SOURCE VS. MECHANISTIC REDUNDANCY IN POINTWISE PARTIAL INFORMATION DECOMPOSITION

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce a framework for pointwise partial information decomposition (PID) that estimates sample- and distribution-level redundancy, uniqueness, and synergy between two modalities and a target. We define an $I_{\min}$-inspired pointwise redundancy measure that is computed directly from predictive models, avoiding explicit high-dimensional input-density estimation. We further introduce a supervised Gács–Körner-inspired objective for learning target-predictive shared representations across modalities. This enables us to decompose predictive redundancy into *source redundancy*, arising from information shared between the modalities, and *mechanistic redundancy*, arising when distinct modality-specific mechanisms provide overlapping predictive information. Across analytically tractable synthetic benchmarks, controlled multimodal tasks, and clinical data, the framework improves pointwise estimation accuracy, recovers known source–mechanistic structure, and yields interpretable multimodal information-sharing patterns. `https://anonymous.4open.science/r/PID-6026/README.md`

1 INTRODUCTION

Multimodal machine learning aims to integrate heterogeneous data modalities to improve predictive performance Baltrušaitis et al. (2018). Beyond prediction, an equally important challenge is understanding how each modality contributes to the final decision and how information is shared across modalities Liang et al. (2022). A modality may provide information unavailable elsewhere (*unique information*), multiple modalities may encode the same predictive signal (*redundant information*), or prediction may only emerge from their interaction (*synergistic information*) Zhang et al. (2024); Castro et al. (2019). Understanding these interaction patterns is essential for interpreting multimodal models. The Partial Information Decomposition (PID) provides a principled framework for quantifying these interactions Williams & Beer (2010). Recent advances have made PID estimation practical in high-dimensional settings Liang et al. (2023); Venkatesh et al. (2023); Zhao et al. (2025), opening the possibility of studying information sharing in modern multimodal learning systems.

Despite this progress, existing methods estimate a single decomposition for an entire dataset. Yet multimodal interaction patterns often vary substantially across samples Luppi et al. (2024). One prediction may rely almost entirely on one modality, another may be explained by redundant information shared across modalities, while a third may genuinely require their interaction. Distribution-level PID cannot distinguish these situations. Sample-level decomposition is therefore essential for understanding individual predictions, improving interpretability, identifying meaningful subgroups, and characterizing individual variability Lizier et al. (2013); Oviatt et al. (2005). Estimating PID at the sample level, however, remains challenging. Existing pointwise formulations offer complementary trade-offs. The common-change-in-surprisal framework Ince (2017a) provides a principled local decomposition but becomes intractable for continuous high-dimensional variables Finn & Lizier (2018). More recently, LSMI Yang et al. (2025) proposed a scalable estimator based on neural entropy estimation. However, LSMI requires estimating the sample-wise marginal surprisal $h(x_m) = -\log p(x_m)$ in high-dimensional representation space, introducing an additional and potentially sensitive density-estimation step.

More fundamentally, we argue that current PID formulations conflate two different forms of redundancy. *Source redundancy* arises when input modalities share common latent factors predictive of the target, and *mechanistic redundancy* arises when sources are independent, but both predictive of 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.

target (e.g., the OR gate) Harder et al. (2013); Pica et al. (2017). Existing PID estimators measure the sum of these two: the redundancy. This distinction is closely related to degeneracy in biological systems Tononi et al. (1999); Li et al. (2012), and Ince (2017b) showed it can be quantified via a partial entropy decomposition, though only for low-dimensional discrete systems. Current estimators fail to separate the two components in high dimensions, even more at the sample level.

In this work, we address both challenges through a unified framework. We first derive a tractable pointwise PID estimator by extending the $I_{\min}$ principle Williams & Beer (2010) to pointwise mutual information, requiring only predictive models. We then introduce a supervised formulation of Gács–Körner common information Gács & Körner (1973); Kleinman et al. (2023) that isolates source redundancy by learning target-aware shared representations, thereby separating source and mechanistic redundancy at both the distribution and sample levels. We additionally show that CLIP Radford et al. (2021) approximately optimizes a relaxed, unsupervised form of the Gács–Körner objective, which motivates the supervised variant we propose. Our contributions are:

1. We define a tractable pointwise PID estimator by extending the $I_{\min}$ principle Williams & Beer (2010) to pointwise mutual information. It requires only predictive models, avoiding the input-density estimation of LSMI (Yang et al., 2025).
2. We introduce a **supervised Gács–Körner objective** (SupGK) that makes a relaxed intersection information (Griffith et al., 2014) estimable in high dimensions and pointwise, separating *source* from *mechanistic* redundancy. We additionally show that CLIP Radford et al. (2021) learns common representations by optimizing a relaxation of the unsupervised Gács–Körner objective, and that supervision captures target-predictive shared structure.
3. We validate the framework on synthetic benchmarks, controlled multimodal tasks, and logic circuits, improving over LSMI at both the distribution and pointwise levels. We compare with human intuition, providing insight into their discrepancy and how to reconcile them.
4. In Huntington's disease, we link a speech representation to structural neuroimaging: SupGK combined with SHAP localizes brain regions whose information about lexical fluency is shared with speech, suggesting what a speech-based biomarker captures.

2 Background and Related Works

Let X_1, X_2, Y be random vectors taking values in $\mathcal{X}_1 \subseteq \mathbb{R}^{d_{X_1}}$, $\mathcal{X}_2 \subseteq \mathbb{R}^{d_{X_2}}$, $\mathcal{Y} \subseteq \mathbb{R}^{d_Y}$. We study the mutual information between modalities X_1, X_2 and a target Y, where samples (x_1, x_2, y) are drawn from their joint distribution. Distribution-Level Mutual Information (MI) is written $I(X;Y) = H(Y) - H(Y|X)$, (or $I(X_1, X_2; Y) = H(Y) - H(Y|X_1, X_2)$ in multimodal settings), where $H(Y)$ is the entropy of the target Y [1]. And pointwise MI is defined as $i(x;y) = \log \frac{p(x,y)}{p(x)p(y)}$.

2.1 Partial Information Decomposition at the Distribution Level

PID Williams & Beer (2010); Bertschinger et al. (2014) posits that $I(X_1, X_2; Y)$ decomposes into non-negative components, redundancy R, uniquenesses U_1, U_2 and synergy S:

$$I(X_1, X_2; Y) = R + S + U_1 + U_2 \qquad I(X_m; Y) = R + U_m. \tag{1}$$

These constraints leave one degree of freedom, so the decomposition is determined whenever one of the four measures is estimated, generally the redundancy. Two redundancy definitions predominate. Williams & Beer (2010) propose: $I_{\min} = \sum_y p(y) \min_m I(Y{=}y; X_m)$, with $I(Y{=}y; X_m) = \mathbb{E}_{x_m \sim p(x_m|y)}\big[\, i(x_m; y) \,\big]$, where $I(Y{=}y; X_m)$ is the *specific information*, the average information the modalities provide about the particular outcome y. Bertschinger et al. (2014) instead define unique information through a constrained optimization over the joint distributions preserving the marginals $P(X_1, Y)$ and $P(X_2, Y)$. It is defined through an optimization over a set of distributions, but admits no natural per-sample (or pointwise) counterpart. Estimators built on these definitions extend to continuous and high-dimensional data (Venkatesh et al., 2023; Pakman et al., 2021; Zhao et al., 2025; Mages & Rohner, 2023), notably the scalable convex estimators (BATCH) of Liang et al. (2023). Griffith et al. (2014) define redundancy as the target information carried by the Gács–Körner common variable.

[1] where $H(Y) = -\mathbb{E}_{y \sim p(y)} \log p(y)$, is a statistical measure of the uncertainty of a random variable.

2.2 Partial Information Decomposition at the Pointwise Level

The estimation of the sample-level enables a better understanding of how each observation contributes to the multimodal information structure Lizier et al. (2013). Denoting pointwise interactions r, u_1, u_2, s, which are not necessarily non-negative, the event-level decomposition reads (Yang et al. (2025); Lizier et al. (2013); Ince (2017a)):

$$i(x_1; y) = r + u_1, \quad i(x_2; y) = r + u_2, \quad i(x_1, x_2; y) = r + u_1 + u_2 + s \tag{2}$$

This system has four unknowns but only three estimable quantities, so the key challenge is defining one of the quantities r, s, u_1, u_2. In practice, related works Yang et al. (2025); Lizier et al. (2013); Ince (2017a) generally propose to estimate r.

Ince (2017a) proposed the common change in surprisal (CCS), which keeps non-null only the cases where both sources are informative about the outcome, or when both are misinformative, so that the shared part is unambiguous. However, it requires a maximum entropy distribution over the full joint state space, which restricts it to low-dimensional discrete variables (Finn & Lizier, 2018).

Finn & Lizier (2018) instead split the pointwise information into two non-negative terms, $i(x_m; y) = h(x_m) - h(x_m \mid y)$, and apply the redundancy lattice (minimum operation) to each separately, yielding $r = \min_m h(x_m) - \min_m h(x_m \mid y)$, which is signed as a difference of non-negative terms. Yang et al. (2025) make this construction tractable with the Lightweight Sample-wise Multimodal Interaction (LSMI) estimator, expressing the second term as $h(x_m \mid y) = h(x_m) - h(y) - \log p(y \mid x_m)$ so that only the marginal entropy $h(x_m)$ and a classifier are required. LSMI requires estimating the marginal entropy $h(x_m) = -\log p(x_m)$ in high-dimensional input space via the KNIFE estimator (Pichler et al., 2022), whose theoretical guarantees only hold at the expectation level; individual pointwise values $h_\theta(x_m)$ lack such control, yet LSMI relies on precisely these values to compute per-sample redundancy making it less accurate. Our formulation estimates modality-wise pointwise mutual information as $i(x_m; y) = \log p(y \mid x_m) - \log p(y)$ and then applies a pointwise version of the $I_{\min}$ principle. $p(y \mid x_m)$ is the output of a standard classifier or regressor trained on each modality, and $p(y)$ is the empirical target distribution (class frequencies for categorical Y, sample mean and variance for Gaussian Y), both readily available in any supervised learning pipeline, with no need to model the input density $p(x_m)$. Finally, LSMI does not distinguish source from mechanistic redundancy, which our objective addresses (Section 3.2).

3 Method

In this section, Sec.3.1 proposes a simpler definition of redundancy at a pointwise level. Sec.3.2 defines *source redundancy* at the distribution and pointwise level using a supervised version of Gács & Körner (1973) which estimates the *common representation* between two modalities.

3.1 Estimating pointwise redundancy via a pointwise version of $I_{\min}$

Inspired by $I_{\min}$ Williams & Beer (2010), we define a pointwise redundancy measure from the pointwise mutual information provided by each modality. For sample j, let

$$i_m^{(j)} = i(x_m^{(j)}; y^{(j)}) = \log p(y^{(j)} \mid x_m^{(j)}) - \log p(y^{(j)}), \qquad m \in \{1, 2\}.$$

We define pointwise predictive redundancy as

$$r^{(j)} = \begin{cases} \min\left(i_1^{(j)}, i_2^{(j)}\right), & \text{if } i_1^{(j)} i_2^{(j)} > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

Inspired by the sign-consistency principle of Ince (2017a), redundancy is non-zero only when both modalities push the prediction in the same direction, and opposite-sign contributions get zero. This measure only needs $p(y \mid x_m)$ and avoids the pointwise input surprisal $h(x_m) = -\log p(x_m)$, which is unreliable in high dimensions. The remaining atoms follow from Eq. 2. However, $r^{(j)}$ is *predictive* redundancy: the two modalities may provide overlapping information about Y through distinct mechanisms without sharing it (*mechanistic redundancy*), which we separate from source redundancy in Sec. 3.2.1.

3.2 SUPERVISING THE GÁCS-KÖRNER OBJECTIVE TO ESTIMATE SOURCE REDUNDANCY

The Gács-Körner common information objective (Gács & Körner, 1973), Kleinman et al. (2023):

$$C_{\mathrm{GK}}(X_1; X_2) = \max_{Z} H(Z^C) \text{ s.t. } Z^C = g_1(X_1) = g_2(X_2) \text{ a.s.} \tag{4}$$

estimates the common representation Z^C, having the largest possible entropy (i.e., the most informative), while remaining predictable from both modalities X_1 and X_2.

We show in Appendix B that the text-image pretraining method CLIP (Radford et al., 2021) implicitly optimizes the following relaxed GK objective, where $Z_1^C = g_{\theta_1}(X_1)$ and $Z_2^C = g_{\theta_2}(X_2)$:

$$\max_{\theta_1,\theta_2} H(Z_1^C) + H(Z_2^C) \quad \text{s.t.} \quad \|z^{C}{}_1^{(i)} - z^{C}{}_2^{(i)}\|_2^2 = 0 \ \ \forall i \tag{5}$$

By approximating entropy via Kernel Density Estimation (KDE) Parzen (1962); Rosenblatt (1956), we show that it recovers the symmetric InfoNCE loss of CLIP (Eq. 20). This interpretation corroborates findings that CLIP only captures redundant information (Dufumier et al. (2025), Fig. 4). This provides a new interpretation of CLIP's objective as an approximate relaxed Gács–Körner objective.

3.2.1 SUPERVISED GÁCS-KÖRNER FOR SOURCE REDUNDANCY ESTIMATION.

To estimate "source-redundant" representations Z_m^R, we seek to capture the common information that is directly predictive of Y. We generalize it, by maximizing $I(Z^R; Y)$ instead of $H(Z^C)$:

$$\max_{\theta_1,\theta_2} I(Z_1^R; Y) + I(Z_2^R; Y) \quad \text{s.t.} \quad \|z^{R}{}_1^{(i)} - z^{R}{}_2^{(i)}\|^2 = 0 \ \ \forall i, \tag{6}$$

where $Z_m^R = g_{\theta_m}(X_m)$ are ℓ_2-normalized. Under a common-source model, Z^R is a function of Z^C with $I(Z^R; Y) = I(Z^C; Y) =: R_{\mathrm{src}}$, and the loss selects its minimal sufficient version (App. D). Under a hard constraint, R_{src} equals the intersection information $I_\wedge$ of Griffith et al. (2014), which is defined only for discrete variables; SupGK makes it estimable for continuous, high-dimensional inputs by relaxing the Gács–Körner constraint

Maximizing $H(Z_m^R)$ aims at estimating the most informative representations possible. Minimizing $H(Z_m^R \mid Y)$ aims at capturing the target information. In Appendix C, we used KDE and Renyi-2 entropy to simplify the objective in Eq. 6 as the minimization of the following loss function:

$$\mathcal{L}_{\mathrm{SupGK}}^{\mathrm{categorical}} = \underbrace{\frac{1}{|\mathcal{P}|}\sum_{(i,j)\in\mathcal{P}}(K_{ij}-1)^2}_{\text{same-class alignment } H(Z^R|Y)\downarrow} + \underbrace{\frac{1}{|\mathcal{N}|}\sum_{(i,j)\in\mathcal{N}}K_{ij}^2}_{\text{different-class separation } H(Z^R)\uparrow} + \underbrace{\frac{\lambda}{N}\sum_i \|z^{R}{}_1^{(i)} - z^{R}{}_2^{(i)}\|^2}_{\text{cross-modal alignment}}, \tag{7}$$

where $K_{ij} = \exp\!\big(-\|z^{R}{}_1^{(i)} - z^{R}{}_2^{(j)}\|^2/2\tau\big)$, $\mathcal{P} = \{(i,j) : y_i = y_j\}$ (same-class pairs), $\mathcal{N} = \{(i,j) : y_i \neq y_j\}$ (different class pairs). For continuous Y, hard membership is replaced by soft weights $S_{ij} = \exp\!\big(-|y_i - y_j|^2/2\sigma_y^2\big)$, see Appendix C. Overall, our pipeline to estimate the pointwise and distribution-level PID with source-mechanistic redundancy decomposition consists of 4 steps:

1) Training joint and unimodal models to estimate $p(y|x_1, x_2)$, $p(y|x_m)$, and $Z_m = f_m(X_m)$. The first step consists of training a) a joint model $h \circ f(.)$, to predict y through encoder $f(.)$ and prediction head $h(.)$, allowing the estimation of $p(y|x_1, x_2)$ at the distribution-level (by averaging), and pointwise-level ($p(y_i|x_{1i}, x_{2i})$). And b) unimodal models $h_m \circ f_m(.)$, to estimate $p(y|x_m)$ (again, at pointwise and distribution level); but also to estimate unimodal representations $Z_m = f_m(X_m)$.

2) Estimating "source-redundant" representations $Z_m^R = g_m^R(Z_m)$. The second step consists of freezing the unimodal encoders $f_m(.)$ to work directly on the compressed representations Z_m. The objective is to train two neural networks $g_m^R(.)$ to produce "source-redundant" representations $Z_m^R = g_m^R(Z_m)$, which are ℓ_2-normalized. To do so, we train the models to minimize the SupGK loss introduced in Eq 7. The "source-redundant" predictors $g_m^R(.)$ are generally parameterized as a two-layer MLP with hidden size of 1024, as this parameterization has been shown to be able to approximate any smooth function. Ultimately, a shared prediction head h_R is trained on *stop-gradient (s.g)* frozen representations: $\mathrm{s.g}(z_m^R)$ to predict y which allows estimating $p(y|z_m^R)$ at the distribution level and $p(y_i|z_{mi}^R)$ at the pointwise-level.

3) Estimating pointwise-level PID with source-mechanistic redundancy separation. At inference, we estimate pointwise source redundancy from the SupGK-aligned embeddings. Because the alignment objective encourages $z_1^R \approx z_2^R$, the shared head h_R predicts Y from target-predictive representations of both modalities. Applying the pointwise $I_{\min}$ principle to these representations, we define

$$r_{\text{source}}^{(j)} = r_{\text{GK}}^{(j)} = \min\left(i(z_1^{R(j)}; y^{(j)}), i(z_2^{R(j)}; y^{(j)})\right), \tag{8}$$

where $i(z_m^{R(j)}; y^{(j)}) = \log p(y^{(j)} \mid z_m^{R(j)}) - \log p(y^{(j)})$, and $p(y \mid z_m^R)$ is estimated by h_R. Since source redundancy is intended to capture the part of the total predictive redundancy that is attributable to information shared between the modalities, we enforce the consistency relation $r^{(j)} \leftarrow \max\left(r^{(j)}, r_{\text{source}}^{(j)}\right)$. We then define $r_{\text{mech}}^{(j)} = r^{(j)} - r_{\text{source}}^{(j)} \geq 0$. Pointwise uniqueness and synergy follow from Eq. 2.

4) Estimating distribution-level PID with source-mechanistic redundancy separation. The total redundancy is estimated by: $R = \frac{1}{N}\sum_{j=1}^{N} r^{(j)}$. The source redundancy is computed by: $R_{\text{source}} = \frac{1}{N}\sum_j r_{\text{source}}^{(j)}$, and the mechanistic redundancy is computed by $R_{\text{mech}} = R - R_{\text{source}}$. Other distribution-level quantities are obtained through the linear system in Eq.1. Additional details are in Appendix E.

4 Experiments

Experimental overview. We evaluate our method across a range of controlled and real-world settings. In Sec. 4.1, we assess both distribution-level and pointwise PID on synthetic data with known interaction structure. In Sec. 4.2, we focus on distribution-level PID estimation on high-dimensional inputs (e.g., images and speech). In Sec. 4.3, we compare our pointwise estimates to theoretically grounded PID values derived from Bertschinger et al. (2014). In Sec. 4.5, we study the relationship between pointwise PID and human attribution. Finally, in Sec. 4.4, we show that our method distinguishes source from mechanistic redundancy in controlled settings.

Ground truth. Ground-truth MI is computed as $I(X_1, X_2; Y) = H(Y) - H(Y \mid X_1, X_2)$, where $H(Y \mid X_1, X_2)$ is the cross-entropy of the joint model, obtained from $h \circ f(x_1, x_2)$. When interaction proportions are controlled (e.g., Sec. 4.1, Sec. 4.2), this allows approximating ground-truth PID.

Baselines. We compare against PID-Batch (Liang et al., 2023), a distribution-level estimator; Flow-PID (Zhao et al., 2025), a normalizing-flow-based method for distribution-level PID; and LSMI (Yang et al., 2025), a pointwise PID estimator.

Hyperparameters. We use the same λ selection criterion across experiments: among $\lambda \in \{0, 1, 10, 25, 50, 100\}$, we select the smallest value achieving high modality agreement while retaining near-maximal source redundancy. This yields $\lambda = 10$ for AV-MNIST and $\lambda = 25$ for the clinical data. On AV-MNIST, where the ground-truth decomposition is known, the selected $\lambda = 10$ recovers 99.99% source attribution. While the estimated magnitude can depend on λ, the clinical decomposition is stable at $\lambda = 10$ (Appendix M). Kernel bandwidths are fixed to standard supervised contrastive values ($\tau = 0.1$, $\sigma_y = 0.1$).

4.1 Synthetic PID benchmark

Setup. We construct a controlled benchmark inspired by Liang et al. (2023), extended to a 4-class setting with explicit control over interaction types.

Pure interaction datasets. We first generate four datasets $\mathcal{D}_I$, $I \in \{R, U_1, U_2, S\}$, each isolating a single PID component: (i) U_1: only X_1 predicts y, (ii) U_2: only X_2 predicts y, (iii) R: both modalities contain the same predictive information, (iv) S: only the joint observation (X_1, X_2) predicts y via a nonlinear XOR-like mechanism (extended to 4 classes). Each dataset contains 20,000 samples.

Mixtures. We then construct mixed datasets by combining samples from the pure datasets *with the same label*. For each sample j, features are obtained as a convex combination of pure samples, with weights $w^{(j)}$ (e.g., $13\%\, r + 87\%\, s$) drawn from a Dirichlet distribution centered on a target ratio (e.g., $33\%\, R + 67\%\, S$). This produces datasets with controlled interaction proportions.

Controlled ground truth. The target mixture (e.g., $33\%\, R + 67\%\, S$) defines the distribution-level PID ($R = 0.33 \times I(X_1, X_2; Y)$, $S = 0.67 \times I(X_1, X_2; Y)$). The per-sample Dirichlet weights $w^{(j)}$ provide controlled generating interaction proportions. Since feature-space interpolation does

not necessarily translate into identical PID proportions, these weights are proxies for the intended pointwise interaction structure rather than exact PID ground truth. Appendix L quantifies this approximation analytically.

Table 1: Synthetic PID benchmark. Δ: distribution-level sum of absolute errors to ground truth ($\downarrow$). ρ: pointwise cosine similarity to the generating Dirichlet interaction weights ($\uparrow$). LSMI produces negative PID values in 8/12 scenarios (12 cells total). Full per-component breakdown in Appendix F.3. Statistical significance between LSMI and our method is assessed via one-sided Wilcoxon signed-rank tests over the 12 paired scenarios (*: $p < 0.05$; **: $p < 0.01$). Column headers list the interactions and their proportions (e.g., $R/U_2/S$ with $50/25/25$ denotes 50% R + 25% U_2 + 25% S).

	Pure				**Two-component mix**				**Three-component mix**				
	R	U_1	U_2	S	U_1/S	R/S	U_1/U_2	U_2/R	$U_1/U_2/R$	$R/U_2/S$	$R/U_1/S$	$U_1/U_2/S$	
mix (%)	100	100	100	100	33/67	33/67	33/67	33/67	50/25/25	50/25/25	25/50/25	25/25/50	**Avg**
Distribution-level error $\Delta\downarrow$													
Batch (distribution-only)	0.689	0.174	2.348	0.097	0.102	0.099	1.620	0.541	0.321	1.021	0.274	0.270	0.630
FlowPID (distribution-only)	0.116	0.174	0.145	0.230	0.182	0.160	0.623	0.548	0.426	0.342	0.296	0.275	0.293
LSMI (distribution)	0.023	0.048	0.080	0.048	0.070	**0.053**	0.296	0.248	0.174	0.292	0.234	0.099	0.139 $\pm$ 0.103
Ours (distribution)	**0.001**	**0.027**	**0.000**	**0.033**	**0.066**	0.071	**0.289**	**0.232**	**0.146**	**0.228**	**0.218**	**0.060**	**0.114**$\pm$ 0.103**
Pointwise cosine similarity $\rho\uparrow$													
LSMI (pointwise)	0.840	0.689	**0.816**	0.553	0.455	0.559	**0.537**	**0.717**	**0.504**	0.686	0.564	0.465	0.615$\pm$ 0.131
Ours (pointwise)	**0.841**	**0.77**	0.77	**0.815**	**0.734**	**0.77**	0.454	0.690	0.464	**0.730**	**0.705**	**0.667**	**0.701**$\pm$ 0.122*

Results. Table 1 summarizes distribution-level error (Δ) and pointwise cosine similarity (ρ) across the 12 synthetic scenarios. At the distribution level, our method achieves the lowest average error ($\Delta = 0.114 \pm 0.103$ vs. 0.139 ± 0.103 for LSMI), which is statistically significant, and outperforms LSMI on 11 out of 12 configurations. On pure regimes, recovery is near-exact ($\Delta \leq 0.033$). At the pointwise level, our method achieves higher average cosine similarity ($\rho = 0.701 \pm 0.122$ vs. 0.615 ± 0.131 for LSMI), with the largest gains on pure and three-component mixtures. In particular, performance improves substantially on fully synergistic settings and 3-way mixtures. LSMI remains competitive on two-component mixtures, but overall exhibits less consistent behavior across regimes.

4.2 Validation on Experiments with Known PID Ground Truth

We first evaluate whether the estimator recovers analytically known PID structures in high-dimensional multimodal settings. We consider AV-MNIST, where both modalities describe the same digit (pure source redundancy), and TriFeatures, where the target is designed to induce pure redundancy or pure uniqueness. **AV-MNIST.** We pair MNIST images with spoken digits from FSDD (Jackson et al., 2018) and predict the shared digit label. The ground-truth decomposition is redundancy-dominated: $R = I(X_1, X_2; Y) \approx 2.25$ nats and $U_1 = U_2 = S = 0$. Our estimator recovers this structure ($R = 2.11$, $U_1 = 0.10$, $U_2 = 0$, $S = 0.04$), outperforming LSMI ($\Delta = 0.28$ vs. 0.555). The remaining error is explained by classifier calibration: redundancy is bounded by the least calibrated modality, and the resulting slack appears as spurious unique information (Appendix A). **TriFeatures.** TriFeatures (Hermann & Lampinen, 2020) provides controlled multimodal pairs with known PID structure. Following Dufumier et al. (2025), we construct settings with either shared shape information (100% redundancy) or modality-specific texture information (100% uniqueness). Table 2 shows that our estimator accurately recovers all regimes, achieving near-zero error on pure redundancy and uniqueness cases, while LSMI introduces larger deviations and negative atoms.

Table 2: PID estimation of PID methods on AVMNIST and TriFeatures. BATCH is compared with pointwise methods (Ours and LSMI) while being restricted to distribution-only PID estimation.

	AV-MNIST					TriFeatures-100%R					TriFeatures-100%U_1					TriFeatures-100%U_2				
	R	U_1	U_2	S	$\Delta\downarrow$	R	U_1	U_2	S	$\Delta\downarrow$	R	U_1	U_2	S	$\Delta\downarrow$	R	U_1	U_2	S	$\Delta\downarrow$
Exact	2.25	0	0	0		2.255	0	0	0		0	2.15	0	0		0	0	2.01	0	
Ours (pointwise)	2.11	0.10	0	0.04	<u>0.28</u>	2.253	0.001	0	0.002	**0.005**	0	2.15	0	0	**0**	0	0	2.0	0.009	**0.019**
LSMI (pointwise)	1.96	0.26	0	-0.005	0.555	2.32	-0.022	-0.025	0	0.111	0	1.83	-0.01	0.056	0.386	0	-0.01	2.01	0.035	0.045
BATCH	2.28	0	0	0.07	**0.1**	2.350	0	0	0	0.105	0	2.27	0	0	0.12	0	0	2.28	2.28	2.45

4.3 Validation on Synthetic Logic Circuits

We evaluate our estimator on binary logic circuits (OR, AND, XOR), a standard benchmark whose exact partial information decomposition (PID) is analytically known (Bertschinger et al., 2014).

Following Liang et al. (2023), inputs are sampled as $x_1, x_2 \sim \text{Bernoulli}(0.5)$ with additive Gaussian noise ($\sigma = 0.1$). We compare against BATCH and the official implementation of LSMI.

Table 3: PID estimation on binary logic circuits (nats). Δ denotes the ℓ_1 error w.r.t. ground truth across (R, U_1, U_2, S). Flow-PID failed to converge (due to numerical instabilities in matrix inversion).

	OR					AND					XOR				
	R	U_1	U_2	S	$\Delta\downarrow$	R	U_1	U_2	S	$\Delta\downarrow$	R	U_1	U_2	S	$\Delta\downarrow$
Exact	0.215	0	0	0.347		0.215	0	0	0.347		0	0	0	0.693	
Ours (pointwise)	0.201	0	0.001	0.344	<u>0.018</u>	0.210	0.000	0.007	0.350	<u>0.015</u>	0	0	0	**0.693**	**0.000**
BATCH	0.208	0	0	0.345	**0.009**	0.213	0.005	0	0.351	**0.011**	0	0	0	0.692	0.001
LSMI (pointwise))	0.340	0.001	0	0.352	0.131	0.289	0.061	0.063	0.279	0.266	0	0	0	0.693	<u>0.000</u>

Global PID accuracy. Table 3 shows that our estimator closely matches the analytical decomposition on all three circuits. On OR and AND, it recovers the correct redundancy with negligible unique information. Both methods correctly identify XOR as purely synergistic. LSMI substantially overestimates redundancy on OR and produces spurious unique information on AND. These results demonstrate that the proposed estimator preserves the accuracy of distribution-level PID.

Beyond distribution-level PID. Unlike BATCH, our method also produces a decomposition for every individual sample. Table 4 illustrates this on the OR gate.

Pointwise structure of OR. Table 4 decomposes the OR gate into its four input configurations, revealing sample-level interaction patterns.

Table 4: Pointwise PID per case for OR.

Cluster	u_1	u_2	r	s
$(0,0) \to 0$	0.023	0.036	0.634	0.676
$(1,0) \to 1$	0.291	-0.398	0	0.399
$(0,1) \to 1$	-0.385	0.292	0	0.387
$(1,1) \to 1$	0.001	0.001	0.291	0.001

$(1,1) \to 1$**: pure redundancy.** Each modality alone guarantees the output. The estimator assigns nearly all information to redundancy ($r \approx 0.29$).

$(0,0) \to 0$**: redundancy with synergy.** Each modality provides identical but insufficient evidence ($p(y = 0|x_i = 0) = 0.5$). Only their joint observation disambiguates the output, leading to redundancy ($r = 0.634$) and synergy ($s = 0.676$).

$(1,0)$ **and** $(0,1)$**: misleading modalities.** These asymmetric cases exhibit *negative unique information*. For ex., in $(0,1)$, observing $x_1 = 0$ reduces confidence in the correct class ($p(y = 1|x_1 = 0) = 0.5 < p(y = 1) = 0.75$), yielding $u_1 < 0$. The second modality enables predicting $y = 1$ with positive unique information, while synergy compensates for the misleading modality.

4.4 Disentangling Source vs. Mechanistic Redundancy

Existing PID estimators do not distinguish source redundancy (shared input information) from mechanistic redundancy (redundancy induced solely by the prediction rule). We evaluate both regimes using shared-source datasets (Synthetic, AV-MNIST, TriFeatures, Food101) and independent-input tasks (logic circuits and digit summation).

Digit summation. We pair MNIST and spoken digits to predict $\mathbb{1}[d_{\text{img}} + d_{\text{aud}} > c]$ ($c \in 6, 7, 8$). Since inputs are sampled independently, any redundancy is purely mechanistic.

Results. Table 5 reports (i) the ground-truth proportion of source redundancy, and (ii) the proportion identified by our estimator. It shows that the proposed estimator correctly identifies both regimes. Source redundancy is recovered almost entirely on shared-source datasets (typically $> 95\%$ SRC), while collapsing to 0% SRC on logic circuits and digit summation despite non-zero redundancy (corroborated analytically in Appendix N); an intermediate case is discussed in App. L.1). On Food101 (no reference), the estimator returns an intermediate split (57% SRC), consistent with recipes and images sharing the dish while retaining modality-specific cues.

Table 5: Estimating if the redundancy is *source* (left) or *mechanistic* (right). ✓: correct identification.

Source redundancy

Setting	True	Pred.
Synthetic (R > 0)	100 % SRC	≈100% SRC ✓
AV-MNIST / TriFeat	100% SRC	99–100% SRC ✓
Food101	unknown	57% SRC n/a

Mechanistic redundancy

Setting	True	Pred.
Logic (OR/AND)	0% SRC	0% SRC ✓
Digit sum (c=6, 7, 8)	0% SRC	0% SRC ✓

4.5 Evaluating Pointwise PID Against Human Attribution

The OR circuit exposes a systematic discrepancy between pointwise PID and human attribution. When one modality is locally misleading, the decomposition assigns it negative unique information and compensates through synergy, whereas humans give it no credit. We test whether this *compensation effect* accounts for a predictable share of the disagreement with human annotations. For this comparison only, we apply a parameter-free readout that, whenever $u_m < 0$ and $u_{m'} \geq 0$, sets $s \leftarrow s - |u_m|$ and $u_m \leftarrow 0$, thereby preserving the sum of the PID atoms without modifying the underlying estimator. We report Pearson correlation with the annotations of Yang et al. (2025); Liang et al. (2023), for which chance level is 0. Our three-seed averaged estimates agree more closely with human attribution than LSMI, both before the readout (0.441 vs. 0.358) and, more substantially, after it (0.742 vs. 0.397). The readout improves both estimators, but the improvement is much larger for ours (+0.301 vs. +0.039). Importantly, the effect is strongest on datasets where cross-modal incongruity is central to the task, and negligible on Food101, where the modalities generally provide consistent evidence. This pattern supports the interpretation that negative unique information captures a systematic form of cross-modal conflict: pointwise PID represents this conflict through negative unique information compensated by synergy, whereas human attribution does not assign credit to the misleading modality. The readout therefore provides a useful diagnostic of where and why pointwise PID can diverge from human attribution, without modifying the proposed estimator.

Table 6: Pearson correlation with human annotations before and after the compensation readout.

	Ours		LSMI	
	Before	After	Before	After
Food101	0.957	0.961	0.898	0.919
MOSEI	-0.415	0.454	-0.029	0.432
UR-Funny	0.772	0.927	0.857	0.631
MUStARD	0.451	0.626	-0.295	-0.393
Avg.	0.441	**0.742**	0.358	0.397

5 What Does Speech Model Capture in Huntington's Disease?

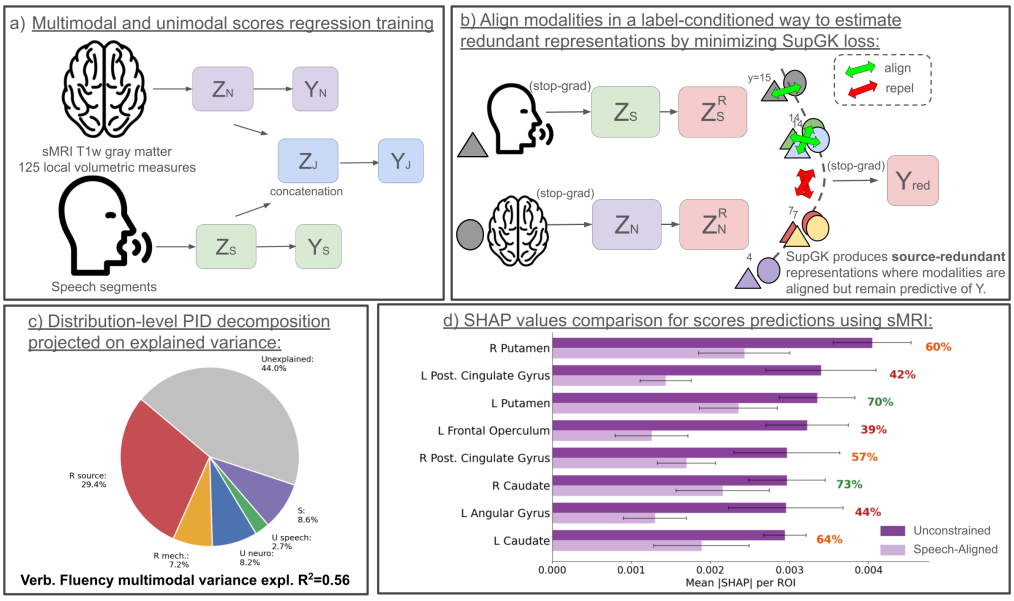


Figure 1: **Multimodal PID analysis of speech and neuroanatomical biomarkers in Huntington's disease. (a)** Unimodal and multimodal regression: clinical scores are predicted from sMRI gray matter volumes (Z_N) (unconstrained), wav2vec2 speech representations (Z_S), and their concatenation (Z_J). **(b)** Supervised Gács–Körner (SupGK) alignment: modality-specific representations are projected into a shared space; a shared head predicts scores from stop-gradient inputs, isolating the target-predictive information shared between the NDSNet representation and neuroanatomy. **(c)** Distribution-level PID projected onto explained variance for lexical fluency ($R^2 = 0.560$). **(d)** SHAP feature attribution comparison between unconstrained and speech-aligned predictors for the top contributing ROIs. Retention percentages indicate the fraction of each region's importance preserved under speech alignment.

Speech is an emerging low-cost biomarker of Huntington's disease (HD) (Le Moine Veillon et al., 2026), but what a speech-based predictor of lexical fluency captures remains unclear: the cortico-striatal motor degeneration that causes dysarthria (Rusz et al., 2014; Skodda et al., 2014), or the higher-order cognitive decline measured by clinical scores? Source redundancy answers this directly: aligning neuroanatomy to speech retains only the brain information that speech also carries. If speech mainly reflects motor degeneration, striatal regions should keep their importance under alignment, and associative cortical regions should lose it.

Setup. Building on NDSNet (Le Moine Veillon et al., 2026), we train unimodal and multimodal regression heads on 341 patients (855 visits, 211 with both speech and sMRI) to predict 12 clinical scores, including lexical fluency (words starting with "P" in 1 min), a measure of executive function and lexical retrieval (Henry & Crawford, 2004; Troyer et al., 1997). Speech is encoded with wav2vec2 (Baevski et al., 2020) and neuroanatomy from 125 gray-matter ROIs (Gaser et al., 2024), over five participant-independent 60/15/25 splits (details in App. I). Unimodal and multimodal score regression training is depicted in Fig. 1a. We then learn source-redundant representations Z_S^R and Z_N^R via SupGK (Fig. 1b). Fig. 1c) projects the PID onto explained variance R^2 [2]. And in Fig. 1d), we compare the SHAP attributions (Lundberg & Lee, 2017) of between the neuroanatomical unconstrained predictor $Y_N = h_N(f_N(X_N))$ (Fig. 1a, purple) and speech-aligned predictor $Y_N^R = h_N^R(Z_N^R)$. Fig. 1d) shows SHAP values of the top 10 neuroanatomical regions in unconstrained and speech-aligned regression settings. Details about the link between R^2 and PID are in Appendix Sec.G.

Distribution-level PID (Fig. 1c). Lexical fluency achieves $R^2 = 0.560$ jointly, with source redundancy at 29.4%, unique neuroanatomical information at 8.2%, and synergy at 8.6%. The moderate source redundancy indicates that part of the target-predictive information is shared between the NDSNet representation and neuroanatomy, but a substantial fraction is not. Under speech alignment, the variance explained drops from $R^2 = 44.8\%$ for the unconstrained predictor $Y_N = h_N(f_N(X_N))$ to $R^2 = 29.4\%$ for the speech-aligned predictor $Y_N^R = h_N^R(Z_N^R)$, a loss of 15.4 percentage points.

Motor versus cognitive substrates of lexical fluency. (Fig. 1d). Retention is the fraction of a region's importance that survives speech alignment: high means its contribution is shared with speech, low means it predicts lexical fluency through information speech does not carry. These retentions are associational: they quantify target-predictive information shared between the NDSNet representation and neuroanatomy, not causative effects. Speech alignment preserves the importance of striatal motor regions while attenuating regions supporting higher-order cognition. The putamen retains 60–70$\%$ of its predictive importance: as the principal hub of the basal ganglia motor circuit (Alexander et al., 1986) and the structure most affected in Huntington's disease (Harris et al., 1992), its degeneration drives both dysarthria (Rusz et al., 2014) and lexical fluency impairment. The caudate retains 64–73$\%$, consistent with its contribution to both cognitive and motor decline in the disease (Aylward, 2013). In contrast, the posterior cingulate (42–57$\%$), angular gyrus (44%) and frontal operculum (39%) exhibit markedly lower retention. These regions support executive control, lexical selection, and episodic memory rather than speech motor production (Leech & Sharp, 2014; Li, 2021; Seghier, 2013; Binder et al., 2009; Baldo et al., 2006), suggesting that their contribution to lexical fluency is independent of the acoustic information captured by NDSNet. A permutation control confirms this pattern reflects genuine speech–neuroanatomy correspondence: random pairings reduce source redundancy to 0% and regional retention to $6 \pm 1\%$ (App. M).

Takeaway. Combining source-redundant representations with SHAP reveals not only *that* speech and neuroanatomy share predictive information in HD Verbal Fluency deficit, but *where* in the brain it lies: in striatal motor circuitry, whereas associative cortical contributions remain speech-independent.

6 Conclusion

We presented a pointwise PID framework that requires only predictive models and separates redundancy into source and mechanistic components via a supervised Gács–Körner objective. Across synthetic and controlled benchmarks, it matches distribution-level estimators while providing signed sample-level decompositions. The approach remains dependent on predictive-model calibration but avoids the high-dimensional input-density estimation required by LSMI (Appendix A). In Huntington's disease, the learned source-redundant representations enable interpretable SHAP analyses of modality-shared and modality-specific neuroanatomical signals, while a cascading extension to five modalities demonstrates applicability to Alzheimer's progression (Appendix J).

[2] We explain how we decompose the variance explained from the PID decomposition in Appendix G.

AI USE STATEMENT

We used generative AI tools for writing assistance and for code assistance. For writing, AI tools were used to draft and edit portions of the appendix, to suggest phrasing and improve clarity in the main text, and to help position the work relative to prior literature; all technical claims, framing, and interpretation are the authors' own, and every AI-suggested statement was verified against the cited sources by the authors. For code, AI tools assisted in debugging existing analysis scripts; all such code was read, tested, and validated by the authors, and results were reproduced independently of AI assistance. We did not use generative AI tools for research ideation, for the design of the proposed method, or for the analysis and interpretation of the clinical results. We have reviewed all AI-assisted work and take responsibility for the final content of this paper, including all text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work involves human subjects. The Huntington's disease cohort comprises longitudinal clinical assessments, speech recordings, and structural MRI collected in the REPAIRHD (NCT03119246) and BIOHD (NCT01412125) studies, and the Alzheimer's analysis uses the publicly released TADPOLE/ADNI dataset. All participants gave informed consent under the protocols of the respective studies, which received approval from the relevant ethics committees. Speech recordings are identifying by nature and may carry sensitive information beyond clinical content; in our experiments, all data were pseudonymised and stored on private, access-controlled servers, and no raw audio, imaging, or individual-level data are released. Only code is released.

REPRODUCIBILITY STATEMENT

Anonymised source code for the estimator and for all synthetic and public-data experiments is available at the repository linked in the abstract. The synthetic benchmark generation, including latent structure, mixture construction and the four-class synergy mechanism, is specified in Appendix F.1, and Appendix L analytically evaluates the relationship between the generating interaction proportions and the resulting pointwise PID. Architectures, optimizers, and training schedules for every dataset, together with our configuration of the LSMI baseline, are given in Appendix H; hyperparameter roles and the label-free selection rule for the alignment weight are described in Appendix M.1. The two derivations supporting the method are given in full in Appendices B and C, with their approximations stated explicitly, and the assumptions and scope of Proposition 1 are set out in Appendix D. Experiments on the logic circuits, AV-MNIST, TriFeatures, Food101 and the public ADNI/TADPOLE cohort can be reproduced from the released code; the Huntington's disease data are not publicly available for the privacy reasons given in the ethics statement, but the preprocessing pipeline, cohort composition and encoder configuration are documented in Appendix I. All experiments run on a single H100 GPU.

REFERENCES

Garrett E Alexander, Mahlon R DeLong, and Peter L Strick. Parallel organization of functionally segregated circuits linking basal ganglia and cortex. *Annual Review of Neuroscience*, 9(1):357–381, 1986.

EH et al. Aylward. Regional atrophy associated with cognitive and motor function in prodromal Huntington disease. *J Huntingtons Dis.*, 2(4):477–89, 2013.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pp. 12449–12460, 2020.

Juliana V Baldo, Sophie Schwartz, David Wilkins, and Nina F Dronkers. Role of frontal versus temporal cortex in verbal fluency as revealed by voxel-based lesion symptom mapping. *Journal of the International Neuropsychological Society*, 12(6):896–900, 2006.

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.

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

Jeffrey R Binder, Rutvik H Desai, William W Graves, and Lisa L Conant. Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral Cortex*, 19(12):2767–2796, 2009.

Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. Towards multimodal sarcasm detection: An obviously perfect paper. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, Florence, Italy, 2019. Association for Computational Linguistics.

Benoit Dufumier, Pietro Gori, Julie Victor, Antoine Grigis, Michele Wessa, Paolo Brambilla, Pauline Favre, Mircea Polosan, Colm McDonald, Camille Marie Piguet, and Edouard Duchesnay. Contrastive learning with continuous proxy meta-data for 3D MRI classification. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2021.

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

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. Understanding dataset difficulty with $\mathcal{V}$-usable information. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 5988–6008. PMLR, 2022.

Conor Finn and Joseph T. Lizier. Pointwise partial information decomposition using the specificity and ambiguity lattices. *Entropy*, 20(4):297, 2018.

Peter Gács and János Körner. Common information is far less than mutual information. *Problems of Control and Information Theory*, 2(2):149–162, 1973.

Christian Gaser, Robert Dahnke, Paul M. Thompson, Florian Kurth, Eileen Luders, and The Alzheimer's Disease Neuroimaging Initiative. Cat: a computational anatomy toolbox for the analysis of structural mri data. *GigaScience*, 13, 2024.

Virgil Griffith, Edwin K. P. Chong, Ryan G. James, Christopher J. Ellison, and James P. Crutchfield. Intersection information based on common randomness. *Entropy*, 16(4):1985–2000, 2014. doi: 10.3390/e16041985.

Malte Harder, Christoph Salge, and Daniel Polani. A bivariate measure of redundant information. *Physical Review E*, 87(1):012130, 2013.

Gregory J Harris, Godfrey D Pearlson, Charles E Peyser, Elizabeth H Aylward, Janet Roberts, Patrick E Barta, Gary A Chase, and Susan E Folstein. Putamen volume reduction on magnetic resonance imaging exceeds caudate changes in mild Huntington's disease. *Annals of Neurology*, 31(1):69–75, 1992.

Julie D Henry and John R Crawford. A meta-analytic review of verbal fluency performance following focal cortical lesions. *Neuropsychology*, 18(2):284–295, 2004.

Katherine Hermann and Andrew Lampinen. What shapes feature representations? exploring datasets, architectures, and training. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9995–10006, 2020.

Robin A. A. Ince. Measuring multivariate redundant information with pointwise common change in surprisal. *Entropy*, 19(7):318, 2017a.

Robin A. A. Ince. The partial entropy decomposition: Decomposing multivariate entropy and mutual information via pointwise common surprisal. *arXiv preprint arXiv:1702.01591*, 2017b.

Clifford R Jack Jr, David A Bennett, Kaj Blennow, Maria C Carrillo, Billy Dunn, Samantha Budd Haeberlein, David M Holtzman, William Jagust, Frank Jessen, Jason Karlawish, Enchi Liu, Jose Luis Molinuevo, Thomas Montine, Creighton Phelps, Katherine P Rankin, Christopher C Rowe, Philip Scheltens, Eric Siemers, Heather M Snyder, and Reisa Sperling. NIA-AA research framework: Toward a biological definition of Alzheimer's disease. *Alzheimer's & Dementia*, 14 (4):535–562, 2018.

Zohar Jackson, César Souza, Jason Flaks, Yuxin Pan, Hereman Nicolas, and Adhish Thite. Free spoken digit dataset (fsdd). *Zenodo*, 2018.

Robert Jenssen, José C. Príncipe, Deniz Erdogmus, and Torbjørn Eltoft. The cauchy–schwarz divergence and parzen windowing: Connections to graph theory and mercer kernels. *Journal of the Franklin Institute*, 343(6):614–629, 2006.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Michael Kleinman, Alessandro Achille, Stefano Soatto, and Jonathan C. Kao. Gács-körner common information variational autoencoder. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Susan M Landau, Danielle Harvey, Cindee M Madison, Robert A Koeppe, Eric M Reiman, Norman L Foster, Michael W Weiner, and William J Jagust. Associations between cognitive, functional, and FDG-PET measures of decline in AD and MCI. *Neurobiology of Aging*, 32(7):1207–1218, 2011.

Clément Le Moine Veillon, S. Fraisse, M. Lunven, A. Fabre, K. Youssov, G. Morgado, H. Titeux, T. Le Ludec, C. Goizet, C. Fougeron, R. Massart, and Anne-Catherine Bachoud-Lévi. A deep learning model for speech-based prediction of clinical scores in people with huntington's disease: a longitudinal study with cross-sectional replication. *The Lancet Digital Health*, 2026.

Robert Leech and David J Sharp. The role of the posterior cingulate cortex in cognition and disease. *Brain*, 137(1):12–32, 2014.

D et al. Li. Abnormal functional connectivity of posterior cingulate cortex correlates with phonemic verbal fluency deficits in major depressive disorder. *Frontiers in Neurology*, 12:724874, 2021.

Yao Li, Gaurav Dwivedi, Wen Huang, Melissa L. Kemp, and Yingfei Yi. Quantification of degeneracy in biological systems for characterization of functional interactions between modules. *Journal of Theoretical Biology*, 302:29–38, 2012.

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

Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 17612–17625, 2022.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*, 2019.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11976–11986, 2022.

Joseph T. Lizier, Benjamin Flecker, and Paul L. Williams. Towards a synergy-based approach to measuring information modification. In *2013 IEEE Symposium on Artificial Life (ALIFE)*, pp. 43–51. IEEE, 2013.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Andrea I. Luppi, Fernando E. Rosas, Pedro A. M. Mediano, David K. Menon, and Emmanuel A. Stamatakis. Information decomposition and the informational architecture of the brain. *Trends in Cognitive Sciences*, 28(4):352–368, 2024.

Tobias Mages and Christian Rohner. Decomposing and tracing mutual information by quantifying reachable decision regions. *Entropy*, 25(7):1014, 2023.

Răzvan V Marinescu, Neil P Oxtoby, Alexandra L Young, Esther E Bron, Arthur W Toga, Michael W Weiner, Frederik Barkhof, Nick C Fox, Polina Golland, Stefan Klein, and Daniel C Alexander. TADPOLE challenge: Accurate Alzheimer's disease prediction through crowdsourced forecasting of future data. In *Predictive Intelligence in Medicine (PRIME)*, volume 11843 of *Lecture Notes in Computer Science*, pp. 1–10, 2019.

Sharon Oviatt, Rebecca Lunsford, and Rachel Coulston. Individual differences in multimodal integration patterns: What are they and why do they exist? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '05)*, pp. 241–249. ACM, 2005.

Ari Pakman, Amin Nejatbakhsh, Dar Gilboa, Abdullah Makkeh, Luca Mazzucato, Michael Wibral, and Elad Schneidman. Estimating the unique information of continuous variables. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Emanuel Parzen. On estimation of a probability density function and mode. In *The Annals of Mathematical Statistics*, volume 33, pp. 1065–1076, 1962.

Giuseppe Pica, Eugenio Piasini, Daniel Chicharro, and Stefano Panzeri. Invariant components of synergy, redundancy, and unique information among three variables. *arXiv preprint arXiv:1706.08921*, 2017. URL `https://arxiv.org/abs/1706.08921`.

Georg Pichler, Pierre Colombo, Malik Boudiaf, Günther Koliander, and Pablo Piantanida. A differential entropy estimator for training neural networks. In *International Conference on Machine Learning*, pp. 17691–17715. PMLR, 2022.

José C. Príncipe. *Information Theoretic Learning: Renyi's Entropy and Kernel Perspectives*. Information Science and Statistics. Springer, 2010.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38 th International Conference on Machine Learning (ICML)*, 2021.

Murray Rosenblatt. Remarks on some nonparametric estimates of a density function. In *The Annals of Mathematical Statistics*, volume 27, pp. 832–837, 1956.

Jan Rusz, Jiri Klempir, Tereza Tykalova, Eva Baborova, Roman Cmejla, Evzen Ruzicka, and Jan Roth. Characteristics and occurrence of speech impairment in Huntington's disease: possible influence of antipsychotic medication. *Journal of Neural Transmission*, 121(12):1529–1539, 2014.

Scott A Schobel, Giuseppe Palermo, Peggy Auinger, Jeffrey D Long, Shiyang Ma, Omar S Khwaja, Dylan Trundell, et al. Motor, cognitive, and functional declines contribute to a single progressive factor in early HD. *Neurology*, 89(24):2495–2502, 2017.

Mohamed L Seghier. The angular gyrus: multiple functions and multiple subdivisions. *The Neuroscientist*, 19(1):43–61, 2013.

Sabine Skodda, Uwe Schlegel, Rainer Hoffmann, and Carsten Saft. Impaired motor speech performance in Huntington's disease. *Journal of Neural Transmission*, 121(4):399–407, 2014.

Giulio Tononi, Olaf Sporns, and Gerald M. Edelman. Measures of degeneracy and redundancy in biological networks. *Proceedings of the National Academy of Sciences*, 96(6):3257–3262, 1999.

Angela K Troyer, Morris Moscovitch, and Gordon Winocur. Clustering and switching as two components of verbal fluency: evidence from younger and older healthy adults. *Neuropsychology*, 11(1):138–146, 1997.

Praveen Venkatesh, Corbett Bennett, Sam Gale, Tamina K. Ramirez, Greggory Heller, Séverine Durand, Shawn Olsen, and Stefan Mihalas. Gaussian partial information decomposition: Bias correction and application to high-dimensional data. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Feng Wang and Huaping Liu. Understanding the behaviour of contrastive loss. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2495–2504, 2021.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning (ICML)*, 2020.

Paul L. Williams and Randall D. Beer. Nonnegative decomposition of multivariate information. In *https://arxiv.org/abs/1004.2515*, 2010.

Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. A theory of usable information under computational constraints. In *International Conference on Learning Representations (ICLR)*, 2020.

Lecun Y., Bottou L., Bengio Y., and Haffner P. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, volume 86, pp. 2278–2324, 1998.

Zequn Yang, Hongfa Wang, and Di Hu. Efficient quantification of multimodal interaction at sample level. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*. PMLR, 2025.

Yedi Zhang, Peter E. Latham, and Andrew M. Saxe. Understanding unimodal bias in multimodal deep linear networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, 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 (NeurIPS)*, 2025.

A CALIBRATION OF THE PREDICTIVE MODELS

Our estimator uses $\log p(y \mid x_m)$ rather than the arg-max, so its accuracy depends on calibration rather than classification performance. Table 7 reports held-out accuracy, cross-entropy and expected calibration error (ECE, 15 equal-width confidence bins) for the Stage 1 heads on AV-MNIST.

Relation to $\mathcal{V}$-information. The pointwise quantity $i(x_m; y) = \log p(y \mid x_m) - \log p(y)$, computed from a trained predictor, coincides with the pointwise $\mathcal{V}$-information (PVI) of Ethayarajh et al. (2022). Its expectation, like the reference quantity $I(X_1, X_2; Y) = H(Y) - H(Y \mid X_1, X_2)$ obtained from the joint model's cross-entropy, is the predictive $\mathcal{V}$-information $I_{\mathcal{V}}(X \to Y)$ of Xu et al. (2020). All atoms reported in this paper should therefore be read as decompositions of the information usable by the fitted predictors, and their accuracy depends on how well these predictors are calibrated.

The two modalities reach near-identical accuracy, yet the audio head has twice the cross-entropy and roughly five times the calibration error. MNIST supplies far more training images than FSDD supplies recordings, and the audio samples are not perfectly i.i.d. despite separating speakers between train and test.

This propagates to the decomposition in a specific direction. Since $r = \min_m i(x_m; y)$, the less calibrated modality caps the estimated redundancy; the deficit then appears as unique information in the better-calibrated modality, which is exactly the pattern observed on AV-MNIST ($R = 2.11$ against an exact 2.25, with $U_1 = 0.10$). Distribution-level estimators that do not pass through per-sample predictive distributions, such as BATCH, are not subject to this effect, which accounts for their lower Δ on this dataset; they are correspondingly unable to produce any pointwise decomposition, and fail elsewhere (TriFeatures-100% U_2, $\Delta = 2.45$).

We regard this as a genuine limitation of prediction-based pointwise PID. It is not one our method incurs more than the alternatives: LSMI requires the same conditional distributions $p(y \mid x_m)$ and,

in addition, marginal entropies $h(x_m)$ estimated in high-dimensional input space, whose guarantees likewise hold only in expectation. Our formulation removes the second requirement, leaving calibration of a standard supervised classifier as the sole estimation dependency.

Table 7: Calibration of the Stage 1 predictive heads on the AV-MNIST test set.

	Accuracy	Cross-entropy	ECE
Visual	0.9633	0.1132	0.0055
Audio	0.9658	0.2387	0.0262
Joint	0.9921	0.0528	0.0053

B Deriving CLIP from a Relaxed Gács-Körner Objective

The Gács-Körner common information Gács & Körner (1973) formalizes the notion of information that is genuinely *common* to two random variables. It is defined as the maximum entropy of a random variable Z that can be recovered as a deterministic function of each source individually:

$$C_{\mathrm{GK}}(X_1; X_2) = \max_{Z} H(Z) \quad \text{s.t.} \quad Z = f(X_1) = g(X_2) \ \text{ a.s.} \tag{9}$$

We show that the symmetric InfoNCE loss optimized by CLIP Radford et al. (2021) can be derived as a Lagrangian relaxation of a Gács-Körner-type entropy maximization problem, under a Gaussian kernel density approximation.

Setup. Consider two modalities X_1, X_2 and learned encoders f_θ, g_ϕ producing ℓ_2-normalized embeddings $Z_1 = f_\theta(X_1)$ and $Z_2 = g_\phi(X_2)$. We formulate the following constrained optimization, analogous to the Gács-Körner objective equation 9:

$$\max_{\theta,\phi} \ H(Z_1) + H(Z_2) \quad \text{s.t.} \quad \|z_1^{(i)} - z_2^{(i)}\|_2^2 = 0 \ \ \forall\, i \tag{10}$$

which seeks maximum-entropy embeddings that agree on matched pairs.

Step 1: Cross-entropy decomposition. For any two distributions p_1, p_2 over the same space, the entropy of p_1 can be written as:

$$H(Z_1) = -\mathbb{E}_{p_1}[\log p_1(z)] = -\mathbb{E}_{p_1}[\log p_2(z)] - D_{\mathrm{KL}}(p_1 \| p_2) \tag{11}$$

where p_1, p_2 denote the marginal distributions of Z_1, Z_2 respectively. This expresses the entropy as the cross-entropy from p_1 to p_2, minus the KL divergence between them.

Step 2: Gaussian KDE approximation. We approximate p_2 using a Gaussian kernel density estimator with bandwidth σ over the batch $\{z_2^{(j)}\}_{j=1}^N$:

$$\hat{p}_2(z) = \frac{1}{N \cdot c_\sigma} \sum_{j=1}^{N} \exp\left(-\frac{\|z - z_2^{(j)}\|_2^2}{2\sigma^2} \right) \tag{12}$$

where c_σ is the normalization constant of the Gaussian kernel (independent of z and j). Substituting into the cross-entropy term:

$$-\mathbb{E}_{p_1}[\log \hat{p}_2(z)] = -\frac{1}{N} \sum_{i=1}^{N} \log \sum_{j=1}^{N} \exp\left(-\frac{\|z_1^{(i)} - z_2^{(j)}\|_2^2}{2\sigma^2} \right) + \text{const.} \tag{13}$$

where the constant absorbs $\log(N \cdot c_\sigma)$.

Step 3: Unit-norm simplification. Since the embeddings are ℓ_2-normalized ($\|z_1^{(i)}\| = \|z_2^{(j)}\| = 1$), we have:

$$\|z_1^{(i)} - z_2^{(j)}\|_2^2 = 2 - 2\, z_1^{(i)} \cdot z_2^{(j)} \tag{14}$$

Substituting into the exponential:

$$\exp\left(-\frac{\|z_1^{(i)} - z_2^{(j)}\|_2^2}{2\sigma^2}\right) = \exp\left(-\frac{1}{\sigma^2}\right) \cdot \exp\left(\frac{z_1^{(i)} \cdot z_2^{(j)}}{\sigma^2}\right) \tag{15}$$

The first factor is constant across j and cancels inside the log-sum-exp. Setting $\tau = \sigma^2$:

$$H(Z_1) \approx -\frac{1}{N}\sum_{i=1}^{N} \log \sum_{j=1}^{N} \exp\left(\frac{z_1^{(i)} \cdot z_2^{(j)}}{\tau}\right) - D_{\mathrm{KL}}(p_1 \| p_2) + \text{const.} \tag{16}$$

The temperature τ thus corresponds to the squared bandwidth of the kernel density estimator.

Step 4: Lagrangian relaxation. Symmetrizing the entropy for both $H(Z_1)$ and $H(Z_2)$, and forming the Lagrangian with multiplier λ for the alignment constraint, we obtain:

$$\begin{aligned}
\mathcal{L} &= -H(Z_1) - H(Z_2) + \lambda \sum_{i=1}^{N} \|z_1^{(i)} - z_2^{(i)}\|_2^2 \\
&\approx \frac{1}{N}\sum_{i=1}^{N} \log \sum_{j=1}^{N} \exp\left(\frac{z_1^{(i)} \cdot z_2^{(j)}}{\tau}\right) + \frac{1}{N}\sum_{i=1}^{N} \log \sum_{j=1}^{N} \exp\left(\frac{z_2^{(i)} \cdot z_1^{(j)}}{\tau}\right) \\
&\quad + D_{\mathrm{KL}}(p_1 \| p_2) + D_{\mathrm{KL}}(p_2 \| p_1) + \lambda \sum_{i=1}^{N} \|z_1^{(i)} - z_2^{(i)}\|_2^2 + \text{const.}
\end{aligned} \tag{17}$$

Step 5: KL vanishing under alignment. When the alignment constraint is sufficiently enforced ($z_1^{(i)} \approx z_2^{(i)}$ for all i), the marginal distributions p_1 and p_2 become approximately equal, so $D_{\mathrm{KL}}(p_1 \| p_2) \approx 0$ and $D_{\mathrm{KL}}(p_2 \| p_1) \approx 0$. Furthermore, the alignment penalty for unit-norm embeddings simplifies:

$$\|z_1^{(i)} - z_2^{(i)}\|_2^2 = 2 - 2\, z_1^{(i)} \cdot z_2^{(i)} \tag{18}$$

so the Lagrangian term becomes (dropping constants):

$$\lambda \sum_{i} \|z_1^{(i)} - z_2^{(i)}\|^2 = -2\lambda \sum_{i} z_1^{(i)} \cdot z_2^{(i)} + \text{const.} \tag{19}$$

Step 6: Final loss. Combining all terms and absorbing 2λ into a single multiplier, we arrive at:

$$\min_{\theta,\phi} \underbrace{\frac{1}{N}\sum_{i=1}^{N} \log \sum_{j=1}^{N} \exp\left(\frac{z_1^{(i)} \cdot z_2^{(j)}}{\tau}\right) + \frac{1}{N}\sum_{i=1}^{N} \log \sum_{j=1}^{N} \exp\left(\frac{z_2^{(i)} \cdot z_1^{(j)}}{\tau}\right)}_{\text{entropy maximization}} - \underbrace{\frac{\lambda}{N}\sum_{i=1}^{N} \frac{z_1^{(i)} \cdot z_2^{(i)}}{\tau}}_{\text{alignment}} \tag{20}$$

This is precisely the symmetric InfoNCE loss optimized by CLIP. The first two terms are the log-partition functions (denominators of the softmax), and the alignment term corresponds to the positive-pair log-numerators. The temperature τ is the squared bandwidth of the Gaussian KDE used to approximate the embedding entropy. This reveals that CLIP can be interpreted as maximizing the entropy of a shared embedding space under a pairwise alignment constraint, which is a relaxed form of the Gács-Körner common information objective equation 4.

Related decompositions of InfoNCE have been derived before: Wang & Isola (2020) show that it splits asymptotically into an alignment term and a *unimodal* uniformity term on the hypersphere, analysing a given loss to expose its structure. Our construction runs in the opposite direction and is intrinsically multimodal: we start from the constrained problem in Eq. equation 10 and express the entropy of each modality as a cross-entropy against the *other* modality's density (Step 1), which is what produces the cross-modal log-partition over $z_1^{(i)} \cdot z_2^{(j)}$. The Gaussian-KDE entropy surrogate is common to both derivations; the constrained objective it is derived from, and its identification with Gács-Körner common information, are however, novel.

Assumptions and limitations. This derivation relies on three approximations: (i) the Gaussian KDE with bandwidth σ (equivalently temperature $\tau = \sigma^2$), (ii) Monte Carlo estimation of the entropy from a finite batch of size N, and (iii) the assumption that the KL divergence between the two marginal distributions vanishes under sufficient alignment. The first is a modeling choice inherent to any kernel-based entropy estimator; the second is standard and improves with batch size; the third holds exactly when the constraint is perfectly satisfied and approximately when it is relaxed via the Lagrangian penalty with large λ.

C Deriving SupGK from a Supervised Gács–Körner Objective

Following Appendix B, we derive the SupGK loss (Eq. equation 7) from the supervised Gács–Körner objective (Eq. equation 6). As in the CLIP derivation, the argument consists of (i) relaxing the hard alignment constraint through a Lagrangian, and (ii) approximating the information terms by kernel density estimation. The only difference is that the marginal entropy is approximated through the quadratic (Rényi-2) entropy, naturally leading to squared Gaussian kernels.

Setup. Let

$$Z_1^R = g_{\theta_1}(X_1), \qquad Z_2^R = g_{\theta_2}(X_2)$$

be ℓ_2-normalized embeddings. Let us recall the Eq. equation 6:

$$\max_{\theta_1,\theta_2} I(Z_1^R; Y) + I(Z_2^R; Y) \qquad \text{s.t.} \qquad z_1^{R(i)} = z_2^{R(i)} \ \ \forall i. \tag{21}$$

Denote

$$K_{ij} = \exp\left(-\frac{\|z_1^{R(i)} - z_2^{R(j)}\|^2}{2\tau}\right),$$

together with the positive and negative pair sets

$$\mathcal{P} = \{(i,j) : y_i = y_j\}, \qquad \mathcal{N} = \{(i,j) : y_i \neq y_j\}.$$

Step 1: Relaxing the alignment constraint. Writing

$$I(Z_m^R; Y) = H(Z_m^R) - H(Z_m^R|Y)$$

and replacing the hard equality constraint by a quadratic penalty gives

$$\mathcal{L}_{\text{align}} = \frac{\lambda}{N} \sum_i \|z_1^{R(i)} - z_2^{R(i)}\|^2,$$

using

$$\|u - v\|^2 = 2 - 2u^\top v$$

for unit-normalized embeddings.

The remaining problem is therefore to construct tractable surrogates for the marginal and conditional entropy terms.

Step 2: Conditional entropy. Using a Gaussian KDE,

$$\hat{p}(z^{R(i)}|y_i) = \frac{1}{|\mathcal{P}_i|} \sum_{j \in \mathcal{P}_i} K_{ij},$$

the re-substitution estimator of the conditional entropy is

$$-H(Z^R|Y) = \frac{1}{N} \sum_i \log \left(\frac{1}{|\mathcal{P}_i|} \sum_{j \in \mathcal{P}_i} K_{ij} \right).$$

Differentiating with respect to any same-class kernel gives

$$\frac{\partial(-H(Z^R|Y))}{\partial K_{ij}} = \frac{1}{N|\mathcal{P}_i|\hat{p}(z_i|y_i)} > 0.$$

Since

$$\frac{\partial(-H(Z^R|Y))}{\partial K_{ij}} > 0,$$

the quantity $-H(Z^R|Y)$ is monotonically increasing in every same-class kernel. Equivalently,

$$K_{ij} \uparrow \implies -H(Z^R|Y) \uparrow \implies H(Z^R|Y) \downarrow .$$

Therefore minimizing the conditional entropy amounts to increasing every same-class kernel toward its maximum value.

Our surrogate is

$$\ell(K_{ij}) = (K_{ij} - 1)^2.$$

Since Gaussian kernels satisfy

$$K_{ij} \in (0, 1],$$

its derivative is

$$\frac{\partial \ell}{\partial K_{ij}} = 2(K_{ij} - 1) \leq 0,$$

with equality only at $K_{ij} = 1$.

Hence

$$K_{ij} \uparrow \implies (K_{ij} - 1)^2 \downarrow .$$

Both objectives therefore induce the same optimization direction on every same-class kernel: they are simultaneously minimized by increasing K_{ij} toward 1.

Step 3: Marginal entropy. Unlike Shannon entropy, the quadratic entropy does not admit a decomposition through cross-entropy and KL divergence. Instead, using the Rényi-2 entropy

$$H_2(p) = -\log \int p^2,$$

the Cauchy–Schwarz inequality Jenssen et al. (2006); Príncipe (2010) gives:

$$\frac{1}{2}\big(H_2(Z_1^R) + H_2(Z_2^R)\big) = -\log \int p_1 p_2 - D_{\mathrm{CS}}(p_1, p_2), \tag{22}$$

where

$$D_{\mathrm{CS}}(p_1, p_2) \geq 0$$

is the Cauchy–Schwarz divergence, vanishing iff $p_1 = p_2$. Under the alignment constraint, $D_{\mathrm{CS}} \to 0$, so maximizing the marginal entropy amounts to maximizing $-\log \int p_1 p_2$. Since $-\log(\cdot)$ is a strictly decreasing function on $(0, \infty)$, this is equivalent to minimizing $\int p_1 p_2$ directly; we therefore drop the logarithm and work with the cross-information potential itself as a monotonic surrogate for the marginal entropy term.

$$\int p_1 p_2.$$

Estimating this quantity by Gaussian KDE yields

$$\int p_1 p_2 \approx \frac{1}{N^2} \sum_{i,j} K_{ij}^2,$$

since squaring a Gaussian kernel simply halves its bandwidth.

Splitting this sum gives

$$\sum_{i,j} K_{ij}^2 = \sum_{(i,j)\in\mathcal{P}} K_{ij}^2 + \sum_{(i,j)\in\mathcal{N}} K_{ij}^2.$$

The positive component is already driven to its optimum ($K_{ij} = 1$) by the conditional entropy term. Keeping it inside the marginal objective would therefore introduce the usual false-negative repulsion between samples sharing the same class Khosla et al. (2020). Following supervised contrastive learning, we therefore retain only the different-class contribution,

$$\frac{1}{|\mathcal{N}|} \sum_{(i,j)\in\mathcal{N}} K_{ij}^2,$$

which acts as the marginal entropy surrogate.

Step 4: Final objective. Combining the conditional entropy surrogate, the marginal entropy surrogate, and the relaxed alignment constraint yields

$$\mathcal{L}_{\mathrm{SupGK}} = \frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} (K_{ij} - 1)^2 + \frac{1}{|\mathcal{N}|} \sum_{(i,j)\in\mathcal{N}} K_{ij}^2 + \frac{\lambda}{N} \sum_i \|z_1^{R(i)} - z_2^{R(i)}\|^2,$$

which is Eq. equation 7.

The continuous case. For continuous targets, the class indicator is replaced by the Gaussian label kernel

$$S_{ij} = \exp\left(-\frac{|y_i - y_j|^2}{2\sigma_y^2}\right),$$

leading to the weighted objective

$$\frac{\sum_{ij} S_{ij}(K_{ij}-1)^2}{\sum_{ij} S_{ij}} + \frac{\sum_{ij}(1-S_{ij})K_{ij}^2}{\sum_{ij}(1-S_{ij})},$$

which is a kernelized relaxation of the discrete formulation. Since $S_{ij} \in [0,1]$, the positive and negative terms remain respectively monotonically decreasing and increasing functions of K_{ij}, so the optimization direction established above is preserved.

Assumptions. As in Appendix B, the derivation relies on three approximations:

(i) Gaussian KDE;

(ii) the Rényi-2 approximation of the marginal entropy through the cross-information potential;

(iii) exclusion of same-class pairs from the repulsion term to avoid false negatives.

The resulting objective is therefore a tractable surrogate sharing the same optimization direction as the relaxed supervised Gács–Körner objective.

D IDENTIFIABILITY OF SOURCE REDUNDANCY

We first define source redundancy at the population level, independently of its practical estimator. For two modalities X_1, X_2 and a target Y, let

$$R_{\mathrm{src}} := \max_{Z^R = f_1(X_1) = f_2(X_2) \text{ a.s.}} I(Z^R; Y), \tag{23}$$

i.e., the largest amount of target information carried by a representation that can be recovered deterministically from either modality.

The ideal supervised Gács–Körner objective of Section 3.2.1,

$$\max_{Z_1^R = Z_2^R} I(Z_1^R; Y) + I(Z_2^R; Y), \tag{24}$$

therefore has optimal value $2R_{\mathrm{src}}$. The practical SupGK loss in Eq. 7 is a tractable relaxation of this ideal objective.

Proposition 1. *Suppose that*

$$X_1 = (Z^C, U_1), \qquad X_2 = (Z^C, U_2),$$

where Z^C is the maximal deterministic common variable of X_1 and X_2 in the sense of Gács–Körner. Then, for any joint distribution of (Z^C, U_1, U_2, Y),

(i)

$$R_{\mathrm{src}} = I(Z^C; Y), \tag{25}$$

attained at $Z^R = Z^C$;

(ii)

$$R_{\mathrm{src}} \le \min_{m \in \{1,2\}} I(X_m; Y); \tag{26}$$

(iii) if $X_1 \perp X_2$, then

$$R_{\mathrm{src}} = 0, \tag{27}$$

even when both modalities are strongly predictive of Y.

Proof. For (i), any feasible Z^R is a deterministic common variable of X_1 and X_2. By maximality of Z^C, there exists a measurable function g such that $Z^R = g(Z^C)$ almost surely Gács & Körner (1973); Kleinman et al. (2023). The data-processing inequality therefore gives

$$I(Z^R; Y) \leq I(Z^C; Y).$$

Equality is attained by choosing $Z^R = Z^C$.

For (ii), since Z^C is a deterministic function of each X_m,

$$I(Z^C; Y) \leq I(X_m; Y),$$

for $m \in \{1, 2\}$.

For (iii), if $X_1 \perp X_2$, their maximal deterministic common variable Z^C is almost surely constant. Hence,

$$R_{\text{src}} = I(Z^C; Y) = 0.$$

□

Proposition 1 separates common information from source redundancy. The Gács–Körner variable Z^C contains all information deterministically shared by the two modalities, whereas

$$R_{\text{src}} = I(Z^C; Y)$$

measures only the part of this common information that is predictive of Y. In particular, when the modalities are independent, no non-trivial source-shared information exists; any predictive redundancy must therefore be mechanistic.

The representation attaining Eq. 23 is not unique. Any sufficient function $Z^R = g(Z^C)$ satisfying $I(Z^R; Y) = I(Z^C; Y)$ achieves the same value, including Z^C itself. Thus, Eq. 23 identifies the *amount* of source redundancy but does not uniquely determine a representation containing only target-relevant common information.

For representation-level analyses, among the maximizers of Eq. 23, the desirable solution is a minimal sufficient function of Z^C for Y. SupGK is explicitly designed to favor such a representation: the same-target term minimizes a surrogate of $H(Z^R \mid Y)$, removing variability that is not needed to predict the target, while the different-target term preserves informative variation and prevents collapse. Thus, among target-predictive common representations, SupGK favors representations that retain as little target-irrelevant common information as possible. We therefore interpret the learned representation as a near-minimal sufficient common representation, although we do not claim exact recovery of the population minimizer.

Finally, Proposition 1 characterizes source redundancy under the ideal Gács–Körner constraint

$$Z^R = f_1(X_1) = f_2(X_2) \qquad \text{a.s.}$$

In practice, SupGK replaces this hard constraint with the soft cross-modal alignment term of Eq. 7, yielding a trainable relaxation of the same population objective. This is particularly appropriate for noisy continuous modalities, for which exact deterministic commonness can be overly restrictive. The controlled experiments then assess whether this relaxation preserves the intended distinction between source and mechanistic redundancy.

E Details on PID estimations.

E.1 Distribution-level consistency constraints

We first average the signed pointwise atoms to obtain distribution-level estimates. These averages need not be non-negative, even with exact predictive probabilities. For the distribution-level comparisons, we therefore apply consistency corrections after aggregation.

Mutual information is estimated from predictive cross-entropies as $\widehat{I}_m = \widehat{H}(Y) - \widehat{H}(Y \mid X_m)$ and $\widehat{I}_{12} = \widehat{H}(Y) - \widehat{H}(Y \mid X_1, X_2)$. Because the predictive heads are fitted separately, their estimated conditional entropies may violate the ordering $\widehat{H}(Y \mid X_1, X_2) \leq \widehat{H}(Y \mid X_m) \leq \widehat{H}(Y)$.

As in Yang et al. (2025), we apply clipping constraints to enforce this ordering. We also bound the distribution-level redundancy by the smaller unimodal information, $R \leftarrow \min(R, \widehat{I}_1, \widehat{I}_2)$, and compute the remaining atoms as

$$U_m = \widehat{I}_m - R, \qquad S = \widehat{I}_{12} - \widehat{I}_1 - \widehat{I}_2 + R.$$

Following the RUS consistency procedure of Yang et al. (2025), if the resulting distribution-level synergy is slightly negative, we set $S \leftarrow 0$, transfer the deficit to R, and recompute U_1 and U_2 from the PID consistency equations. These adjustments enforce distribution-level non-negativity while preserving the estimated total information. They do not modify the pointwise quantities used in the sample-level analyses.

Example: the OR gate. For independent fair binary inputs and $Y = X_1 \vee X_2$, our pointwise definition gives redundancy $\log 2$ for $(0, 0)$, $\log(4/3)$ for $(1, 1)$, and zero for the two mixed configurations. Its average is therefore $\bar{R} \approx 0.2452$ nats, whereas each unimodal mutual information is approximately 0.2158 nats, yielding $\bar{U}_1 = \bar{U}_2 \approx -0.0294$ nats. The distribution-level correction caps redundancy at 0.2158 nats and recomputes the remaining atoms, giving $U_1 = U_2 = 0$ and $S \approx 0.3466$ nats, matching the reference decomposition for this circuit.

E.2 Consistency between total and source redundancy

Our distribution-level predictive redundancy is obtained by averaging the pointwise redundancy estimates,

$$R = \mathbb{E}[r] \approx \frac{1}{N} \sum_{j=1}^{N} r^{(j)}, \qquad r^{(j)} = \min\Big(i(x_1^{(j)}; y^{(j)}), i(x_2^{(j)}; y^{(j)})\Big).$$

Appendix D independently establishes that the population source redundancy satisfies

$$R_{\text{source}} \leq \min_{m \in \{1,2\}} I(X_m; Y).$$

Importantly, this does not imply $R_{\text{source}} \leq R$, since in general

$$\mathbb{E}[\min(i(X_1; Y), i(X_2; Y))] \leq \min_{m \in \{1,2\}} I(X_m; Y).$$

Nevertheless, source redundancy is operationally intended to represent the part of the total predictive redundancy attributable to information shared between the sources, with the remaining contribution defined as mechanistic redundancy. Because R and R_{source} arise from distinct operational estimators, their definitions do not by themselves guarantee $\widehat{R}_{\text{source}} \leq \widehat{R}$. To preserve the intended non-negative source–mechanistic decomposition, we therefore apply the consistency correction

$$\widehat{R} \leftarrow \max\Big(\widehat{R}, \widehat{R}_{\text{source}}\Big), \qquad \widehat{R}_{\text{mech}} = \widehat{R} - \widehat{R}_{\text{source}} \geq 0.$$

This correction should be understood as an estimator-level consistency constraint rather than as a consequence of the population result in Appendix D.

F Results on Synthetic PID benchmark in full details

F.1 Synthetic data generation details

Latent variables and modalities. Each sample is generated from latent Gaussian variables $z_1, z_2, z_c \sim \mathcal{N}(0_d, \sigma^2 I_d)$, encoding modality-specific (z_1, z_2) and shared (z_c) information, with $d = 100$. For each modality $m \in \{1, 2\}$, inputs are obtained via random linear projections:

$$x_m = [z_m, z_c]\, T_m, \tag{28}$$

where $T_m \in \mathbb{R}^{2d \times D}$ is a random matrix and $D = 200$.

Pure interaction datasets. We construct four datasets $\mathcal{D}_I$, $I \in \{U_1, U_2, R, S\}$, each with $20{,}000$ samples and labels $y \in \{0, 1, 2, 3\}$:

- **Uniqueness U_1:** y is generated solely from z_1; X_2 is non-informative.
- **Uniqueness U_2:** y depends only on z_2.
- **Redundancy R:** y depends only on z_c, so both modalities contain identical predictive information.
- **Synergy S:** y depends jointly on z_1 and z_2 via a nonlinear XOR-like mechanism, such that neither modality alone is predictive.

For U_1, U_2, and R, labels are obtained by applying a nonlinear transformation to the relevant latent variables followed by discretization into four classes. For S, we construct two independent nonlinear projections of z_1 and z_2, and combine them using an XOR-like rule extended to four classes, ensuring that predictive information is only accessible jointly.

Mixture datasets. To generate mixed-interaction datasets, we combine samples from the four pure datasets. For each sample j, we draw mixing weights:

$$w^{(j)} \sim \mathrm{Dir}(c \cdot \bar{w}), \tag{29}$$

where $\bar{w} \in \mathbb{R}^4$ is the target interaction ratio (e.g., $(0, 0, 0.33, 0.67)$ for $33\%\, R + 67\%\, S$) and $c = 10$ controls concentration.

For each component $I \in \{U_1, U_2, R, S\}$, we select a sample with the same label y and construct:

$$x_m^{(j)} = \sum_I w_I^{(j)}\, x_{m,I}^{(j)}. \tag{30}$$

This ensures that: (i) the label is preserved, (ii) the interaction structure varies continuously across samples.

Ground-truth interactions. At the distribution level, the PID components correspond to the target mixture $\bar{w}$. At the pointwise level, the Dirichlet weights $w^{(j)}$ define the true interaction proportions for each sample.

Datasets. We generate 4 pure datasets and 8 mixed datasets (four two-component and four three-component mixtures), each containing 20,000 samples split into 70/15/15 train/validation/test sets.

Synergy mechanism (4-class XOR). For the synergistic dataset S, the label is constructed such that neither modality alone is predictive, but their joint observation is.

We first extract two disjoint subsets of latent features from each modality-specific variable z_1 and z_2, and map them to scalar probabilities via nonlinear transformations:

$$p_1^{(a)}, p_2^{(a)} = \sigma(f_a(z_1)), \sigma(g_a(z_2)), \quad p_1^{(b)}, p_2^{(b)} = \sigma(f_b(z_1)), \sigma(g_b(z_2)), \tag{31}$$

where σ is the logistic function and $f., g.$ denote random projections.

We then define two continuous XOR-like interactions:

$$p_{\text{on}} = p_1^{(a)}(1 - p_2^{(a)}) + (1 - p_1^{(a)})p_2^{(a)}, \tag{32}$$

$$p_{\text{off}} = p_1^{(b)}(1 - p_2^{(b)}) + (1 - p_1^{(b)})p_2^{(b)}. \tag{33}$$

These correspond to soft XOR operations taking values in $[0, 1]$.

Finally, the label is obtained by combining the two interactions:

$$y = \min\left(2\, p_{\text{off}} + p_{\text{on}},\ 3\right), \tag{34}$$

which yields four discrete classes $y \in \{0, 1, 2, 3\}$.

This construction can be interpreted as a composition of two XOR-like interactions, producing a 4-class label where predictive information is only accessible jointly from (X_1, X_2).

F.2 Architecture and training details

Each modality is encoded by a linear projection from input dimension $D = 200$ to a hidden dimension of 512. The two encoded representations are concatenated and passed through a three-layer MLP (dimensions $1024 \to 512 \to 512$) to produce the joint representation. The joint prediction head and two unimodal prediction heads are each two-layer MLPs ($512 \to 512 \to 4$). All three heads are trained simultaneously by minimizing the sum of cross-entropy losses $\mathcal{L}_{\text{joint}} + \mathcal{L}_1 + \mathcal{L}_2$.

Training uses AdamW (lr=10^{-4}, weight decay=10^{-2}), batch size 32, for 30 epochs. The best model is selected on validation loss. After training, hidden representations (dim 512) from each modality encoder are extracted and frozen for Stage 2 (SupGK), which uses the default hyperparameters: two-layer MLP projectors with hidden dimension 1024, ℓ_2-normalized output of dimension $\min(d_1, d_2) = 512$, shared prediction head ($512 \to 1024 \to 4$), trained with Adam (lr=10^{-4}), batch size 1024, 100 epochs, $\lambda = 10$.

F.3 Discussion of results

Pure regimes. On the four pure datasets, our method achieves near-perfect recovery. In 100% R, the estimated redundancy ($R = 0.755$) closely matches the ground truth (0.756) with $\Delta = 0.001$, while LSMI slightly overestimates ($R = 0.77$) and produces a small negative $U_2 = -0.005$. In 100% U_1 and 100% U_2, our method correctly assigns zero redundancy and recovers the dominant uniqueness component, whereas LSMI introduces negative redundancy values ($R = -0.016$ and $R = -0.024$ respectively).

Source redundancy in pure regimes. Our method correctly identifies the source proportion in all four pure settings: 100% source in 100% R (redundancy arises from the shared latent z_c), and 0% source in the remaining three (no shared latent contributes to the label). This confirms that SupGK does not spuriously detect source redundancy when none exists.

Two-component mixtures. Performance remains strong on most two-component mixtures. On 33% R + 67% S, our method correctly identifies 100% source redundancy for the R component and closely matches the ground-truth PID. On 33% U_1 + 67% U_2, both methods struggle with the uniqueness–uniqueness interaction (our $\Delta = 0.289$, LSMI $\Delta = 0.296$), likely because the mixing of two uniqueness components creates ambiguity between redundancy and uniqueness in the interpolated features. The source attribution correctly collapses to 0% in this setting, as no shared latent is involved.

Three-component mixtures. These are the most challenging settings. Our method shows the largest improvements over LSMI on 25% U_1 + 25% U_2 + 50% S ($\Delta = 0.060$ vs. 0.099) and 50% R + 25% U_2 + 25% S ($\Delta = 0.228$ vs. 0.292). In the latter, the gain comes mainly from a better recovery of the unique component ($U_2 = 0.048$ vs. 0.004 for LSMI; ground truth 0.082). Both methods overestimate redundancy ($R = 0.28$ for ours and 0.293 for LSMI, against a ground truth of 0.164) and recover almost none of the synergy (ground truth $S = 0.082$), which suggests that part of the synergistic component is absorbed into redundancy in this setting. Our method attributes all of the recovered redundancy to source redundancy (100%). On 50% U_1 + 25% U_2 + 25% R, our method recovers the redundancy accurately ($R = 0.093$ vs. 0.09), and the source proportion is estimated at 61.5%. This is the only setting where the source–mechanistic split is not binary. It reflects the fact that the R component contributes 25% of the mixture, while the interaction between the two uniqueness components introduces a small amount of mechanistic redundancy through the mixing process.

	100% R					100% U_1					100% U_2					100% S				
PID	R	U_1	U_2	S	$\Delta\downarrow$	R	U_1	U_2	S	$\Delta\downarrow$	R	U_1	U_2	S	$\Delta\downarrow$	R	U_1	U_2	S	$\Delta\downarrow$
Batch	0.741	0.014	0	0.66	0.689	0.027	0.69	0	0.042	0.174	0.762	0	0	0.804	2.348	0.044	0.007	0	0.187	0.097
FlowPID	0.797	0.006	0	0.069	0.116	0.033	0.762	0	0.108	0.174	0.035	0	0.762	0.09	0.145	0.032	0	0	0.035	0.230
LSMI	0.77	0.003	-0.005	0	0.023	-0.016	0.817	0	0.01	0.048	-0.024	0	0.823	-0.015	0.08	-0.024	0.008	0	0.217	0.048
Ours	0.755 (100%)	0	0	0	**0.001**	0.0 (0%)	0.781	0	0.013	**0.027**	0.0 (0%)	0	0.782	0.0	**0.000**	0.0 (0%)	0.013	0.003	0.216	**0.033**
Truth	0.756	0	0	0		0	0.795	0	0		0	0	0.782	0		0	0	0	0.233	

	33% U_1 + 67% S					33% R + 67% S					33% U_1 + 67% U_2					33% U_2 + 67% R				
PID	R	U_1	U_2	S	$\Delta\downarrow$	R	U_1	U_2	S	$\Delta\downarrow$	R	U_1	U_2	S	$\Delta\downarrow$	R	U_1	U_2	S	$\Delta\downarrow$
Batch	0.052	0.09	0	0.15	0.102	0.158	0	0	0.119	0.099	0.62	0.02	0	0.58	1.620	0.66	0	0	0.033	0.541
FlowPID	0.052	0.120	0	0.054	0.182	0.14	0.01	0	0.05	0.160	0.15	0	0.51	0.113	0.623	0.66	0	0.03	0.07	0.548
LSMI	0.022	0.087	0	0.103	0.07	0.08	0.004	0.0	0.091	**0.053**	0.064	0	0.365	0.017	0.296	0.426	0	0.032	-0.006	0.248
Ours	0 (0%)	0.098	0.004	0.10	**0.066**	0.053 (100%)	0.021	0.016	0.093	0.071	0.043 (0%)	0	0.373	0.023	**0.289**	0.415 (100%)	0	0.031	0	**0.232**
Truth	0	0.063	0	0.127		0.06	0	0	0.12		0	0.145	0.295	0		0.299	0	0.147	0	

	50% U_1 + 25% U_2 + 25% R					50% R + 25% U_2 + 25% S					25% R + 50% U_1 + 25% S					25% U_1 + 25% U_2 + 50% S				
PID	R	U_1	U_2	S	$\Delta\downarrow$	R	U_1	U_2	S	$\Delta\downarrow$	R	U_1	U_2	S	$\Delta\downarrow$	R	U_1	U_2	S	$\Delta\downarrow$
Batch	0.17	0.3	0	0.031	0.321	0.485	0	0	0.7	1.021	0.094	0.40	0	0.052	0.274	0.126	0	0	0.144	0.270
FlowPID	0.22	0.3	0	0.086	0.426	0.46	0	0.068	0.05	0.342	0.128	0.4	0	0.064	0.296	0.139	0	0.004	0.14	0.275
LSMI	0.11	0.224	0	0.02	0.174	0.293	0	0.004	-0.003	0.292	0.055	0.289	0	0	0.234	0.049	0.023	0.035	0.1	0.099
Ours	0.093 (61.5%)	0.237	0.014	0.01	**0.146**	0.28 (100%)	0	0.048	0.004	**0.228**	0.054 (100%)	0.276	0	0.004	**0.218**	0 (0%)	0.064	0.069	0.073	**0.060**
Truth	0.09	0.18	0.09	0		0.164	0	0.082	0.082		0.084	0.168	0	0.084		0	0.052	0.052	0.104	

Table 8: **Multimodal PID benchmark on the synthetic dataset.** R, U_1, U_2, and S denote redundancy, unique information from each modality, and synergy, respectively. Δ is the absolute error between the estimated and reference PID decomposition (lower is better). For our method, percentages in parentheses indicate the fraction of estimated redundancy attributed to source redundancy. "Truth" denotes the reference distribution-level PID induced by the generating process.

G Connecting Partial Information Decomposition to Variance Explained in Regression cases

We derive the mapping between the information-theoretic Partial Information Decomposition (PID) and the variance-explained (R^2) decomposition used throughout this work. All results assume jointly Gaussian variables.

G.1 Mutual information and R^2 under Gaussianity

Let Y be a scalar target and $\hat{Y}$ a predictor of Y. Under the Gaussian assumption $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$ and $Y \mid \hat{Y} \sim \mathcal{N}(\hat{Y}, \sigma_\varepsilon^2)$, the mutual information between Y and $\hat{Y}$ is

$$I(Y;\hat{Y}) = H(Y) - H(Y \mid \hat{Y}) = \frac{1}{2}\log\frac{\sigma_Y^2}{\sigma_\varepsilon^2} = -\frac{1}{2}\log\big(1 - R^2\big), \tag{35}$$

where $R^2 = 1 - \sigma_\varepsilon^2/\sigma_Y^2$ is the coefficient of determination. The inverse mapping is

$$R^2 = 1 - e^{-2I}. \tag{36}$$

This relationship is *monotone* but *nonlinear*: mutual information is additive when its components are independent, whereas R^2 is not.

G.2 PID in information space

Consider two source modalities X_1 (neuroanatomy) and X_2 (speech) predicting a clinical target Y. The PID of $I(Y; X_1, X_2)$ decomposes the joint mutual information into four non-negative atoms (Williams & Beer, 2010):

$$I(Y; X_1, X_2) = \underbrace{R}_{\text{redundancy}} + \underbrace{U_1}_{\text{unique to } X_1} + \underbrace{U_2}_{\text{unique to } X_2} + \underbrace{S}_{\text{synergy}}, \tag{37}$$

with the marginal constraints

$$I(Y; X_1) = R + U_1, \qquad I(Y; X_2) = R + U_2. \tag{38}$$

In practice, we estimate each mutual information term via Eq. equation 35 applied to the corresponding model's predictions. Let $R_1^2, R_2^2, R_{\text{joint}}^2$ denote the test-set R^2 of the first-modality-only, second-modality-only, and joint models, respectively. We compute

$$I_1 = -\tfrac{1}{2}\log(1 - R_1^2), \quad I_2 = -\tfrac{1}{2}\log(1 - R_2^2), \quad I_{\text{joint}} = -\tfrac{1}{2}\log(1 - R_{\text{joint}}^2). \tag{39}$$

The redundancy R is estimated via our pointwise extension of our $i_{\min}$ approach (Williams & Beer, 2010): for each test sample k, we compute the pointwise mutual information

$$i(x_m^{(k)}; y^{(k)}) = \frac{1}{2}\log\frac{\sigma_Y^2}{\sigma_{\varepsilon,m}^2} + \frac{(y^{(k)} - \mu_Y)^2}{2\sigma_Y^2} - \frac{(y^{(k)} - h_m(f_m(x_m^{(k)})))^2}{2\sigma_{\varepsilon,m}^2}, \quad m \in \{1, 2\}, \tag{40}$$

where $\sigma_{\varepsilon,m}^2 = \frac{1}{N}\sum_k (y^{(k)} - \hat{y}_m^{(k)})^2$ is the mean squared error of modality m.

We use the sign-filtered pointwise $I_{\min}$ redundancy of Eq. 3, inspired by Williams & Beer (2010):

$$r^{(k)} = \begin{cases} \min\big(i(x_1^{(k)}; y^{(k)}),\, i(x_2^{(k)}; y^{(k)})\big), & \text{if } i(x_1^{(k)}; y^{(k)})\, i(x_2^{(k)}; y^{(k)}) > 0, \\ 0, & \text{otherwise,} \end{cases} \tag{41}$$

and the distribution-level redundancy is $R = \frac{1}{N}\sum_{j=1}^{N} r^{(j)}$. The unique and synergy atoms follow from the PID constraints in Eqs. equation 1–equation 38:

$$U_1 = I_1 - R, \qquad U_2 = I_2 - R, \qquad S = I_{\text{joint}} - I_1 - I_2 + R. \tag{42}$$

G.3 Source vs. mechanistic redundancy

Redundancy can arise from two distinct mechanisms (Sec. 3.2.1). *Source redundancy* R_{src} is the part of the redundancy carried by target-predictive information shared between the modalities, i.e., a common latent factor recoverable from both X_1 and X_2 (App. D). *Mechanistic redundancy* R_{mech} arises when X_1 and X_2 predict the same aspects of Y through distinct, unshared information, as in the logic circuits and the digit-summation task (Sec. 4.4).

We estimate source redundancy from the SupGK representations $Z_m^R = g_m^R(Z_m)$ and the shared prediction head h_R, trained on stop-gradient inputs (Sec. 3.2.1, steps 2–3). For test sample k, let $\hat{y}_{R,m}^{(k)} = h_R\big(z_m^{R(k)}\big)$ and $\sigma_{\varepsilon,R,m}^2 = \frac{1}{N}\sum_k \big(y^{(k)} - \hat{y}_{R,m}^{(k)}\big)^2$. The pointwise mutual information $i\big(z_m^{R(k)}; y^{(k)}\big)$ is given by Eq. 40, with $h_m(f_m(x_m^{(k)}))$ and $\sigma_{\varepsilon,m}^2$ replaced by $\hat{y}_{R,m}^{(k)}$ and $\sigma_{\varepsilon,R,m}^2$. Pointwise source redundancy is the Gaussian counterpart of Eq. 8, with the same sign filter as Eq. 41:

$$r_{\text{src}}^{(k)} = \begin{cases} \min\big(i(z_1^{R(k)}; y^{(k)}),\, i(z_2^{R(k)}; y^{(k)})\big), & \text{if } i(z_1^{R(k)}; y^{(k)})\, i(z_2^{R(k)}; y^{(k)}) > 0, \\ 0, & \text{otherwise,} \end{cases} \tag{43}$$

and the distribution-level source redundancy is the test-set average $R_{\text{src}} = \frac{1}{N}\sum_{k=1}^{N} r_{\text{src}}^{(k)}$. As in App. E.2, we enforce $R \leftarrow \max(R, R_{\text{src}})$ and define the mechanistic component as the remainder:

$$R_{\text{mech}} = R - R_{\text{src}} \;\geq\; 0. \tag{44}$$

Both quantities are mapped to explained variance in App. G.4.

G.4 Mapping PID atoms to variance explained

To express PID atoms in R^2 (variance-explained) space, we exploit the relationship in Eq. equation 36. For each marginal modality, the explained variance decomposes naturally. Since $I_1 = R + U_1$, Eq. equation 36 gives

$$R_1^2 = 1 - e^{-2I_1} = 1 - e^{-2(R+U_1)}. \tag{45}$$

We define the R^2 attributable to redundancy as the explained variance obtained from R alone:

$$R_R^2 \;=\; 1 - e^{-2R}. \tag{46}$$

The unique contribution of modality 1 in variance-explained space is the *incremental* R^2 gained by adding U_1 nats of unique information on top of the R nats of redundancy:

$$R^2_{U_1} \;=\; R^2_1 - R^2_R \;=\; \left(1 - e^{-2(R+U_1)}\right) - \left(1 - e^{-2R}\right) \;=\; e^{-2R} - e^{-2(R+U_1)}, \tag{47}$$

and symmetrically for $R^2_{U_2}$. Synergy in R^2 space is obtained as the residual:

$$R^2_S \;=\; R^2_{\text{joint}} - R^2_R - R^2_{U_1} - R^2_{U_2}. \tag{48}$$

Source vs. mechanistic redundancy in R^2 space. Since $R = R_{\text{src}} + R_{\text{mech}}$ in nats, the same incremental logic yields:

$$R^2_{R_{\text{src}}} = 1 - e^{-2R_{\text{src}}}, \tag{49}$$

$$R^2_{R_{\text{mech}}} = R^2_R - R^2_{R_{\text{src}}} = e^{-2R_{\text{src}}} - e^{-2R}. \tag{50}$$

This decomposition is exact and preserves additivity: $R^2_{R_{\text{src}}} + R^2_{R_{\text{mech}}} = R^2_R$.

Rescaling to R^2_{joint}. In practice, estimation noise can cause $R^2_R + R^2_{U_1} + R^2_{U_2} + R^2_S$ to differ slightly from R^2_{joint}. We therefore apply a multiplicative procedure: each component is scaled by $R^2_{\text{joint}}/(R^2_R + R^2_{U_1} + R^2_{U_2} + R^2_S)$ so that the decomposition sums exactly to the observed joint explained variance. The unexplained fraction is $1 - R^2_{\text{joint}}$.

Summary. The total variance of Y is partitioned as:

$$1 = \underbrace{R^2_{R_{\text{src}}} + R^2_{R_{\text{mech}}}}_{\text{redundancy}} + \underbrace{R^2_{U_1}}_{\substack{\text{unique}\\ \text{(modality 1)}}} + \underbrace{R^2_{U_2}}_{\substack{\text{unique}\\ \text{(modality 2)}}} + \underbrace{R^2_S}_{\text{synergy}} + \underbrace{(1 - R^2_{\text{joint}})}_{\text{unexplained}}. \tag{51}$$

This is the decomposition displayed in the pie charts of Figure 1c.

H IMPLEMENTATION DETAILS

All experiments can be run with a single H100 GPU, 8 CPUs, and 32 GB RAM.

Common training protocol. All experiments follow the same two-stage pipeline (Algorithm 2). In Stage 1, we train a multimodal model with a joint prediction head and two unimodal heads, optimizing the sum of three cross-entropy losses $\mathcal{L}_{\text{joint}} + \mathcal{L}_1 + \mathcal{L}_2$. We use AdamW with gradient clipping (max norm 8) unless stated otherwise. After training, we extract mean-pooled encoder representations for each modality, which serve as frozen inputs to Stage 2 (SupGK, described below).

AV-MNIST. We pair MNIST images (Y. et al., 1998) with spoken digits from the Free Spoken Digit Dataset (Jackson et al., 2018) (mel-spectrograms, 64 bins, hop length 128, padded/cropped to a fixed length). The model is a shared CNN architecture with separate visual and audio streams, trained for 30 epochs with Adam (lr=10^{-3}, $\gamma = 0.996$ step decay). Batch size is 1024. Encoder representation dimension: task-dependent (final convolutional feature maps).

TriFeatures. We use the BimodalTriFeatures dataset (Hermann & Lampinen, 2020; Dufumier et al., 2025), pairing two 224×224 synthetic images. Each modality is encoded by an AlexNet backbone (no pretraining) followed by a 1×1 Conv2d projection with 2D sinusoidal positional embeddings, producing 36 spatial tokens of dimension $d = 512$. Fusion uses a single-layer Transformer with 8 heads and a CLS token. Training: 50 epochs, AdamW (lr=10^{-4}, weight decay=10^{-2}), batch size 64. Pairs are reshuffled each epoch for data augmentation.

CMU-MOSEI, UR-Funny, MUStARD. For temporal multimodal datasets, we use the CoMM architecture (Dufumier et al., 2025): each modality is projected via a 1×1 Conv1d to $d = 40$, then processed by a 5-layer Transformer encoder (5 heads, pre-norm). Fusion concatenates both token sequences with a CLS token and applies a single-layer Transformer (8 heads). Sequences are truncated or padded to length 50 (MOSEI) or 20 (UR-Funny). Training: 10 epochs, AdamW (lr=10^{-3}, weight decay=10^{-2}), batch size 32. Data augmentation consists of additive Gaussian noise ($\sigma = 0.1$) and temporal dropout (up to 80% of timesteps). Modality pairs: vision–text for MOSEI, MUStARD, and UR-Funny.

UPMC-Food101. Images are encoded by a pretrained ConvNeXt-Base (Liu et al., 2022) producing spatial tokens projected to $d = 512$ via a 1×1 Conv2d with 2D sinusoidal positional embeddings. Text titles are encoded by RoBERTa-Base (Liu et al., 2019) and linearly projected to $d = 512$. Fusion follows the same CLS-token Transformer as above (8 heads, 1 layer). Training: 20 epochs, AdamW (lr=10^{-4}, weight decay=10^{-2}), batch size 32, cosine annealing schedule. A 10% validation split is held out from the training set.

SupGK (Stage 2). Frozen encoder representations from Stage 1 are fed to two-layer MLP projectors (g_1, g_2) with hidden dimension $h = 1024$ and ℓ_2-normalized outputs of dimension $\min(d_1, d_2)$. The shared prediction head h_R is a two-layer MLP ($h = 1024$). Training uses Adam (lr=10^{-4}) for 100 epochs with batch size 1024. We use $\lambda = 10$ for the alignment penalty (and $\lambda = 25$ for the speech + neuroanatomical experiments) and $\sigma = 0.1$ for the kernel bandwidth of the Kernel on Y for the continuous version of SupGK. We use $\tau = 0.1$ for the kernel bandwidth of the Kernel on Z^R for the continuous and categorical version of SupGK. Appendix M discusses the hyperparameters choice. Early stopping is based on validation loss; the best checkpoint is used for test-time PID estimation. For continuous targets (Huntington's disease), we use MSE loss and the continuous SupGK variant with σ_y set to the training-set standard deviation of the target score.

LSMI baseline implementation We use the official LSMI implementation[3] with default hyperparameters. To ensure a fair comparison, LSMI operates on the same frozen encoder representations as our method (extracted after Stage 1 training), rather than on raw inputs.

LSMI requires two sets of estimators. First, three discriminator networks $p(y|z_1)$, $p(y|z_2)$, and $p(y|z_1, z_2)$ are trained to estimate pointwise mutual information via $i(z_m; y) = \log K + \log p(y|z_m)$

[3] https://github.com/GeWu-Lab/LSMI_Estimator

where K is the number of classes. Each discriminator is a two-layer MLP with hidden dimension 32 and output dimension equal to the number of classes. They are trained jointly for 30 epochs with Adam (lr=10^{-3}), step decay (factor 0.1 every 15 epochs), and batch size 512.

Second, two KNIFE-based marginal entropy estimators $H(Z_1)$ and $H(Z_2)$ are trained using the MargKernel module from the LSMI codebase, for 30 epochs with Adam (lr=10^{-3}) and step decay (factor 0.1 every 20 epochs).

Redundancy is then computed as $r = \min(H(Z_1), H(Z_2)) - \min(H(Z_1|Y), H(Z_2|Y))$, with the remaining components derived as $u_m = I(Z_m; Y) - r$ and $s = I(Z_1, Z_2; Y) - r - u_1 - u_2$. Distribution-level values are adjusted using the RUS procedure from Yang et al. (2025), which shifts all components by a constant to enforce non-negativity of R and S at the expectation level.

BATCH baseline implementation We use the official implementation of BATCH (Liang et al., 2023) with its default hyperparameters, operating on the same frozen Stage 1 representations as our method and LSMI. We observed that BATCH occasionally returns decompositions that are internally inconsistent, i.e. whose atoms do not sum to the total mutual information. On the pure U_2 synthetic dataset, it returns $R = 0.762$ and $S = 0.804$, so that $R + U_1 + U_2 + S = 1.566$ nats, about twice $I(X_1, X_2; Y) = 0.782$ nats (Table 1). Similarly, on TriFeatures-100% U_2, it returns $U_2 = S = 2.28$, summing to 4.56 nats against a total of 2.01 nats (Table 2). These values violate the consistency constraint of Eq. 1, which holds by construction for our estimator and for LSMI. We report them as obtained rather than discarding them.

I Speech x Neuroanatomy Experimental Details

Developmental cohort, eligibility, and clinical variables. The dataset comprised 341 people with Huntington's disease (PwHD) and 855 visits with clinical assessments. Among these visits, 634 had speech recordings, 220 had structural MRI only, and 211 had both speech and structural MRI. Unimodal prediction heads were trained using all visits available for the corresponding modality, whereas multimodal and speech-neuroanatomy alignment analyses were restricted to visits with paired speech and structural MRI. Importantly, all quantities entering the decomposition, namely the unimodal, joint, and source-redundant R^2 and the pointwise redundancies $r^{(k)}$, are evaluated on the *same* held-out paired test visits of each split.

Participants were considered PwHD if they carried at least 36 CAG repeats in the *HTT* (Huntingtin) gene and were excluded if they had a neurological disorder other than HD. Only speech recordings with sufficient sound quality were retained. Demographic and genetic variables included sex, age, education, and CAG repeat length. Clinical targets covered motor, cognitive, functional, and global disease severity: cUHDRS (Schobel et al., 2017), Total Motor Score (TMS), Symbol Digit Modalities Test (SDMT), Stroop Word Reading (SW), Total Functional Capacity (TFC), Stroop Colour Naming (SC), Stroop Interference (SI), lexical fluency (LF), and Independence Scale (IS). Higher cUHDRS, SDMT, Stroop, LF, TFC, and IS values indicate better clinical status, whereas higher TMS indicates greater motor impairment.

Structural MRI features. Structural MRI was represented by regional gray-matter measures extracted from 125 regions of interest defined by the Neuromorphometrics atlas in the CAT12 toolbox for SPM. For each visit with available MRI, this yielded a neuroanatomical feature vector:

$$\mathbf{x}_{\mathrm{sMRI}}^{p,v} \in \mathbb{R}^{125}.$$

The neuroanatomical encoder mapped these 125 ROI features through a two-layer MLP with hidden size 32. Only visits with paired speech and MRI were used for multimodal prediction and SupGK-based speech–neuroanatomy alignment.

Speech protocol and audio preprocessing. Speech recordings were obtained during standardized sessions conducted by trained neuropsychologists. REPAIRHD (NCT03119246) `https://jnnp.bmj.com/content/89/Suppl_1/A59.1` and BIOHD (NCT01412125) `https://clinicaltrials.gov/study/NCT01412125` recordings used a microphone, whereas more recent BIOHD recordings used a tablet. We did not exclude speech tasks a priori, since no

single task has been established as optimal for capturing HD-related speech alterations and deep learning models benefit from large and diverse training data.

Speech collection followed the Basal Voice Protocol, a fixed-order battery of 13 tasks. The protocol includes constrained productions (e.g., sustained vowels, phoneme production, counting, and automatic recitation), which limit lexical variability and facilitate comparisons across participants, as well as sentence-level and narrative productions (e.g., backward counting, picture description, past-24-hours recall, emotional narratives, and story retelling), which probe motor, prosodic, cognitive, and functional dimensions of speech.

Each recording was first segmented using task annotations provided by speech therapists. Examiner instructions, non-task intervals, and long silences were removed. A second segmentation step was then applied using pyannote-audio voice activity detection `https://huggingface.co/pyannote/voice-activity-detection` with default parameters to extract continuous active-speech segments of 4–10 seconds. Segments were resampled to 16 kHz before wav2vec 2.0 feature extraction. This two-stage procedure yielded task-aligned, speech-only inputs of comparable duration for NDSNet training and evaluation.

NDSNet speech encoder. For the clinical speech–neuroanatomy experiments, we used the Neurodegenerative Speech Network (NDSNet) as a supervised speech encoder. As a front end, NDSNet uses a pretrained wav2vec 2.0 speech model to transform each raw speech segment into a sequence of frame-level representations. In the configuration used here, the front end was the Hugging Face checkpoint `facebook/wav2vec2-xlsr-53-espeak-cv-ft`, a wav2vec 2.0 XLSR model optimized for phonetic transcription. For a speech segment $x^{p,v,k}$, recorded from participant p at visit v, the front end produces

$$\{\mathbf{w}_t^{p,v,k}\}_{t=1}^{T} = \mathrm{W2V2}(x^{p,v,k}), \qquad \mathbf{w}_t^{p,v,k} \in \mathbb{R}^{d_w},$$

with $d_w = 1024$. These frame-level wav2vec 2.0 representations, rather than decoded transcriptions, are used as input to the NDSNet temporal encoder. Static demographic and genetic variables $\mathbf{y}^{p,v} \in \mathbb{R}^D$, when used, are broadcast over time and concatenated channel-wise to the speech representation,

$$\mathbf{q}_t^{p,v,k} = [\mathbf{w}_t^{p,v,k}; \mathbf{y}^{p,v}] \in \mathbb{R}^{d_w+D}.$$

The temporal encoder consists of two bidirectional LSTM layers with 128 hidden units per direction and dropout 0.2:

$$\mathbf{H}^{(0)} = (\mathbf{q}_1^{p,v,k}, \ldots, \mathbf{q}_T^{p,v,k}), \qquad \mathbf{H}^{(\ell)} = \mathrm{BiLSTM}^{(\ell)}(\mathbf{H}^{(\ell-1)}), \quad \ell = 1, 2.$$

Let $\mathbf{h}_t^{(2)} \in \mathbb{R}^{256}$ denote the output of the second bidirectional LSTM at frame t. A single-head attention pooling layer aggregates the recurrent sequence,

$$a_t = \mathbf{w}_a^\top \mathbf{h}_t^{(2)} + b_a, \qquad \alpha_t = \frac{\exp(a_t)}{\sum_{u=1}^{T} \exp(a_u)}, \qquad \mathbf{c}^{p,v,k} = \sum_{t=1}^{T} \alpha_t \mathbf{h}_t^{(2)}.$$

A linear projection maps the context vector to the NDSNet latent space,

$$\mathbf{h}^{p,v,k} = \mathbf{W}_{\mathrm{emb}} \mathbf{c}^{p,v,k} + \mathbf{b}_{\mathrm{emb}} \in \mathbb{R}^{32}.$$

During NDSNet pretraining, a batch-normalized version of this embedding is passed to a prediction head with sigmoid output activation,

$$\hat{\mathbf{s}}^{p,v,k} = \sigma\big(\mathbf{W}_o \,\mathrm{BN}(\mathbf{h}^{p,v,k}) + \mathbf{b}_o\big)\,,$$

to predict the visit-level clinical target vector $\mathbf{s}^{p,v}$. We therefore trained NDSNet using only a Mean Squared Error prediction loss on all clinical variables.

$$\mathcal{L}_{\mathrm{NDSNet}} = \frac{1}{|\mathcal{B}|} \sum_{(p,v,k)\in\mathcal{B}} \left\| \hat{\mathbf{s}}^{p,v,k} - \mathbf{s}^{p,v} \right\|_1 .$$

Training used batches of size 32, Adam with learning rate 10^{-5}, and $\lambda_{L_1} = 1$. After training, the prediction head was discarded and the latent vector $\mathbf{h}^{p,v,k}$, before batch normalization, was

retained as the speech representation. For visits with multiple speech segments, we used the average segment-level embedding

$$\mathbf{x}_{\text{speech}}^{p,v} = \frac{1}{K_{p,v}} \sum_{k=1}^{K_{p,v}} \mathbf{h}^{p,v,k}.$$

This visit-level embedding was paired with the corresponding neuroanatomical representation in the downstream multimodal prediction and PID analyses.

Broader Impact. Our framework demonstrates that speech signals can recover a substantial portion of the neuroanatomical information predictive of clinical status in Huntington's disease. This finding has direct translational implications: speech can be recorded remotely using consumer devices, offering a non-invasive, low-cost alternative to structural MRI for monitoring disease progression. In practice, this could enable at-home assessment of patients' cognitive and motor status, reducing the burden of hospital visits and increasing the temporal resolution of clinical follow-up. Such remote speech-based biomarkers could also support clinical trial enrichment by identifying patients whose disease trajectory is progressing more rapidly, and enable continuous at-home tracking of cognitive status between scheduled visits. More broadly, the ability to quantify precisely the target-predictive information shared between the NDSNet representation and neuroanatomy, and the one that is not, may strengthen neurological follow-up protocols for a range of neurodegenerative diseases beyond HD, by clarifying when speech-based monitoring suffices and when imaging remains necessary. On the other hand, over-reliance on automated speech-based predictions without clinician oversight could lead to misdiagnosis or inappropriate treatment decisions, particularly for patients whose speech patterns fall outside the training distribution. Privacy considerations also arise from the collection and storage of longitudinal speech recordings, which may contain sensitive information beyond clinical content. In our experiments, data were anonymized and stored on private and protected servers, thus preventing sensitive information from falling into the public domain.

J Application to Alzheimer's Disease Biomarker Analysis

We apply our cascading PID framework to the ADNI dataset to characterize how multimodal biomarkers share and combine predictive information when estimating current cognitive status and subsequent cognitive decline.

Dataset and modalities. We use the TADPOLE D1/D2 dataset (Marinescu et al., 2019), aggregating data from ADNI-1/GO/2. We define six modality groups: demographics (age, sex, education, APOEε4), baseline cognitive scores (ADAS-13, MMSE, CDR-SB, FAQ, RAVLT, logical memory, and trail-making; longitudinal task only), CSF amyloid (Aβ_{42}), CSF tau (total and phospho-tau), volumetric MRI (ICV-normalized hippocampal, entorhinal, fusiform, ventricular, and mid-temporal volumes), and FDG-PET (MetaROI composite (Landau et al., 2011)). After filtering for complete data, the analyses include $N = 386$ participants for the cross-sectional task and $N = 342$ for the longitudinal task, pooling CN, EMCI, LMCI, and AD participants.

Tasks. **Cross-sectional**: predicting current ADAS-13 from baseline biomarkers. **Longitudinal**: predicting annual ADAS-13 slope (pts/yr) from baseline biomarkers, computed as the average rate of change across follow-up visits occurring at least 12 months after baseline.

Cascading PID. Starting from demographics, we greedily add the remaining modality that maximizes joint validation R^2 at each step. For every addition, we compute a bivariate PID between the accumulated modality set and the newly added modality. Three XGBoost regressors are trained at each step, with hyperparameters selected by grid search on a validation set. PID values are mapped to variance-explained space through $R^2 = 1 - e^{-2I}$. Because modality selection is based on validation performance, test-set R^2 is not constrained to increase monotonically.

Cross-sectional results (Fig. 2, upper). MRI is the first non-demographic modality selected, explaining 37.6% of the variance in current ADAS-13. Adding FDG-PET increases joint performance to $R^2 = 0.477$ and yields 22.0% mechanistic redundancy with MRI, together with 9.8% unique information from FDG-PET. No source redundancy is recovered at this step, indicating that, within

the learned representations, the predictive overlap between MRI and FDG-PET is attributed to mechanistic rather than source redundancy.

Adding CSF tau further increases performance to $R^2 = 0.516$. The decomposition again contains predominantly mechanistic redundancy (16.5%), with little unique information from tau (1.9%) and a small synergistic contribution (2.0%). Amyloid is selected next based on validation performance but does not improve test performance ($R^2 = 0.479$). At this step, 8.7% of the variance is attributed to mechanistic redundancy, while the accumulated modalities account for 39.2% unique information. Overall, no source redundancy is recovered throughout the cross-sectional cascade. Thus, according to the proposed decomposition, MRI, FDG-PET, tau, and amyloid contain overlapping information associated with current cognitive status, but this overlap is predominantly represented as mechanistic rather than source redundancy.

Longitudinal results (Fig. 2, lower). The interaction structure differs when predicting future cognitive decline. Baseline cognitive scores are selected first and yield $R^2 = 0.154$, including 6.6% unique information and 8.8% synergy with demographics. CSF tau is selected next, increasing performance to $R^2 = 0.284$. Tau contributes 5.9% unique information and 7.1% synergy with the accumulated baseline representation. This indicates that the combination of baseline cognitive status and tau measurements carries predictive information about subsequent decline that is not captured by either component independently.

MRI is selected next, yielding $R^2 = 0.296$. In contrast to the cross-sectional analysis, this step recovers 6.9% source redundancy between MRI and the accumulated cognition–tau representation, together with 1.6% unique information from MRI. Within the proposed framework, this suggests that part of the information associated with future cognitive decline can be represented as a target-predictive component shared between MRI and the accumulated modalities. Finally, amyloid increases test performance to $R^2 = 0.321$ and contributes 2.5% synergy with the accumulated representation.

Cross-sectional vs. longitudinal contrast. The two cascades reveal different patterns of multimodal information sharing. For current cognitive status, MRI provides the largest initial predictive contribution, and the redundancy recovered after adding FDG-PET, tau, and amyloid is entirely mechanistic. For future decline, baseline cognition and tau provide complementary predictive information, with a larger role for synergy, while source redundancy emerges when MRI is added to the accumulated cognition–tau representation. These results are associational, but the contrast is compatible with the broad organization of the AT(N) framework (Jack Jr et al., 2018): neurodegeneration-related measurements are strongly associated with current clinical impairment, whereas pathological biomarkers and cognitive status may provide complementary information about subsequent progression.

K Common information estimation via CLIP is a bad estimator of source redundancy

Redundancy quantifies the information that makes X_1 and X_2 both predictive of the target Y. It decomposes further into a *mechanistic* part, quantifying the part where both modalities predict Y without carrying information genuinely shared between X_1 and X_2, and a *source* part, the share of redundancy that both predicts Y *and* stems from a source common to both X_1 and X_2. We denote a representation identifying source redundant latent vectors $Z_{R_{\text{source}}}$. This is inspired by the Gács–Körner common information, whose representation Z_C captures everything shared between X_1 and X_2 *regardless of its relevance to* Y. Source redundancy is therefore the target-relevant slice of the common information, and $Z_{R_{\text{source}}}$ is contained in Z_C; the two coincide only when every shared factor happens to predict Y.

That containment is what the following experiment exploits. Because CLIP optimizes a relaxed Gács–Körner objective (Appendix B), the representation it learns targets Z_C rather than $Z_{R_{\text{source}}}$. When the modalities share structure that is uninformative about Y, that structure is admitted into Z_C and corrupts any source-redundancy estimate read from it, even though the underlying partial information decomposition is unchanged. Our supervised objective $\mathcal{L}_{\text{SupGK}}$ constrains the shared representation to remain predictive of Y and so targets $Z_{R_{\text{source}}}$ directly. We construct a dataset in which

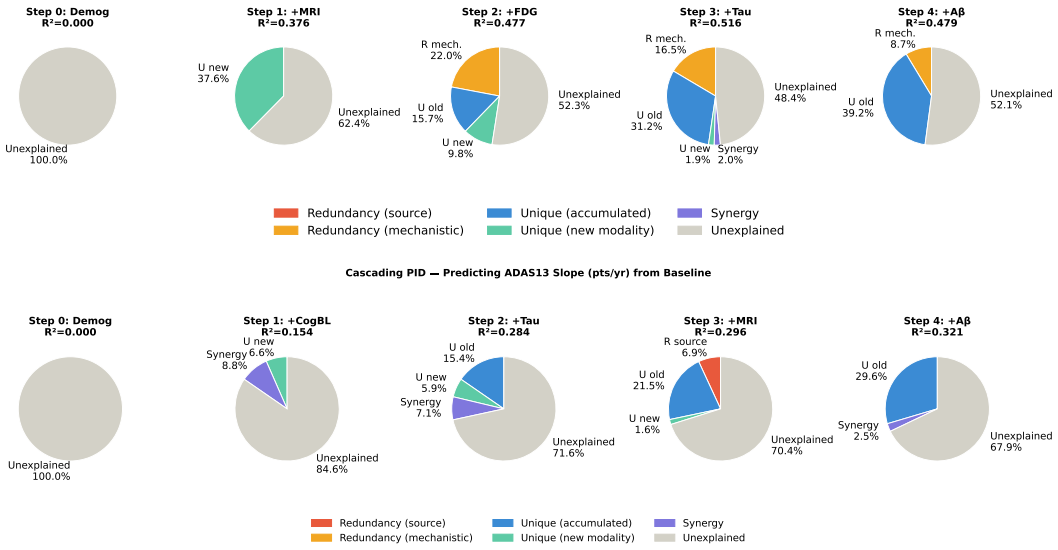


Figure 2: Cascading PID on the ADNI cohort. **Upper**: prediction of current ADAS-13. **Lower**: prediction of annual ADAS-13 slope (pts/yr). Each pie chart shows the variance partition at one cascade step: mechanistic redundancy (orange), source redundancy (red), unique accumulated information (blue), unique information from the newly added modality (green), synergy (purple), and unexplained variance (gray). In the cross-sectional cascade, the recovered redundancy is entirely mechanistic. In the longitudinal cascade, source redundancy is additionally recovered when MRI is added to the accumulated cognition–tau representation. These decompositions describe predictive associations and should not be interpreted causally.

the decomposition is fixed analytically. At the same time, only the target-irrelevant shared content varies, and we study the difference between SupGK and CLIP in quantifying source redundancy.

K.1 Construction

Each observation is generated from three latent parts: a *signal* z_{sig} that determines the label, a *nuisance* z_{nui} statistically independent of the label, and a *private* part z^m_{pri} specific to modality $m \in \{1, 2\}$. Each is mapped into a disjoint block of the observed input,

$$\text{base} = \tanh(z_{\text{sig}} Q), \qquad y = \text{quantile-bin}(\text{base}\, v), \tag{52}$$

$$\text{signal}_m = \text{base}\, R_m, \qquad \text{nuisance}_m = \tanh(\alpha\, z^m_{\text{nui}} T^m_n), \qquad \text{private}_m = \tanh(z^m_{\text{pri}} T^m_p), \tag{53}$$

with $x_m = [\,\text{signal}_m \mid \text{nuisance}_m \mid \text{private}_m\,] + \varepsilon$ and small observation noise ε. The projections Q, R_m and the read-out v are drawn once from a fixed seed and reused in every condition, so the signal block, the label, and the dependence of the label on each modality are held identical throughout. Since the label is a linear read-out of a signal that both modalities carry (up to the orthogonal rotation R_m), and the private part occupies separate coordinates, neither modality holds unique information, and no synergy is present. The label Y is derived from source redundancy only,

$$R = I(X_1, X_2; Y) = H(Y) - H(Y \mid X_1, X_2), \qquad R_{\text{source}} = R, \tag{54}$$

so the ground-truth source redundancy should always represent 100% of $I(Y; (X1, X2))$. In terms of representation, the signal should be captured in $Z_{R_{\text{source}}}$, the nuisance is a shared-but-irrelevant factor that should be captured in Z_C but not in $Z_{R_{\text{source}}}$, and the private part in neither.

We vary the target-irrelevant shared content linearly. The parameter $\rho \in [0, 1]$ controls how much of the nuisance is common: modality 1 observes z_{nui} and modality 2 observes $\rho\, z_{\text{nui}} + \sqrt{1 - \rho^2}\, \xi$ for independent ξ. At $\rho = 0$ the nuisance is present in both modalities but shared by neither, so it is not common information at all; at $\rho = 1$ it is a perfect common factor. Crucially, each modality's marginal law is identical for every ρ (matched dimension and variance), so any movement in an

Table 9: Configuration of the CLIP vs. SupGK study. Both estimators share every setting; only the training objective differs.

Data	**Representation training (both estimators)**
$n = 20{,}000$, 4 balanced classes	trunk Linear$(400 \to 1024)$
signal / nuisance / private $= 100/200/100$	projector $(1024 \to 1024 \to d_{\text{emb}})$, ℓ_2-normalised
latent variance 0.5, obs. noise 0.02	Adam, lr 10^{-4}, batch 512, 100 epochs
$70/15/15$ split	$\lambda = 10$ (warm-up), τ median-heuristic
$H(Y) = \log 4 \approx 1.386$ nats	linear probe: Adam, lr 10^{-3}, 60 epochs

estimate along this axis is attributable solely to the nuisance becoming shared, and not to any change in the inputs. We also sweep the embedding width d_{emb} at $\rho = 1$ to isolate whether the effect is an artifact of a constrained bottleneck.

Estimation protocol. The experiment reduces to training the two shared representations and probing them. CLIP and SupGK each train two modality projectors from the raw inputs into a common embedding, under the symmetric InfoNCE loss and $\mathcal{L}_{\text{SupGK}}$ respectively; we then fit a linear head on the frozen embedding and read source redundancy off it with the sign-filtered $i_{\min}$ estimator. Everything is shared between the two branches except the training objective. Recovered source redundancy is reported relative to $H(Y) = \log 4 \approx 1.386$ nats. The attainable ceiling is not 100% but $\approx 86\%$: indeed, we computed that a joint predictor on both modalities recovers $I(X_1, X_2; Y) \approx 0.86\, H(Y)$, which thus upper-bounds what any shared representation can reach, and we mark this level in Figure 3. All quantities are reported as the mean and standard deviation over three seeds, with the remaining hyperparameters given in Table 9.

K.2 Results

Figure 3 reports both sweeps.

A constant offset separates the two estimators even at $\rho = 0$, where the nuisance is present but unshared: CLIP recovers 53% of $H(Y)$ against SupGK's 84%. This offset is expected and is not the finding. An unsupervised objective sees no labels, and therefore has no criterion for preferring the target-relevant directions of the common space over any other; it captures the signal only partially. This shows that CLIP is limited for capturing source redundancy, and that SupGK enhances it by supervising through Y.

SupGK is invariant to the sharing of target-irrelevant content; CLIP is not. As ρ increases from 0 to 1, SupGK holds at its ceiling ($84\% \to 84\%$) while CLIP declines by eleven points ($53\% \to 42\%$). This decline cannot be ascribed to the absence of supervision, which is constant across the sweep, nor to any change in the inputs, whose marginals are held fixed by construction. It is caused only by the nuisance becoming a common factor (but not predictive of Y) and thereby entering Z_C, the representation CLIP targets. One might attribute CLIP's failure to an embedding too constrained to hold both the signal and the nuisance. However, the shared content spans at most 300 dimensions (signal 100, nuisance 200), so at $d_{\text{emb}} = 512$ the embedding already exceeds the dimensionality of everything it could represent, yet CLIP attains only 42%. That cannot be attributed to a too small embedding size.

The experiment separates two effects that a single comparison at fixed ρ would conflate. The level gap present already at $\rho = 0$ shows that not using a target Y affects the estimation of the source redundancy. The ρ-related decline shows that a common-information objective is corrupted by shared structure that is irrelevant to the target, because such structure is captured by Z_C, whereas $Z_{R_{\text{source}}}$ excludes it. The capacity sweep reinforces the same conclusion from a second direction, showing that these results are independent of the embedding width. Moreover, the plot shows that SupGK does not lose estimation capability in the presence of shared content being target-irrelevant.

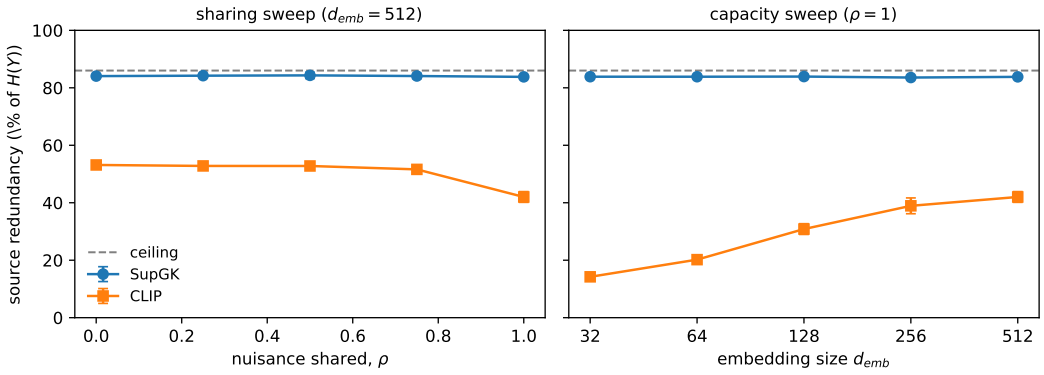


Figure 3: Recovered source redundancy against the $\approx 86\%$ ceiling (dashed). SupGK tracks the ceiling throughout. **Left**: as the nuisance is shared more (ρ), CLIP declines even though the ground truth and the input marginals are fixed. **Right**: the decline persists at every embedding size, including $d_{\text{emb}} = 512$, which exceeds the ≤ 300 dimensions of shared content, ruling out a capacity explanation. Mean $\pm$ std over three seeds.

L Validating the pointwise ground truth of the synthetic benchmark

Section 4.1 uses the Dirichlet weights $w^{(j)}$ as proxy ground truth for the pointwise interaction proportions. Since PID is a nonlinear functional of the joint distribution, convexly mixing features does not necessarily induce the same convex combination of PID atoms. Rather than assuming this correspondence, we validate it analytically. To avoid validating the benchmark with the estimator under evaluation, we compute the exact pointwise PID under two independent redundancy definitions: our own and the definition underlying our main pointwise baseline.

Setup. We retain the generative process of Section 4.1, replacing the neural encoders by class-conditional Gaussian variables so that every density and every pointwise mutual information can be computed analytically. Specifically, Y is uniformly distributed over four classes, with class means $\mu = (-1.5, -0.5, 0.5, 1.5)$ and within-class standard deviation $\sigma = 0.45$, and we draw $N = 10^5$ samples per scenario. Five latent variables encode redundant, unique-to-X_1, unique-to-X_2, and synergistic information, and are mixed according to

$$x_1 = w_R z_c + w_{U_1} z_1 + w_S z_a, \qquad x_2 = w_R z_c + w_{U_2} z_2 + w_S z_b, \qquad w \sim \mathrm{Dir}(c\bar{w}). \tag{55}$$

The synergistic variables satisfy $b = a \oplus Y$ with a drawn independently of Y, such that neither modality is individually informative whereas the pair jointly determines the target. The resulting conditional densities are finite Gaussian mixtures, allowing all pointwise mutual informations and PID atoms to be computed exactly without density estimation, entropy estimation, or trained models. We verified the implementation against an independent histogram estimator of $I(X_1; Y)$, agreeing to within 0.002 nats, and recovered the analytic special cases exactly.

Ground-truth computation. Since the pointwise PID consistency equations contain one degree of freedom, any ground truth must commit to a redundancy definition. To avoid circularity, we compute the exact pointwise redundancy under two independent definitions,

Definition A (ours):

$$r = \min(i(x_1; y), i(x_2; y))$$

Definition B (Finn & Lizier (2018), used by LSMI):

$$r = \min_m(-\log p(x_m)) - \min_m(-\log p(x_m|y))$$

where Definition B corresponds to the redundancy measure used by LSMI, evaluated here directly from the exact densities. For each sample, let $A^{(j)}$ denote the normalized PID atom vector. We report two complementary metrics. The *scenario cosine*,

$$\cos_{\text{scen}} = \cos\left(\frac{1}{N}\sum_j A^{(j)},\, w\right),$$

compares the *average* interaction proportions with the generating weights and therefore evaluates whether the benchmark is correctly labeled. The *pointwise cosine*,

$$\cos_{\text{pw}} = \frac{1}{N}\sum_j \cos(A^{(j)}, w),$$

compares every realization with its generating weights and therefore measures how faithfully a single sample reflects its underlying interaction proportions.

Table 10: Exact pointwise atoms versus the Dirichlet weights, $K = 4$ classes, $N = 10^5$ samples per scenario, both definitions computed from exact densities. The scenario cosine $\cos_{\text{scen}}$ tests whether w is the right label for the scenario; the pointwise cosine $\cos_{\text{pw}}$ is the metric of Table 1. Bold rows are the two scenarios whose ground truth we find to be invalid.

	Definition A (ours)		**Definition B (LSMI)**		**A vs. B**
Scenario	$\cos_{\text{scen}}$	$\cos_{\text{pw}}$	$\cos_{\text{scen}}$	$\cos_{\text{pw}}$	agreement
100% R	0.996	0.917	1.000	0.917	0.917
100% U_1	1.000	0.917	1.000	0.917	0.917
100% S	1.000	0.869	1.000	0.869	0.869
33% U_1 + 67% S	0.999	0.801	0.983	0.775	0.900
33% R + 67% S	0.868	0.628	0.902	0.646	0.960
33% U_1 + 67% U_2	**0.321**	0.257	**0.162**	0.160	0.869
33% U_2 + 67% R	0.961	0.839	0.945	0.862	0.955
50% U_1 + 25% U_2 + 25% R	**0.540**	0.481	**0.432**	0.412	0.882
50% R + 25% U_2 + 25% S	0.972	0.734	0.983	0.861	0.856
25% R + 50% U_1 + 25% S	0.986	0.749	0.980	0.768	0.920
25% U_1 + 25% U_2 + 50% S	0.979	0.698	0.962	0.681	0.961
Average (all)	0.875	0.717	0.850	0.715	0.910

Findings. Table 10 leads to four observations. First, the validation is largely independent of the redundancy definition. The two exact definitions produce highly consistent atom vectors, with an average agreement of 0.91 at the pointwise level.

Second, the benchmark is well calibrated for almost all interaction scenarios. The pure regimes are recovered almost perfectly ($\cos_{\text{scen}} \geq 0.996$). For nine of the eleven scenarios, the average interaction proportions closely match the generating Dirichlet weights ($\cos_{\text{scen}} = 0.973$ under both redundancy definitions). Thus, for these scenarios, the Dirichlet weights provide an accurate proxy for the expected pointwise interaction proportions.

Third, the lower values of $\cos_{\text{pw}}$ should not be interpreted as benchmark errors. Unlike $\cos_{\text{scen}}$, which compares average interaction proportions, $\cos_{\text{pw}}$ compares individual realizations, whose PID atoms fluctuate naturally because they depend on a single realization (x_1, x_2, y). Consequently, even an oracle possessing the exact densities reaches only $\cos_{\text{pw}} = 0.795$ under our definition and 0.811 under Definition B, as measured in this analytically tractable instance. These values therefore represent the intrinsic ceiling of the pointwise metric rather than estimation error. On these nine scenarios, our method reaches $\cos_{\text{pw}} = 0.747$ and LSMI 0.614, against an oracle ceiling of 0.795: our estimator attains 94% of the ceiling, versus 77% for LSMI.

Finally, the two uniqueness-dominated mixtures reveal a limitation of the benchmark construction rather than of the estimators. Convexly mixing informative U_1 and U_2 components causes both

modalities to become informative about the same target, inevitably inducing redundancy despite the absence of any shared generative latent. For example, in the 33% U_1 + 67% U_2 mixture we measure $I(X_1; Y) = 0.92$ and $I(X_2; Y) = 0.93$ nats, so any information-theoretic redundancy measure must report a non-zero overlap. Our estimator attributes it to the mechanistic component (0% source): although both modalities predict Y, neither is individually sufficient and their errors are uncorrelated across samples, so no representation satisfying the per-sample alignment constraint of Eq. 7 can carry the label. In the 50% U_1 + 25% U_2 + 25% R mixture, where a shared latent is present, the recovered redundancy is split into 61.5% source and 38.5% mechanistic. This mirrors the logic gates, where independent inputs likewise yield purely mechanistic redundancy. To our knowledge, no previous pointwise PID estimator can identify this distinction in high-dimensional settings.

We therefore regard the Dirichlet weights as a controlled proxy rather than exact pointwise ground truth. They are exact at the interaction vertices and accurately approximate the expected pointwise PID for nine of the eleven mixture types under two independent redundancy definitions. The remaining two scenarios reveal an intrinsic limitation of convex feature mixing, which our method identifies as mechanistic rather than source redundancy, providing additional empirical support for the proposed decomposition.

L.1 An intermediate source–mechanistic case

Beyond the fully shared and independent regimes of Sec. 4.4, the $50\%U_1{+}25\%U_2{+}25\%R$ mixture combines a shared latent (z_c) with redundancy induced by mixing independent uniqueness components (App. F.1). Our estimator recovers the total redundancy ($R = 0.093$ vs. 0.09) and splits it into 61.5% source and 38.5% mechanistic (Sec F.3), showing that the decomposition is not restricted to binary outputs. As the exact split is not analytically available for this mixture, this result is qualitative; an exact reference, e.g., via the partial entropy decomposition (Ince, 2017b), is left for future work.

M Role and selection of the hyperparameters

Our supervised Gács–Körner objective (Eq. 7) involves three hyperparameters: the alignment weight λ, the embedding-kernel bandwidth τ, and, for continuous targets, the target-kernel bandwidth σ_y. This appendix briefly describes their roles and how they are chosen.

M.1 Data-driven selection of the alignment weight λ

The alignment weight λ is selected using a data-driven criterion that does not require access to the unknown source/mechanistic decomposition. Specifically, we consider (i) cross-modal agreement of the source-redundancy estimates and (ii) retention of target-predictive source information after alignment. We select the smallest λ that achieves strong cross-modal agreement while retaining near-maximal source information, corresponding to the knee of the agreement–information trade-off (Fig. 4). The procedure is applied independently to each dataset; the numerical value of λ is not expected to transfer across datasets because the relative scale of the alignment and predictive terms depends on the learned representations and data distribution.

Validation on controlled data. We validate this criterion on settings with known source/mechanistic structure (Table 12). On AV-MNIST, the criterion selects $\lambda = 10$ and recovers 99.99% source attribution. Importantly, this result is stable across a broad low-λ regime: $\lambda \in 0, 1, 10$ all yield 99.99%. At larger values, source attribution decreases as increasingly strong alignment suppresses target-predictive information. The existence of this over-alignment regime is precisely what makes the criterion non-trivial: there is a wrong operating point, and the agreement–information trade-off reliably identifies and avoids it, rather than the estimate being flat and uninformative with respect to λ. On digit summation, which is purely mechanistic, the estimated source attribution remains 0% across all tested values of λ. These results validate the selection criterion in settings where the correct decomposition is known. Because the relative scale of the alignment and predictive terms is dataset-dependent, these λ values are not comparable across datasets: the criterion selects the knee at $\lambda = 10$ for AV-MNIST and at $\lambda = 25$ for HD, and the AV-MNIST degradation at $\lambda = 25$ therefore carries no implication for the HD operating point, which sits at its own knee (Fig. 4).

Robustness of the Huntington's disease analysis. On the Huntington's disease data, the criterion selects $\lambda = 25$. Repeating the analysis with $\lambda = 10$ produces an essentially identical distribution-level decomposition: source redundancy for lexical fluency changes from 29.4% to 29.5%, while all other components change by less than one percentage point. The regional SHAP-retention analysis is also stable (Table 11), with a maximum change of 12 percentage points and preserved regional ranking (Spearman $\rho = 0.93$, $p < 0.001$). By contrast, $\lambda = 0$ lies outside this regime: cross-modal disagreement between the speech- and neuroanatomy-based source-redundancy estimates is largest at $\lambda = 0$ (MAE ≈ 0.21, Fig. 4), so the per-sample agreement assumed by Proposition 1 is not met. Consistent with this, estimated source redundancy is highest at $\lambda = 0$ and decreases as agreement is enforced, as expected when redundancy that the two modalities cannot carry jointly per sample is no longer attributed to source. Thus, the main HD findings are robust within the empirically supported alignment regime.

Table 11: SHAP retention ratios for lexical fluency: $\lambda = 25$ (main text) vs. $\lambda = 10$.

Region	$\lambda = 25$	$\lambda = 10$	$\lvert\Delta\rvert$
R Putamen	60%	53%	7 pp
L Post. Cingulate G.	42%	49%	6 pp
L Putamen	70%	64%	6 pp
L Frontal Operculum	39%	37%	2 pp
R Post. Cingulate G.	57%	55%	2 pp
R Caudate	73%	61%	12 pp
L Angular Gyrus	44%	46%	2 pp
L Caudate	64%	58%	6 pp
Spearman rank correlation: $\rho = 0.93$, $p < 0.001$			

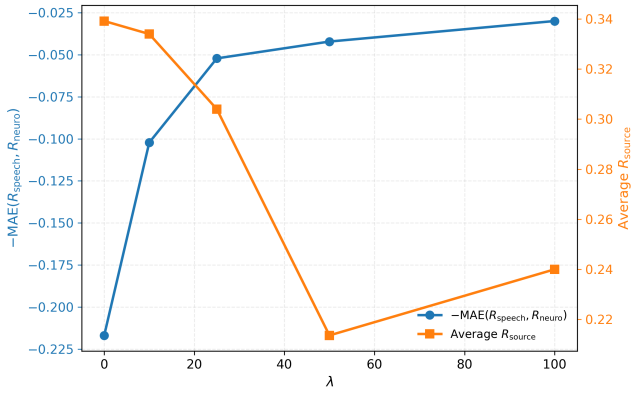


Figure 4: Agreement–information trade-off used to select λ. **Blue**: cross-modal agreement of the source-redundancy estimates (higher is better; reported as negative MAE). **Orange**: average estimated source redundancy. The selected operating point corresponds to the onset of strong agreement before further alignment substantially reduces predictive source information.

Table 12: Estimated source attribution across alignment strengths on controlled data. The selection criterion chooses $\lambda = 10$ for AV-MNIST, within the stable 99.99% source-attribution regime.

Setting	True	$\lambda = 0$	$\lambda = 1$	$\lambda = 10$	$\lambda = 25$	$\lambda = 50$	$\lambda = 100$
Digit summation	0%	0%	0%	0%	0%	0%	0%
AV-MNIST	100%	99.99%	99.99%	**99.99%**	60.8%	62.6%	68.1%

M.2 The bandwidths τ and σ_y

The remaining hyperparameters follow standard practice from supervised contrastive learning and are not tuned. The temperature τ controls the sharpness of the similarity kernel in the embedding

space. Following prior work on supervised contrastive learning (Khosla et al., 2020; Wang & Liu, 2021), we fix $\tau = 0.1$ for all experiments.

For continuous targets, σ_y defines the width of the Gaussian kernel used to measure target similarity, replacing hard positive and negative pairs with soft similarities (Dufumier et al., 2021). We set it to a standard value of $\sigma_y = 0.1$, close to the training-set standard deviation of Y (of the experiment with a continuous Y, i.e, on Huntington's disease and Alzheimer's disease, where clinical scores were re-normalized between 0 and 1).

N ANALYTIC CORROBORATION OF THE SOURCE/MECHANISTIC SPLIT

The ground truth in Section 4.4 follows from the construction is confirmed independently by the Partial Entropy Decomposition (PED) of Ince (2017b), which quantifies source and mechanistic redundancy exactly on small discrete systems. It assigns the redundancy of AND/OR (0.10 bits) and SUM (0.50 bits, the analytic counterpart of our digit summation task) entirely to the mechanistic component, since the inputs are independent in both cases. Our estimator recovers the same attribution (0% source). The PED decomposition is defined only for low-dimensional discrete variables, which is why it does not appear as a baseline elsewhere.