1.00$ on the enumerated table. The plug in mean absolute error of $\widehat{\text{FVI}}$ falls from $0.054$ nats at $n = 50$ to $0.005$ nats at $n = 3200$, which is Figure 9b. Figure 9 places the two panels together.

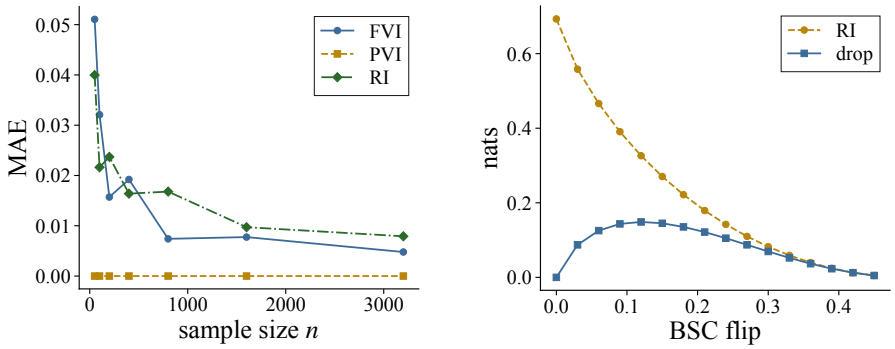



(a) Fusion value versus Bayes gain.

(b) Plug in error of fusion value.

Figure 9: Fusion value ranks Bayes gain, while alignment ranks that gain in the opposite direction. The plug in estimator follows the usual entropy rate.

On a mixture with weights $0.4$ on XOR, $0.3$ on AND, and $0.3$ on COPY, the mean absolute error of all three indices falls with $n$, as shown in Figure 10a. Pairing value is the noisiest of the three at $n = 50$, and it remains of order $n^{-1/2}$. Figure 10b shows that on noisy COPY the residual floor and the drop move together, which is the Blackwell nested case in which $\mathrm{RI} = R_{\mathrm{MMI}}$. Figure 10 collects the two panels.




(a) Plug in error of the three indices.

(b) Noisy COPY floor and drop.

Figure 10: Plug in error of the three indices on a mixed world, together with the residual floor on a Blackwell nested family.

### K.4 Parity and random features

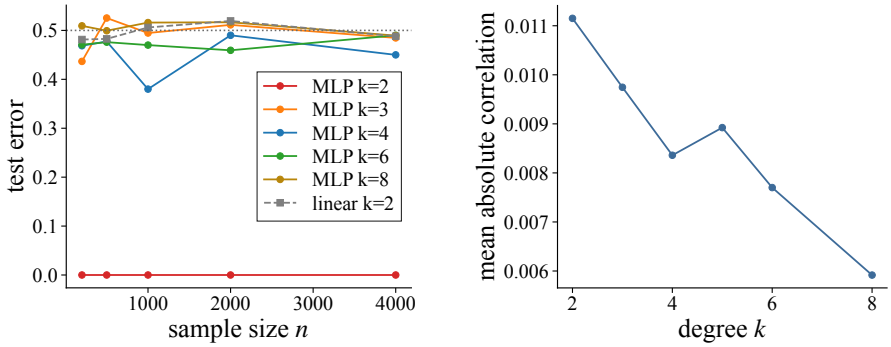
Degree 2 cross modal XOR is learned by a multilayer perceptron of width $64$ at $n = 200$, while linear models remain at one half. For $k \geq 3$ and $n \leq 4000$ the network error stays near one half. Table 8 lists the values, which lie in $[0.38, 0.53]$, and Figure 11a plots the same sweep. Linear models stay in $[0.47, 0.54]$ for every $k \geq 2$, as Lemma 16 requires. Two thousand random $\tanh$ features $\sigma(w^\top Z)$, with $w \sim \mathcal{N}(0, I/d)$ and $d = 20$, have mean absolute correlation with $Y = \chi_S(Z)$ that decreases in $k$. Table 9 and Figure 11b record the decay, which matches the Hermite calculation of Appendix G. Figure 11 places the sample size sweep next to the correlation decay.

Table 8: Parity test error of the width $64$ network. Every linear entry lies in $[0.47, 0.54]$.

| $k$ at $n$ | 200 | 500 | 1000 | 2000 | 4000 |
|---|---|---|---|---|---|
| 2 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 3 | 0.44 | 0.53 | 0.49 | 0.51 | 0.48 |
| 4 | 0.47 | 0.48 | 0.38 | 0.49 | 0.45 |
| 6 | 0.47 | 0.48 | 0.47 | 0.46 | 0.49 |
| 8 | 0.51 | 0.50 | 0.52 | 0.52 | 0.49 |

Table 9: Absolute correlation of random features with a degree $k$ parity label.

| $k$ | 2 | 3 | 4 | 5 | 6 | 8 |
|---|---|---|---|---|---|---|
| mean \|corr\| | 0.0112 | 0.0097 | 0.0084 | 0.0089 | 0.0077 | 0.0059 |
| max \|corr\| | 0.045 | 0.056 | 0.032 | 0.032 | 0.030 | 0.022 |



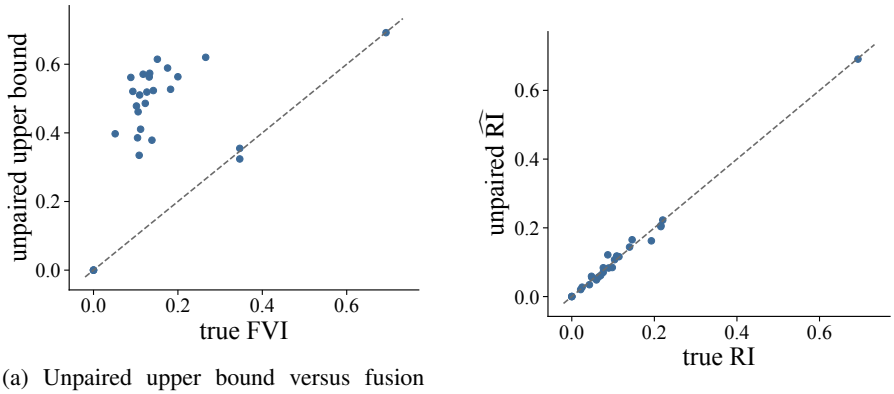

(a) Parity test error versus sample size.

(b) Mean absolute feature correlation.

Figure 11: Degree 2 is interpolated at these sample sizes, while degree $k \geq 3$ remains near chance. The mean absolute correlation of the random features decreases with $k$.

### K.5 Unpaired samples and three sources

Unpaired samples $\{(x_i, y)\}$ identify $I(X_i; Y)$ and therefore RI, with Spearman correlation 0.978 against the paired truth. They also identify the Shannon upper bound $H(Y) - \max_i I(X_i; Y)$ on FVI, which is valid in 25 of 27 draws. Figure 12a plots that upper bound, and Figure 12b plots the residual floor. The same samples do not identify PVI, which is the content of Proposition 5. Figure 12 collects the two panels.




(a) Unpaired upper bound versus fusion value.

(b) Unpaired floor versus the paired floor.

Figure 12: Unpaired samples recover the residual floor and an upper bound on fusion value. Pairing value stays unidentified from those samples.

With $Y = V \oplus A$ and $T$ independent of $Y$, a multilayer perceptron on $(V, A)$ interpolates the two bit XOR, while $(V, T)$ stays at chance for every $n \leq 1280$. With $Y = T$ and $(V, A)$ independent of the label, the pair $(V, T)$ interpolates immediately and $(V, A)$ stays at chance. Figure 13 shows both regimes. Low order aligned pairs are fit at these sample sizes, and high order synergistic pairs are not.

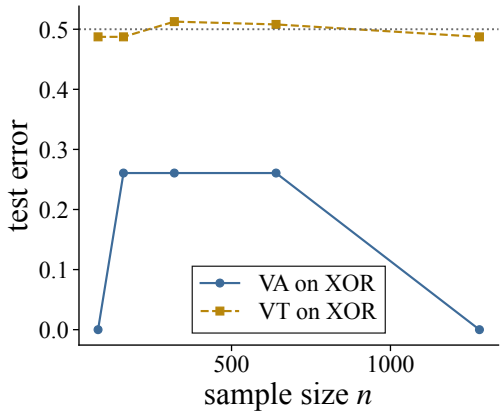


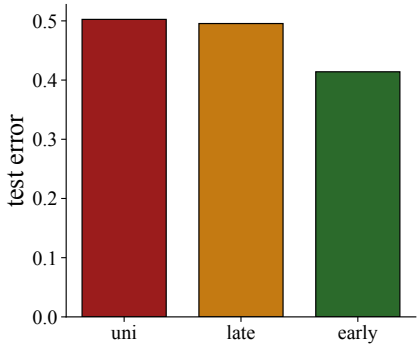

Figure 13: Three source test error versus sample size. The dashed line marks chance.

### K.6 Pixels, CLIP heads, and sentiment

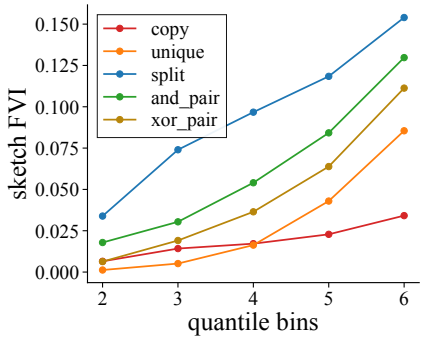
Table 10 records the two view MNIST sketch. The Spearman correlation of the sketch FVI with empirical early gain is 0.77, and the correlation of $I(X_1; X_2)$ with the same gain is $-0.14$. XOR pair is the only row in which both unimodal errors sit at chance while early fusion still improves. The same coarse sketch reports pairing value near zero on that row. A trained polynomial stack of the two unimodal probabilities restores a strictly positive pairing value of the same sign as the observed late versus early gap, which is Figure 14a. Absolute sketch FVI grows with the number of quantile bins, as shown in Figure 14b, while ranks against the four bin reference stay in $[0.90, 1.00]$. Figure 14 collects the two panels.

Table 10: Two view MNIST. Fusion value is in nats, and errors are means over three seeds.

| World | FVI | PVI | Unimodal | Early | Late |
|---|---|---|---|---|---|
| copy | 0.017 | 0 | 0.150 | 0.147 | 0.158 |
| copy heavy | 0.020 | 0 | 0.150 | 0.155 | 0.151 |
| unique | 0.016 | 0 | 0.150 | 0.161 | 0.151 |
| split | 0.097 | 0 | 0.208 | 0.146 | 0.169 |
| AND pair | 0.056 | 0.000 | 0.243 | 0.154 | 0.244 |
| XOR pair | 0.037 | 0 | 0.488 | 0.400 | 0.493 |




(a) Operational errors on the MNIST XOR pair.




(b) Sketch fusion value versus the number of bins.

Figure 14: The pixel XOR pair has a positive operational pairing value of the same sign as the late versus early gap. Ranks of the sketch stay stable as the number of bins changes.

Frozen CLIP ViT-B/32 features of CIFAR 10 recover the same three regimes, and Table 11 records the errors. Copy and the unique world gain 0.000 and 0.001. AND gains 0.216, and score fusion matches that gain. XOR gains 0.366 from unimodal heads at chance, while score fusion stays at chance. Across thirty one versus rest tasks, a binned fusion value ranks the gain at 0.74, while $I(X_1; X_2)$ ranks it at $-0.39$.

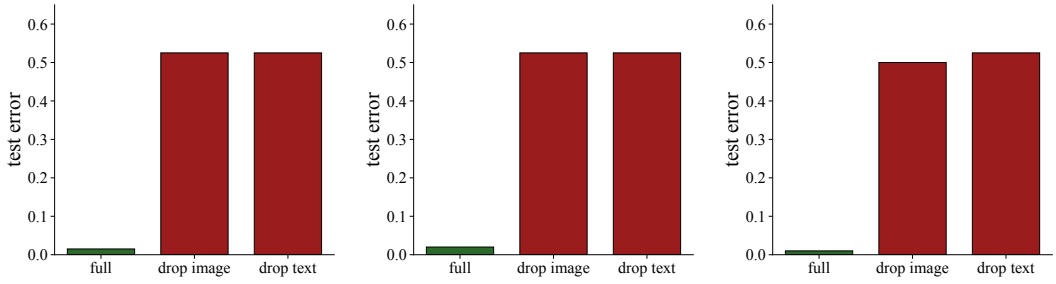
Table 11: Frozen CLIP on CIFAR 10, animal versus vehicle. Errors and gains are accuracy, and the sketch quantities are in nats.

| World | FVI | $I(X_1; X_2)$ | Unimodal | Early | Late | Gain |
|---|---|---|---|---|---|---|
| copy | 0.000 | 1.14 | 0.000 | 0.000 | 0.000 | 0.000 |
| unique | 0.003 | 0.011 | 0.009 | 0.008 | 0.010 | 0.001 |
| split | 0.081 | 0.703 | 0.041 | 0.019 | 0.035 | 0.022 |
| AND | 0.329 | 0.011 | 0.225 | 0.009 | 0.011 | 0.216 |
| XOR | 0.559 | 0.011 | 0.401 | 0.035 | 0.555 | 0.366 |

The same matching prompt, with LoRA tuning of Qwen3-VL at 2B, 4B, and 8B, keeps the full model error at 0.015, 0.020, and 0.010. Dropping the image or the caption returns the error to chance at every size. Table 12 records the three sizes, and the logarithmic loss gaps lie between 0.65 and 0.70 nats. Figure 15a, Figure 15b, and Figure 15c show the three sizes separately, and Figure 15 collects them.

Table 12: LoRA tuned Qwen3-VL on caption matching. Errors are held out, and $\Delta$ll is the logarithmic loss gap in nats.

| Size | Full error | Drop image | Drop text | $\Delta$ll image | $\Delta$ll text |
|---|---|---|---|---|---|
| 2B | 0.015 | 0.525 | 0.525 | 0.680 | 0.698 |
| 4B | 0.020 | 0.525 | 0.525 | 0.650 | 0.648 |
| 8B | 0.010 | 0.500 | 0.525 | 0.678 | 0.673 |




(a) Qwen at 2B. (b) Qwen at 4B. (c) Qwen at 8B.

Figure 15: Caption matching error for Qwen3-VL at three sizes. Full fine tuning is near zero, and dropping either view returns the error to chance.

Table 13 compares the operational protocol with the coarse sketch. On copy and on the unique world the operational fusion value is near zero, while the sketch remains on the order of a tenth of a nat. On XOR the operational pairing value is 0.65 nats against a sketch value of 0.07 nats. On caption matching both operational indices equal about 0.43 nats. On MOSI and MOSEI the operational remainder is about a hundredth of a nat, while the sketch sits near 0.43 nats in every reported cell. Minimum mutual information and $I_{\min}$ are closed form. BROJA uses the transportation parameterization above. Mean processor time over eight Dirichlet draws is 0.037 seconds on $2 \times 2 \times 2$, 0.15 seconds on $3 \times 3 \times 2$, and 0.39 seconds on $5 \times 3 \times 2$. The maximum of $|\mathrm{FVI} - (S + \min(U_1, U_2))|$ is at machine precision except for $2.5 \times 10^{-5}$ nats on $3 \times 2 \times 3$, which is Figure 16.

Table 13: Operational remainder versus a coarse sketch. Entries are in nats. The mark n.r. means that the sketch pairing value was not computed.

| World | $\widehat{\mathrm{FVI}}_{\mathrm{op}}$ | Sketch FVI | $\widehat{\mathrm{PVI}}_{\mathrm{op}}$ | Sketch PVI |
|---|---|---|---|---|
| CLIP copy | $9 \times 10^{-5}$ | 0.103 | 0.074 | 0.037 |
| CLIP unique | $-0.005$ | 0.035 | 0.082 | 0.027 |
| CLIP XOR | 0.650 | 0.143 | 0.653 | 0.070 |
| CLIP match | 0.438 | 0.123 | 0.427 | 0.065 |
| MOSI language vision | 0.011 | 0.437 | 0.021 | 0.212 |
| MOSI language audio | 0.010 | 0.433 | 0.020 | 0.206 |
| MOSI vision audio | 0.012 | 0.435 | 0.031 | 0.171 |
| MOSEI language vision | 0.009 | 0.481 | 0.018 | n.r. |
| MOSEI language audio | $-0.002$ | 0.481 | 0.008 | n.r. |
| MOSEI vision audio | 0.000 | 0.515 | 0.006 | n.r. |

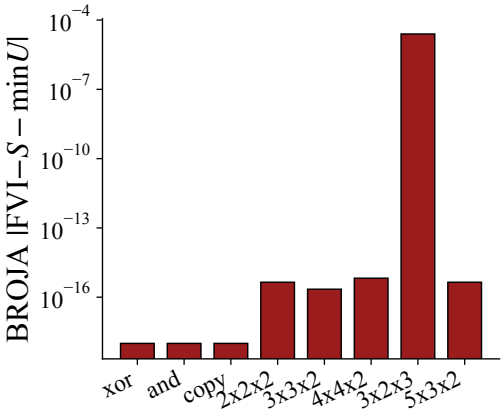



Figure 16: BROJA translation remainder on the canonical worlds and on Dirichlet joints.