SYNERGY-AWARE MULTIMODAL DISENTANGLEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Disentangling a multimodal representation into factors for different information types makes it inspectable and reusable: the evidence each factor carries can be probed or recombined. Partial Information Decomposition (PID) provides natural coordinates for this, dividing task information into redundant, unique, and synergistic parts. However, to our knowledge, existing self-supervised multimodal disentanglement stops at a shared/private split, and interaction-oriented methods capture synergy only inside one fused embedding. We propose **URSyn** (**U**nique, **R**edundant, **Syn**ergistic), a self-supervised framework that learns separately accessible factors for all three types: unique factors are repelled from redundant ones, redundant factors are supported by every single modality, and the synergy factor is anchored to the full input while repelled from all single-modality references. We show that idealized versions of these constraints bound the task information of the redundant and synergy factors by redundancy- and synergy-related information quantities, without requiring PID values to be estimated. On a controlled benchmark, each factor best predicts its designated target, and the synergy factor *alone* reaches 83.5% on a joint-only task, exceeding the complete representations of CoMM (70.4%) and InfMasking (77.0%). On six real-world benchmarks, URSyn matches or exceeds our CoMM reproduction in mean performance under the same protocol and is close to InfMasking's reported results.

1 INTRODUCTION

A representation that disentangles different types of information is widely argued to be more useful than one that merely contains them (Bengio et al., 2013). In multimodal learning, this matters for three reasons. *Inspection*: when information types occupy separate factors, one can ask more directly which modality combination supports a prediction and test that claim by probing the corresponding factor. *Learning*: a factor dedicated to one information type receives its own training signal; we hypothesize that in a single entangled embedding, information that is hard to learn can be crowded out by information that is easier to align. *Reuse*: datasets differ in how much they rely on each information type (Liang et al., 2023a), and disentangled factors can be recombined accordingly (tested here only via routing), whereas an entangled embedding offers one fixed mixture.

Partial Information Decomposition (PID) provides principled coordinates for multimodal disentanglement (Williams & Beer, 2010; Harder et al., 2013; Bertschinger et al., 2014). It divides the information that modalities carry about a target into *redundant* information R, obtainable from any single modality; *unique* information U_i, obtainable only from modality i; and *synergistic* information S, obtainable only when the modalities are observed jointly. The object category mentioned in both an image and its caption is typically redundant, vocal tone can be unique to audio, and sarcasm can emerge from the mismatch between words and facial expression (Castro et al., 2019).

To our knowledge, existing self-supervised methods disentangle only part of this structure. Multimodal disentanglement methods split representations into shared and private factors, e.g., MISA (Hazarika et al., 2020), FactorCL (Liang et al., 2023b), QUEST (Song et al., 2024), DisentangledSSL (Wang et al., 2025), DeCUR (Wang et al., 2024), and DecAlign (Qian et al., 2026); this separates R from U_i but gives no factor to information that no single modality carries. Interaction-oriented methods such as CoMM (Dufumier et al., 2025) and InfMasking (Wen et al., 2025) do capture synergy, but entangled with redundant and unique information inside one fused embedding Z.

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.

Synergy is plausibly the hardest information type to learn, since no single modality exposes it, and it is also the one that current disentanglement leaves out. On a controlled benchmark, our disentangled synergy factor alone predicts a joint-only target better than the complete representations of CoMM and InfMasking (Sec. 4.3).

Disentangling synergy is not straightforward: exact PID values require the target Y and an optimization over joint distributions (Bertschinger et al., 2014), and Y is unobserved in self-supervised pretraining. URSyn therefore does not estimate PID values; it approximates the PID *access pattern* (available from every modality, from one modality, or only jointly) with a representation-level surrogate. Using singleton and full-modality references from a shared backbone, it learns a unique factor per modality, a family of redundant factors supported by every singleton view (pooled into Z_R), and a synergy factor anchored to the full view while repelled from all singleton references. The output is a disentangled set of factors $Z_{U_1}, \ldots, Z_{U_M}, Z_R, Z_S$ that can be probed or recombined individually. We adopt a *target-selective* notion of disentanglement: each factor should be the best single-factor carrier of its designated information type. The factors are linearly selective, not independent or exclusive: concatenating factors from different modalities can re-form joint-only evidence (Sec. 4.3).

Our contributions are: **(1)** URSyn, a self-supervised method designed to separate unique, redundant, *and* synergistic information into three separately accessible types of factors, whereas shared/private methods have no synergy factor and CoMM and InfMasking keep all types entangled in one embedding. URSyn couples unique/redundant objectives with a singleton-repelled, full-view-anchored synergy objective. **(2)** An accessibility-based formulation of this disentanglement, with a proposition showing what the idealized constraints guarantee, including when the bound reduces to the synergy term and a tight example of when it does not (Sec. 2.2). **(3)** Evidence for the usefulness of typed factors: single-factor cross-probing on a controlled benchmark shows target-selective factors; the disentangled synergy factor alone outperforms the complete representations of CoMM and InfMasking on a joint-only target; and sample-wise recomposition (Relative-Diversity Routing, downstream only, no learned parameters) raises the mean accuracy of URSyn factors (per-dataset gains within one standard deviation) but not of CoMM's views. We compare with CoMM on seven benchmarks and with InfMasking on Trifeature and via its reported real-world results (Appendix C.4).

2 PID-Based Formulation of Multimodal Disentanglement

2.1 Preliminaries

PID and its access pattern. Let $\mathbf{X} = (X_1, \ldots, X_M)$ be an M-modal input and Y a downstream target. For two modalities, PID writes (Williams & Beer, 2010; Bertschinger et al., 2014)

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

Eq. (1) has one more unknown than equations, so a PID requires an additional definition. The BROJA decomposition (Bertschinger et al., 2014) fixes it by

$$S = I_P(X_1, X_2; Y) - \min_{Q \in \Delta_P} I_Q(X_1, X_2; Y), \qquad \Delta_P = \{Q : Q(X_i, Y) = P(X_i, Y),\ i = 1, 2\}, \tag{2}$$

which requires Y and an optimization over Q. Neither is available during self-supervised pretraining, so URSyn does not compute Eq. (2). What Eq. (1) does determine, independently of the choice of PID measure, is the access pattern: singleton modality i can access $\mathcal{A}_i = R + U_i$, whereas the full input can access $\mathcal{A}_J = R + \sum_i U_i + S$. We use $\mathcal{A}_i$ and $\mathcal{A}_J$ as shorthand for these access relations, not as set algebra over bits. For $M > 2$, the PID lattice has many more atoms and no single decomposition is agreed upon. We then use R, U_i, and S as coarse access groups, not PID atoms. S is the information that no single modality provides, including synergy within strict subsets, i.e., the union-information synergy $S_\cup = I(\mathbf{X}; Y) - I_\cup$ with $I_\cup = \min_{Q \in \Delta_P} I_Q(\mathbf{X}; Y)$ and Δ_P preserving every $P(X_i, Y)$ (Griffith & Koch, 2014); for $M = 2$, $S_\cup$ is the BROJA S of Eq. (2). Information shared by a strict subset of modalities is left to the unique factors, since Z_R must be predictable from every modality. Throughout, "synergy" means this access-defined quantity: the BROJA S for $M = 2$ and the broader $S_\cup$ for $M > 2$. The synergy factor Z_S is designed to concentrate it; neither the idealized constraint nor the loss guarantees this (Proposition 1(c)).

Self-supervised surrogate for the target. Following Dufumier et al. (2025), we replace Y with a second label-preserving view. Let $t \in \mathcal{T}^\star$ be a multimodal augmentation with $\mathbf{X}' = t(\mathbf{X})$. Idealizing

the multi-view assumptions of Tsai et al. (2021) as in Dufumier et al. (2025), we assume that the view retains the task information, $I(\mathbf{X}; Y \mid \mathbf{X}') = 0$, and that the input and its view share no task-irrelevant information, $I(\mathbf{X}; \mathbf{X}' \mid Y) = 0$. Since $I(\mathbf{X}; \mathbf{X}') = I(\mathbf{X}; Y) + I(\mathbf{X}; \mathbf{X}' \mid Y) - I(\mathbf{X}; Y \mid \mathbf{X}')$, these assumptions imply $I(\mathbf{X}; \mathbf{X}') = I(\mathbf{X}; Y)$ and, unlike this equality alone, make the information shared by the views task-relevant. We assume the same for any pair of augmented views $(\mathbf{X}', \mathbf{X}'')$, which is the pair used by the contrastive objectives. For an encoder f_θ, the data processing inequality gives $I(f_\theta(\mathbf{X}'); f_\theta(\mathbf{X}'')) \leq I(\mathbf{X}'; \mathbf{X}'')$, so maximizing agreement between two views targets the information they share, which under these assumptions is task-relevant. Under the analogous assumptions for singleton views, the same argument targets what is accessible from modality i alone. Pure cross-modal alignment (Tian et al., 2020; Radford et al., 2021) additionally relies on multi-view redundancy, $I(Y; X_1 \mid X_2) \approx I(Y; X_2 \mid X_1) \approx 0$ (Liang et al., 2023b), and therefore tends to discard U_i and S.

2.2 What idealized access constraints guarantee

URSyn is motivated by two idealized constraints: a redundant factor should be recoverable from every singleton, and a synergy factor should share no information with any singleton. The following proposition (proof in Appendix A) states what these constraints imply about task information for an *arbitrary* target Y.

Proposition 1 (Idealized access constraints)**.** *Let Z_R and Z_S be (possibly randomized) functions of $\mathbf{X}$, with randomness independent of Y given $\mathbf{X}$. **(a) Common-part constraint.** If Z_R is a deterministic function of X_i for every i, then $I(Z_R; Y) \leq I_\wedge(X_1, \ldots, X_M \to Y) \leq \min_i I(X_i; Y)$, where $I_\wedge$ is the common-randomness redundancy of Griffith et al. (2014), built on the Gács–Körner common variable (Gács et al., 1973). **(b) Singleton-exclusion constraint.** If $I(Z_S; X_i) = 0$ for all i, then for every i, $I(Z_S; Y) \leq I(Z_S; Y \mid X_i) \leq I(\mathbf{X}; Y) - I(X_i; Y)$; that is, the co-information $I(Z_S; X_i; Y) = I(Z_S; Y) - I(Z_S; Y \mid X_i)$ is nonpositive, so Z_S and every singleton are not net-redundant about Y. For any M, this gives $I(Z_S; Y) \leq S_\cup + [I_\cup - \max_i I(X_i; Y)]$, where the bracket is nonnegative; for $M = 2$ and any PID satisfying Eq. (1), $I(Z_S; Y) \leq S + \min(U_1, U_2)$. **(c) Tightness.** The bound in (b) is attained: for independent uniform bits X_1, X_2 and $Y = (X_1, X_2)$, BROJA gives $R = S = 0$, $U_1 = U_2 = 1$, while $Z_S = X_1 \oplus X_2$ satisfies the exclusion constraint and $I(Z_S; Y) = 1$.*

Proposition 1 gives PID-related upper bounds for the idealized constraints. Exact singleton exclusion would guarantee that a synergy factor carries no task information beyond $S + \min_i U_i$ (for $M = 2$); whenever some modality has no unique information about Y, the bound reduces to $I(Z_S; Y) \leq S$, and otherwise its slack above S is $\min(U_1, U_2)$. The proposition gives no lower bound: preserving S is left to full-view anchoring, which carries no guarantee either. The trainable losses of Sec. 3 further relax these constraints, as summarized in Table 1.

3 Method

Given two label-preserving views $\mathbf{X}'$ and $\mathbf{X}''$, a CoMM-style backbone (Dufumier et al., 2025) produces singleton references Z_i', Z_i'' by exposing only modality i, and full references Z', Z'' from all modalities (Fig. 1). Typed branches on top of the shared backbone output Z_{U_i}, Z_{R_i}, and Z_S. $\mathcal{L}_{\text{NCE}}$ denotes the symmetric InfoNCE loss (Oord et al., 2018) (Appendix E), $\widehat{H}(\cdot)$ an entropy-gradient surrogate (Appendix B), and $\rho(A, B) = \frac{1}{N}\sum_{n=1}^{N} \max(0, \cos(a_n, b_n) - \epsilon)$ a cosine hinge with margin ϵ over a batch $A = \{a_n\}_{n=1}^N$, $B = \{b_n\}_{n=1}^N$, and $\langle A, B \rangle = \frac{1}{N}\sum_n a_n^\top b_n$.

3.1 Coupled unique and redundant factors

Each singleton reference can expose at most $\mathcal{A}_i = R + U_i$ (for $M > 2$, in the coarse sense above), so matching a branch to Z_i alone cannot separate shared from modality-specific content. URSyn therefore learns the two jointly. The unique branch first preserves singleton-accessible information,

$$\mathcal{L}_U = \frac{1}{2M} \sum_{i=1}^{M} \left[\mathcal{L}_{\text{NCE}}(Z_i', Z_{U_i}'') + \mathcal{L}_{\text{NCE}}(Z_i'', Z_{U_i}') \right]. \tag{3}$$

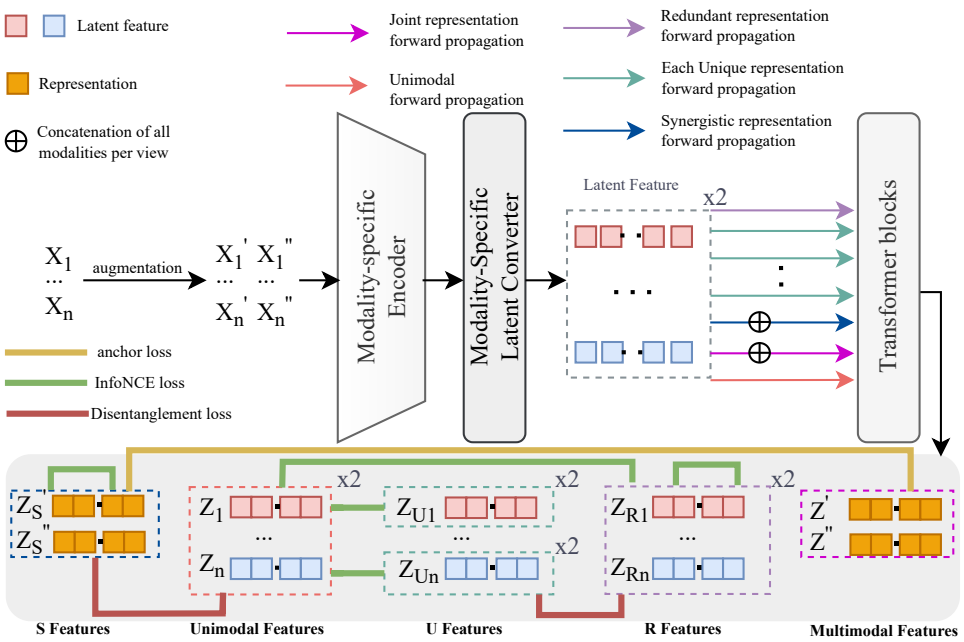

Figure 1: Pretraining pipeline of URSyn. Two augmented views $\mathbf{X}', \mathbf{X}''$ are processed by modality-specific encoders and latent converters to obtain latent features. Shared Transformer blocks are evaluated on each singleton input and on the concatenated all-modality input, producing singleton references $Z_1, \ldots, Z_M$ and joint references Z', Z''. These references supervise the unique factors Z_{U_i}, redundant factors Z_{R_i}, and the synergy factor Z_S, while the joint path also supplies the multimodal representations shown at the right. The propagation arrows distinguish unimodal, joint, unique, redundant, and synergistic forward paths. In the loss legend, yellow denotes the cross-view anchor $-\lambda_a[\langle Z'_S, Z''\rangle + \langle Z''_S, Z'\rangle]$, green denotes InfoNCE preservation, and red denotes the disentanglement penalties. Relative-Diversity Routing (Sec. 3.4) is applied *only after pretraining* on frozen factors and has no effect on representation learning.

Redundant branches must be predictable from *every* singleton reference, which is motivated by, but does not enforce, the common-part constraint of Proposition 1(a), and must agree with each other:

$$\begin{aligned}
\mathcal{C}_{R,i} &= \frac{1}{2M}\sum_{j=1}^{M}\left[\mathcal{L}_{\mathrm{NCE}}(Z'_i, Z''_{R_j}) + \mathcal{L}_{\mathrm{NCE}}(Z''_i, Z'_{R_j})\right], \\
\mathcal{P}_{R,ij} &= \frac{1}{2}\left[\mathcal{L}_{\mathrm{NCE}}(Z'_{R_i}, Z'_{R_j}) + \mathcal{L}_{\mathrm{NCE}}(Z''_{R_i}, Z''_{R_j})\right], \\
\mathcal{L}_R &= \frac{1}{M + \binom{M}{2}}\left(\sum_{i=1}^{M}\mathcal{C}_{R,i} + \sum_{i<j}\mathcal{P}_{R,ij}\right).
\end{aligned} \tag{4}$$

We average over the M consistency and $\binom{M}{2}$ pair entries together. Thus the two groups are weighted by their entry counts (for $M = 2$, $2:1$), rather than normalized to equal group weight. To separate U_i from R, directly minimizing $I(Z_{U_i}; Z_{R_j})$ is unsafe because a collapsed representation also has low mutual information. Motivated by $I(A;B) = H(A) - H(A \mid B)$, we keep Z_{U_i} high-entropy and penalize its angular overlap with the redundant family:

$$\mathcal{L}_{UR} = \frac{1}{M}\sum_{i=1}^{M}\Big\{ -\lambda_h\left[\widehat{H}(Z'_{U_i}) + \widehat{H}(Z''_{U_i})\right] + \frac{\lambda_p}{2M}\sum_{j=1}^{M}\left[\rho(Z'_{U_i}, Z'_{R_j}) + \rho(Z''_{U_i}, Z''_{R_j})\right]\Big\}. \tag{5}$$

The hinge is inactive once the cosine similarity falls below ϵ, and it updates both branches; it is intended to push unique factors away from the consensus and redundant factors away from modality-specific residue.

Table 1: Ideal constraints targeted by URSyn and the trainable terms that relax them. R, U_i, and S are defined by PID access relations and are not observed; the loss is a relaxation, not a PID estimator.

Factor	Ideal constraint	Trainable term	Role
Z_{U_i}	preserve $\mathcal{A}_i$: $\max I(Z_{U_i}; X_i)$	$\mathcal{L}_U$ (InfoNCE)	preservation
	$I(Z_{U_i}; Z_R) = 0$ without collapse	$-\lambda_h \widehat{H}$ and $\lambda_p \rho(Z_{U_i}, Z_{R_j})$ in $\mathcal{L}_{UR}$	anti-collapse, overlap
Z_R	$Z_R = g_i(X_i)\ \forall i$ (Prop. 1a)	$\mathcal{C}_{R,i}$, $\mathcal{P}_{R,ij}$ in $\mathcal{L}_R$	consensus
Z_S	preserve $\mathcal{A}_J$	$\mathcal{L}_S$, $-\lambda_a \langle Z_S, Z \rangle$	preservation, anchoring
	$I(Z_S; X_i) = 0\ \forall i$ (Prop. 1b)	$\lambda_p \rho(Z_S, Z_i)$, $-\lambda_h \widehat{H}(Z_S)$	penalty, anti-collapse

3.2 Singleton-repelled synergy factor

Only the full path can access synergy, so Z_S is computed from the joint input and preserves augmentation-stable full-view information, $\mathcal{L}_S = \mathcal{L}_{\mathrm{NCE}}(Z'_S, Z''_S)$. On its own, this would yield a generic fused representation of $\mathcal{A}_J$. Motivated by the singleton-exclusion constraint of Proposition 1(b), URSyn additionally penalizes the alignment of Z_S with every singleton reference, while anchoring Z_S to the opposite full view and counteracting collapse with an entropy term:

$$\mathcal{L}_{S\perp} = -\lambda_h \big[\widehat{H}(Z'_S) + \widehat{H}(Z''_S)\big] - \lambda_a \big[\langle Z'_S, Z'' \rangle + \langle Z''_S, Z' \rangle\big] + \frac{\lambda_p}{2M} \sum_{i=1}^{M} \big[\rho(Z'_S, Z'_i) + \rho(Z''_S, Z''_i)\big]. \tag{6}$$

The anchor term is intended to keep Z_S grounded in joint semantics, and the hinge penalizes per-sample similarities $\cos(z_S, z_i) > \epsilon$ for all singleton references, which ideally cover $R + \sum_i U_i$; its gradient also updates Z_i. This is a geometric proxy rather than an exclusion guarantee: the hinge does not penalize anti-aligned or perpendicular embeddings (e.g., $z_S = -z_i$ and $z_S = Jz_i$ with invertible skew-symmetric J both have zero penalty although each carries the same information as z_i), so whether singleton information is actually removed must be checked empirically (Secs. 3.3 and 4.3). The pretraining objective is

$$\mathcal{L}_{URSyn} = \mathcal{L}_U + \lambda_R \mathcal{L}_R + \lambda_{UR} \mathcal{L}_{UR} + \lambda_S \mathcal{L}_S + \lambda_{S\perp} \mathcal{L}_{S\perp}. \tag{7}$$

Appendix B maps these symbols to the configuration keys; $\lambda_a = 1$ in all main experiments.

3.3 From ideal constraints to a trainable surrogate

Table 1 separates what URSyn ideally targets from what it optimizes. The surrogate has three gaps: InfoNCE only lower-bounds mutual information; the singleton hinge acts on augmentation-stable references Z_i and is not an MI bound; and capacity and optimization are finite. Hence the gap to BROJA synergy cannot be bounded in general, and we evaluate the factors empirically (Sec. 4.3).

Why a geometric separation term. Geometric separation is common in disentangled learning: Domain Separation Networks and MISA impose soft orthogonality (Bousmalis et al., 2016; Hazarika et al., 2020), DMD and FOCAL use cosine penalties (Li et al., 2023; Liu et al., 2023), and DeCUR uses a correlation penalty in the spirit of Barlow Twins/VICReg (Wang et al., 2024; Zbontar et al., 2021; Bardes et al., 2022). Geometric terms are practical because MI upper bounds such as CLUB require density learning (Cheng et al., 2020) and MI is difficult to estimate from finite samples (McAllester & Stratos, 2020); FOCAL also reports degradation, and possible collapse, when replacing geometric orthogonality with distributional independence (Liu et al., 2023). The cosine hinge ignores anti-aligned and nonlinear dependence, so we test separation by single-factor probing (Sec. 4.3).

Relation to prior objectives. Table 2 summarizes the difference in what is learned. CoMM optimizes full–full and singleton–full agreement, preserving mixtures $R + \sum_i U_i + S$ and $R + U_i$ in one fused Z; its singleton views Z_i serve only as supervisory references, are not trained to carry a specific information type, and are not part of its standard downstream output. InfMasking also targets synergy, but strengthens it *inside* that fused Z by aligning randomly masked with unmasked fusions, whereas URSyn aims to *concentrate* it in a dedicated factor by penalizing its alignment with singleton references, using deterministic singleton masks as accessibility sets for Z_{U_i}, Z_R, and Z_S. URSyn thus moves from capturing interactions to disentangling them, enabling factor-level tests unavailable to one fused Z (Appendix C.2).

Table 2: What each self-supervised method outputs. ✓: a separately accessible factor for that information type; "mixed": the information is captured but entangled with other types in one representation; –: not modeled as a target of the objective; *: linearly selective, non-exclusive.

Method	R	U_i	S	Output used downstream
FactorCL (Liang et al., 2023b)	✓	✓	–	shared + unique factors
DisentangledSSL (Wang et al., 2025)	✓	✓	–	shared + modality-specific factors
CoMM (Dufumier et al., 2025)	mixed	mixed	mixed	one fused Z
InfMasking (Wen et al., 2025)	mixed	mixed	mixed	one fused Z
URSyn (ours)	✓*	✓*	✓*	$Z_{U_1}, \ldots, Z_{U_M}, Z_R, Z_S$

3.4 Relative-Diversity Routing (downstream only)

After pretraining, the encoder and branches are frozen, and a downstream probe must combine the factors. Concatenation followed by a linear probe applies coefficients that are fixed across samples. Because the factors are typed, a simple sample-dependent reweighting is available. We pool $Z_R = \frac{1}{M}\sum_i Z_{R_i}$, form $\mathcal{Z} = \{Z_{U_1}, \ldots, Z_{U_M}, Z_R, Z_S\}$, and for each sample x compute

$$d_k(x) = -\sum_{j \neq k} \mathrm{sim}\left(v_k(x), v_j(x)\right), \qquad w_k(x) = \mathrm{softmax}_k\left(d_k(x)/\tau_r\right), \\ F(x) = \mathrm{Concat}\left(w_1(x)v_1(x), \ldots, w_K(x)v_K(x)\right), \tag{8}$$

where $v_k(x)$ denotes the k-th factor in $\mathcal{Z}$ for sample x (distinct from the singleton references Z_i), $K = M + 2$, sim is cosine similarity between frozen pre-projection factors, and a single global temperature $\tau_r = 0.5$, fixed a priori rather than tuned, is used for all datasets. Intuitively, a factor similar to the others for a given sample is largely covered by them and receives less weight. Three properties of this simple rule should be noted: because d_k sums $K - 1$ similarities, its scale grows with the number of modalities, so the same τ_r corresponds to different effective sharpness in the two- and three-modality settings; a factor anti-aligned with another is scored as diverse even if it carries redundant information; and a factor that is intrinsically similar to the others, such as Z_R, tends to be down-weighted regardless of its task relevance. The rule has no learned parameters. A linear probe on the unrouted concatenation cannot reproduce it, since $\sum_k W_k w_k(x) v_k(x)$ involves a sample-dependent multiplicative interaction that fixed coefficients W_k cannot represent. We motivate the rule by the conditional gain $I(v_k; Y \mid v_{\setminus k})$, but relative diversity is *not* an estimator of that quantity; routing is a secondary, empirically tested use of the typed factors (Sec. 4.4).

4 Experiments

Setup. We evaluate on Trifeature (Hermann & Lampinen, 2020), five MultiBench datasets (Liang et al., 2021), and MM-IMDb (Arevalo et al., 2017) with frozen linear probes and seeds 42–46. Baselines are Cross, Cross+Self, FactorCL, DisentangledSSL, CMC, CoMM, and InfMasking, plus SimCLR, CLIP, and SLIP (Mu et al., 2022) on MM-IMDb. Markers †/♢ denote numbers reported in the CoMM/DisentangledSSL papers; unmarked CoMM numbers are our public-code reproduction under its protocol, with paper-reported values in Appendix C.3. InfMasking is reproduced on Trifeature with its public configuration; its reported real-world results are in Appendix C.4. *Comparison with CoMM.* The methods share encoders, Transformer, and projection head; they differ in objective, typed branches, coefficients, and routing. Unrouted features of approximately matched dimension provide a partial probe-size control (Sec. 4.4). Real-world results use Relative-Diversity Routing; Trifeature uses concatenation. Augmentations, hyperparameters, the selection protocol, and training cost are in Appendix B.

4.1 Controlled benchmark: Trifeature

Trifeature targets are designed to depend on redundant, unique, or joint-only information, so the benchmark directly tests our claim, although both of its modalities are images. Redundancy is saturated for all contrastive methods (Table 3); on the joint-only target, URSyn reaches 83.56% vs. 70.41% for CoMM and 77.00% for InfMasking. InfMasking is stronger on uniqueness (90.55% vs.

Table 3: Linear probing accuracy (%) on Trifeature. Red.: shape; Uniq.: texture (mean of two probes); Syn.: joint texture–color target (chance 50%). URSyn uses all-factor concatenation; single-factor results are in Table 7. † rows are taken from CoMM and may differ in setup. InfMasking Uniq. averages its two probe means (Table 7), so no standard deviation is given; URSyn Uniq. pools all ten probes.

Model	Red.	Uniq.	Syn.
Cross†	**100.0**	11.6	50.0
Cross+Self†	99.7	86.9	50.0
FactorCL†	99.8	62.5	46.5
CoMM	99.93±0.04	87.23±2.40	70.41±1.90
InfMasking	99.90±0.09	**90.55**	77.00±4.22
URSyn	99.90±0.10	87.76±1.81	**83.56±1.97**

Table 4: Linear probing on MultiBench: MSE ($\times 10^{-4}$, lower is better) for Vision&Touch (V&T) end-effector (EE) regression and accuracy (%) for classification. The average is over classification tasks. †/♢ rows are taken from the respective papers and may differ in backbone and training setup; CoMM and URSyn share backbone and protocol. Comparison with InfMasking: Appendix C.4.

Model	V&T EE ↓	MIMIC ↑	MOSI ↑	UR-FUNNY ↑	MUSTARD ↑	Avg. ↑
Cross†	33.09±3.67	66.70±0.10	47.80±1.80	50.10±1.90	53.50±2.90	54.52
Cross+Self†	7.56±0.31	65.49±0.00	49.00±1.10	59.90±0.90	53.90±4.00	57.07
FactorCL†	10.82±0.56	67.30±0.00	51.20±1.60	60.50±0.80	55.80±0.90	58.70
DisentangledSSL♢	–	66.44	65.16	64.24	61.60	64.36
CoMM	7.65±1.83	66.47±0.40	64.46±3.05	62.82±1.10	63.35±6.49	64.28
URSyn	**5.24±1.26**	**68.00±0.20**	**67.06±1.20**	**64.93±0.62**	**66.99±2.48**	**66.75**

87.76%), so the clearest gain is on joint-only information; Sec. 4.3 shows that this information is most directly decodable from Z_S.

4.2 Real-world benchmarks

On the real-world benchmarks (Tables 4–6), URSyn has higher mean accuracy than our CoMM reproduction on every two-modality dataset and MM-IMDb (within one CoMM standard deviation on MOSI and MUSTARD), ties it on three-modality V&T contact prediction, raises the classification average from 64.28% to 66.75%, and lowers V&T EE error from 7.65×10^{-4} to 5.24×10^{-4} (CoMM reports 4.55×10^{-4}; Appendix C.3). It is close to InfMasking's reported average (66.75% vs. 67.05%; Appendix C.4): InfMasking leads on MOSI by 1.94 points and on V&T EE (4.23 vs. 5.24×10^{-4} MSE), while other gaps in either direction are at most 0.63 points; per-seed InfMasking results are unavailable.

Similar overall accuracy, and where the methods differ. Both InfMasking and URSyn add objectives that contrast masked or singleton inputs with the full input; we hypothesize that this shared ingredient is why their complete-representation accuracy is similar. The largest gap in favor of URSyn is on the joint-only Trifeature target, where URSyn reaches 83.56% vs. 77.00% for InfMasking, while Z_S alone reaches 83.46% with lower variance (Sec. 4.3).

4.3 Factor-level analysis

A single fused embedding does not expose which access path carries information. We therefore test target selectivity and subset sufficiency on the frozen checkpoints of all five seeds.

Single-factor cross-probing. With one factor per probe (Table 7, top), each target is best predicted by its designated factor, by margins of 75.15/54.22/63.62/33.46 points for R/U_1/U_2/S. Z_{U_1}, Z_{U_2}, and the pooled Z_R are at chance on the joint-only target (49.95–50.00%), whereas Z_S alone reaches 83.46%, close to all factors (83.56%). Selectivity is imperfect: Z_S predicts shape and texture at about 24%, Z_{U_1} predicts shape at 25%, and Z_R predicts U_1 at 33%.

Table 5: Accuracy (%) with two and three modalities; V&T CP: contact prediction (modalities in Appendix D). Comparison with InfMasking: Appendix C.4.

Model	#Mod.	V&T CP	UR-FUNNY
Cross	2	85.82±0.31	50.10†
Cross+Self	2	86.74±0.55	59.90†
CoMM	2	86.14±1.28	62.82±1.10
URSyn	2	**88.76±0.74**	**64.93±0.62**
CMC†	3	**94.10**	59.20
CoMM	3	**94.10±0.10**	64.76±1.35
URSyn	3	**94.10±0.05**	**65.58±0.88**

Table 6: Weighted/macro F1 (%) on MM-IMDb (V+L). △: extra unlabeled pretraining on MM-IMDb; CLIP without △ uses the original pretrained encoders. Comparison with InfMasking: Appendix C.4.

Model	weighted-F1	macro-F1
SimCLR†△ (V only)	40.35±0.23	27.99±0.33
CLIP†	58.90	50.90
SLIP†△	56.54±0.19	47.35±0.27
CLIP†△	54.49±0.19	44.94±0.30
CoMM (CLIP)	60.72±0.37	53.48±0.38
URSyn (CLIP)	**62.62±0.67**	**55.76±0.67**

Table 7: Trifeature accuracy (%) from probing *one* URSyn factor (top), InfMasking's complete fused representation (middle, InfM.), or URSyn concatenations after post-hoc factor removal (bottom). Chance: 10% for R/U_1/U_2, 50% for S. $Z_R = M^{-1}\sum_i Z_{R_i}$.

Input	R	U_1	U_2	S
Z_{U_1}	24.82±1.44	**87.57±2.99**	6.54±0.59	49.97±0.05
Z_{U_2}	13.96±1.12	5.19±0.31	**86.71±2.96**	49.95±0.09
Z_R	**99.97±0.02**	33.35±1.52	22.88±1.36	50.00±0.00
Z_S	24.31±1.28	23.72±1.44	23.09±1.19	**83.46±1.15**
Margin	+75.15	+54.22	+63.62	+33.46
InfM. Z	99.90±0.09	90.30±1.52	90.80±2.88	77.00±4.22
All factors	99.90±0.10	87.04±1.70	88.47±1.41	83.56±1.97
w/o Z_R	99.87±0.07	86.70±3.12	88.02±1.14	82.02±2.58
w/o Z_{U_1}	99.80±0.26	86.65±1.45	86.84±2.61	83.84±2.00
w/o Z_{U_2}	99.92±0.04	88.00±2.32	86.25±1.90	80.58±1.54
w/o Z_S	99.87±0.12	86.70±2.90	84.82±3.73	70.20±12.56

Separated factor versus fused embedding. Under the same probe, Z_S (83.46±1.15%) exceeds complete InfMasking (77.00±4.22%) and CoMM (70.41±1.90%) on the joint-only target with lower variance. Z_S thus makes the target linearly decodable from one factor, although InfMasking's fused Z retains more texture information (90.30/90.80 vs. 87.57/86.71%); the role of the separation terms is not isolated.

Post-hoc factor removal. The bottom of Table 7 removes one factor from the concatenation of a trained model and refits the probe; it tests subset sufficiency, not the effect of removing an objective. Removing Z_R, Z_{U_1}, or Z_{U_2} changes accuracy on that factor's target by at most 2.2 points (99.87/86.65/86.25%): this information remains decodable from the other factors, so it is not stored exclusively in the designated factor. Removing Z_S lowers synergy accuracy to 70.20% with high variance, but accuracy stays above chance, likely because Z_{U_1} and Z_{U_2} come from different modalities and their concatenation is a joint input. One explanation, not tested directly, is that a linear probe only partially reads this texture–color conjunction from separate factors, whereas Z_S can form it through nonlinear fusion (83.46±1.15%). Thus the partition is not atom-wise exclusive, but Z_S remains the most direct and stable carrier (13.26 points above the non-S subset, whose std is 12.56).

4.4 Routing analysis

Table 8 evaluates routing on frozen encoders (standard deviations in Appendix C.5). For CoMM, we concatenate its M singleton views and fused Z from unchanged checkpoints ($M{+}1$ vs. $M{+}2$ vectors, so dimensions match only approximately). Unrouted URSyn factors have higher mean accuracy than the concatenation of CoMM's views in all four settings (65.30 vs. 64.65 average; within one standard deviation) with the same probe; routing adds 0.84 points to URSyn but lowers the routed CoMM views by 0.22. This suggests that the routing gain is not a generic reweighting effect, although all per-dataset differences are within one standard deviation and this control does not isolate disentanglement. With an MLP probe on MOSI, UR-FUNNY, MUSTARD, and MIMIC,

Table 8: Accuracy (%) with (+div.) and without (Raw) Relative-Diversity Routing on frozen features, with the same rule applied to CoMM singleton/full views. UR-FUNNY$_3$: three modalities.

Features	MOSI	UR-FUNNY	MUSTARD	UR-FUNNY$_3$	Avg.
CoMM, Raw	65.03	64.04	64.60	64.91	64.65
CoMM, +div.	64.62	63.64	64.46	64.98	64.43
URSyn, Raw	65.86	64.42	65.44	65.46	65.30
URSyn, +div.	**67.06**	**64.93**	**66.99**	**65.58**	**66.14**

routing still improves all four means, while $\tau_r \in [0.1, 1]$ changes the average by at most 1.22 points (Appendix C.5).

Sensitivity. On UR-FUNNY, 27 of 30 non-default settings of the eight coefficients exceed CoMM (62.82%); on MOSI, only 16 of 27 non-default settings exceed CoMM (64.46%), so sensitivity is dataset-dependent (Appendix C.6).

Objective ablation. We retrain five objective variants on MOSI and two-modality UR-FUNNY and report overall linear-probe accuracy (Appendix C.1). Removing the entropy term gives the largest average drop and removing the pairwise-agreement term $\mathcal{P}_R$ the smallest; yet removing ρ or $\mathcal{P}_R$ changes accuracy by under one standard deviation on both datasets, as do most UR-FUNNY variants.

5 Related Work

Multimodal self-supervised representation learning. Cross-view agreement methods, from early joint encoders and CLIP/ALIGN to fusion-based and many-modality objectives (Ngiam et al., 2011; Srivastava & Salakhutdinov, 2012; Radford et al., 2021; Jia et al., 2021; Li et al., 2021; Singh et al., 2022; Girdhar et al., 2023), mainly learn shared, entangled representations.

Multimodal interaction and synergy modeling. PID provides the language for multimodal interactions (Williams & Beer, 2010; Harder et al., 2013; Bertschinger et al., 2014; Griffith et al., 2014; Kolchinsky, 2022) and has also been used to analyze self-supervised learning (Mohamadi et al., 2024). C-MCR and LoReTTa connect modality pairs not observed together through additional cross-modal objectives (Wang et al., 2023; Tran et al., 2023), while CoMM and InfMasking capture interactions in a fused embedding through augmented views or masking (Dufumier et al., 2025; Wen et al., 2025); in all of them, synergy remains implicit. URSyn instead gives synergy an explicit full-view factor repelled from singleton references.

Disentangled multimodal representation learning. Disentanglement is a long-standing goal of representation learning (Bengio et al., 2013; Locatello et al., 2019). Shared/private methods such as MISA, FactorCL, QUEST, DisentangledSSL, and FOCAL separate invariant or unique subspaces (Hazarika et al., 2020; Liang et al., 2023b; Song et al., 2024; Wang et al., 2025; Liu et al., 2023); further shared/private decoupling methods address modality heterogeneity, missing modalities, or inconsistency (Wang et al., 2024; Yao et al., 2024; Qian et al., 2026; Lin & Hu, 2023; Li et al., 2023), while DnR uses supervised $U/R/S$ embeddings (Mai et al., 2026). URSyn uses a PID-based self-supervised objective to make all three factors separately accessible and reusable.

6 Conclusion

We presented URSyn, a self-supervised method that makes unique, redundant, and synergistic information separately accessible by turning the PID access pattern into trainable constraints. On a controlled benchmark, its factors are target-selective and the synergy factor alone predicts a joint-only target better than the complete representations of CoMM and InfMasking; on real data, its mean performance matches or exceeds our CoMM reproduction and is close to InfMasking's reported results. The separation terms are geometric proxies rather than information-theoretic guarantees, and direct evidence of factor selectivity currently comes from a controlled benchmark; stronger dependence controls and factor-level evaluation on real-world tasks are natural next steps.

Ethics Statement

URSyn is a representation learning method, not a deployed decision system. All datasets are public benchmarks used under their licenses and standard preprocessing; MIMIC-III is de-identified and accessed under its credentialed data use agreement. Improved predictive accuracy does not guarantee fairness, calibration, privacy, or safety; in clinical or behavioral applications, models built on URSyn require task-specific auditing and human oversight. Factor-level access may help such auditing by showing which modality combinations carry decodable information, although factor selectivity has so far been verified only on a synthetic benchmark.

Reproducibility Statement

The objective is specified in Eqs. (3)–(7), routing in Eq. (8), and the training loop in Algorithm 1. Appendix B lists augmentations for each modality, the regularization coefficients and learning rates, the mapping of all loss coefficients to configuration keys, the selection protocol, and compute. Dataset splits and preprocessing are described in Appendix D, and the proof of Proposition 1 is in Appendix A. Code will be made publicly available upon acceptance.

AI Use Statement

We used generative AI tools to polish the language of the manuscript and improve its readability. All mathematical claims and proofs were checked by the authors, and every cited reference was verified against the original publication. We did not use generative AI tools to design the method, implement code, run experiments, generate data, or produce experimental results; all results were obtained by the authors. We have reviewed all AI-assisted content and take full responsibility for the final content of this work.

References

John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A González. Gated multimodal units for information fusion. *arXiv preprint arXiv:1702.01992*, 2017.

Adrien Bardes, Jean Ponce, and Yann Lecun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. In *ICLR 2022-International Conference on Learning Representations*, 2022.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

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

Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. *Advances in neural information processing systems*, 29, 2016.

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*, pp. 4619–4629, 2019.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020.

Pengyu Cheng, Weituo Hao, Shuyang Dai, Jiachang Liu, Zhe Gan, and Lawrence Carin. Club: A contrastive log-ratio upper bound of mutual information. In *International conference on machine learning*, pp. 1779–1788. PMLR, 2020.

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

Peter Gács, Janos Korner, et al. Common information is far less than mutual information. *Problems of Control and Information Theory*, 2:149–162, 1973.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind one embedding space to bind them all. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15180–15190. IEEE, 2023.

Virgil Griffith and Christof Koch. Quantifying synergistic mutual information. In *Guided self-organization: inception*, pp. 159–190. Springer, 2014.

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

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

F Maxwell Harper and Joseph A Konstan. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, 5(4):1–19, 2015.

Md Kamrul Hasan, Wasifur Rahman, AmirAli Bagher Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, and Mohammed Ehsan Hoque. Ur-funny: A multimodal language dataset for understanding humor. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 2046–2056, 2019.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM international conference on multimedia*, pp. 1122–1131, 2020.

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

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pp. 4904–4916. PMLR, 2021.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):160035, 2016.

Artemy Kolchinsky. A novel approach to the partial information decomposition. *Entropy*, 24(3):403, 2022.

Michelle A Lee, Yuke Zhu, Krishnan Srinivasan, Parth Shah, Silvio Savarese, Li Fei-Fei, Animesh Garg, and Jeannette Bohg. Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks. In *2019 International conference on robotics and automation (ICRA)*, pp. 8943–8950. IEEE, 2019.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021.

Yingzhen Li and Richard E. Turner. Gradient estimators for implicit models. In *International Conference on Learning Representations*, 2018.

Yong Li, Yuanzhi Wang, and Zhen Cui. Decoupled multimodal distilling for emotion recognition. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6631–6640. IEEE, 2023.

Paul Pu Liang, Yiwei Lyu, Xiang Fan, Zetian Wu, Yun Cheng, Jason Wu, Leslie Chen, Peter Wu, Michelle A Lee, Yuke Zhu, et al. Multibench: Multiscale benchmarks for multimodal representation learning. *Advances in neural information processing systems*, 2021(DB1):1, 2021.

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

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

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

Shengzhong Liu, Tomoyoshi Kimura, Dongxin Liu, Ruijie Wang, Jinyang Li, Suhas Diggavi, Mani Srivastava, and Tarek Abdelzaher. Focal: Contrastive learning for multimodal time-series sensing signals in factorized orthogonal latent space. *Advances in Neural Information Processing Systems*, 36:47309–47338, 2023.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pp. 4114–4124. PMLR, 2019.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Anh-Tuan Mai, Cam-Van Thi Nguyen, and Duc-Trong Le. Divide and refine: Enhancing multimodal representation and explainability for emotion recognition in conversation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2700–2709, 2026.

David McAllester and Karl Stratos. Formal limitations on the measurement of mutual information. In *International Conference on Artificial Intelligence and Statistics*, pp. 875–884. PMLR, 2020.

Salman Mohamadi, Gianfranco Doretto, and Donald A Adjeroh. Rethinking self-supervised learning within the framework of partial information decomposition. *arXiv preprint arXiv:2412.02121*, 2024.

Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. SLIP: Self-supervision meets language-image pre-training. In *European conference on computer vision*, pp. 529–544. Springer, 2022.

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, Andrew Y Ng, et al. Multimodal deep learning. In *Icml*, volume 11, pp. 689–696, 2011.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Chengxuan Qian, Shuo Xing, Li Li, Yue Zhao, and Zhengzhong Tu. DecAlign: Hierarchical cross-modal alignment for decoupled multimodal representation learning. In *International Conference on Learning Representations*, volume 2026, pp. 152802–152824, 2026.

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

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15617–15629. IEEE, 2022.

Qi Song, Tianxiang Gong, Shiqi Gao, Haoyi Zhou, and Jianxin Li. Quest: quadruple multimodal contrastive learning with constraints and self-penalization. *Advances in Neural Information Processing Systems*, 37:28889–28919, 2024.

Nitish Srivastava and Russ R Salakhutdinov. Multimodal learning with deep boltzmann machines. *Advances in neural information processing systems*, 25, 2012.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *European conference on computer vision*, pp. 776–794. Springer, 2020.

Manuel Tran, Yashin Dicente Cid, Amal Lahiani, Fabian Theis, Tingying Peng, and Eldad Klaiman. Training transitive and commutative multimodal transformers with LoReTTa. *Advances in Neural Information Processing Systems*, 36:42918–42931, 2023.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Chenyu Wang, Sharut Gupta, Xinyi Zhang, Sana Tonekaboni, Stefanie Jegelka, Tommi Jaakkola, and Caroline Uhler. An information criterion for controlled disentanglement of multimodal data. *International Conference on Learning Representations*, 2025.

Yi Wang, Conrad M Albrecht, Nassim Ait Ali Braham, Chenying Liu, Zhitong Xiong, and Xiao Xiang Zhu. Decoupling common and unique representations for multimodal self-supervised learning. In *European Conference on Computer Vision*, pp. 286–303. Springer, 2024.

Zehan Wang, Yang Zhao, Xize Cheng, Haifeng Huang, Jiageng Liu, Aoxiong Yin, Li Tang, Linjun Li, Yongqi Wang, Ziang Zhang, and Zhou Zhao. Connecting multi-modal contrastive representations. *Advances in Neural Information Processing Systems*, 36:22099–22114, 2023.

Liangjian Wen, Qun Dai, Jianzhuang Liu, Jiangtao Zheng, Yong Dai, Dongkai Wang, Zhao Kang, Jun Wang, Zenglin Xu, and Jiang Duan. InfMasking: Unleashing synergistic information by contrastive multimodal interactions. *Advances in Neural Information Processing Systems*, 38: 15529–15555, 2025.

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

Wenfang Yao, Kejing Yin, William K Cheung, Jia Liu, and Jing Qin. Drfuse: Learning disentangled representation for clinical multi-modal fusion with missing modality and modal inconsistency. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 16416–16424, 2024.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*, 2016.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International conference on machine learning*, pp. 12310–12320. PMLR, 2021.

A Proof of Proposition 1

Throughout, Z_R and Z_S are (possibly randomized) functions of $\mathbf{X}$ whose randomness is independent of Y given $\mathbf{X}$, so that $Y - \mathbf{X} - (Z_R, Z_S)$ is a Markov chain.

Part (a). Griffith et al. (2014) define the common-randomness redundancy $I_\wedge(X_1, \ldots, X_M \to Y) = \max_Q\{I(Q;Y) : Q \text{ is a deterministic function of } X_i\ \forall i\}$; every such Q is a function of the Gács–Körner common random variable (Gács et al., 1973), which therefore attains the maximum. If Z_R is a deterministic function of every X_i, it is feasible for this maximization, hence $I(Z_R;Y) \le I_\wedge$. For any feasible Q, Q is a function of X_i, so $I(Q;Y) \le I(X_i;Y)$ by data processing; taking the minimum over i gives $I_\wedge \le \min_i I(X_i;Y)$. □

Part (b). Fix i. By the chain rule applied to $I(Z_S; Y, X_i)$ in two orders,

$$I(Z_S;X_i) + I(Z_S;Y \mid X_i) = I(Z_S;Y) + I(Z_S;X_i \mid Y).$$

With $I(Z_S;X_i) = 0$, this gives $I(Z_S;Y \mid X_i) - I(Z_S;Y) = I(Z_S;X_i \mid Y) \ge 0$, i.e., the co-information $I(Z_S;X_i;Y) = I(Z_S;Y) - I(Z_S;Y \mid X_i)$ is nonpositive, which is the standard signature of net synergy. Next, since $Y - \mathbf{X} - Z_S$ is a Markov chain and X_i is a component of $\mathbf{X}$, $I(Z_S;Y \mid X_i) \le I(\mathbf{X}, Z_S;Y \mid X_i) = I(\mathbf{X};Y \mid X_i) + I(Z_S;Y \mid \mathbf{X}) = I(\mathbf{X};Y \mid X_i)$, and because X_i is a component of $\mathbf{X}$, $I(\mathbf{X};Y \mid X_i) = I(\mathbf{X};Y) - I(X_i;Y)$ by the chain rule. Combining, $I(Z_S;Y) \le I(\mathbf{X};Y) - I(X_i;Y)$ for every i. For general M, taking the minimum over i gives $I(Z_S;Y) \le I(\mathbf{X};Y) - \max_i I(X_i;Y) = S_\cup + [I_\cup - \max_i I(X_i;Y)]$. Every $Q \in \Delta_P$ preserves $P(X_i,Y)$, so $I_Q(\mathbf{X};Y) \ge I_Q(X_i;Y) = I(X_i;Y)$ for all i; hence $I_\cup \ge \max_i I(X_i;Y)$ and the bracket is nonnegative. For $M = 2$ under BROJA, $I_\cup = R + U_1 + U_2$ and the bracket equals $\min(U_1, U_2)$. For $M = 2$, Eq. (1) gives $I(X_1, X_2;Y) - I(X_1;Y) = U_2 + S$ and $I(X_1, X_2;Y) - I(X_2;Y) = U_1 + S$; taking the minimum yields $I(Z_S;Y) \le S + \min(U_1, U_2)$. □

Part (c). Let X_1, X_2 be independent uniform bits and $Y = (X_1, X_2)$. Then $I(X_1, X_2;Y) = 2$ and $I(X_i;Y) = 1$. We compute BROJA directly from Eq. (2). Write $Y = (Y_1, Y_2)$. Any $Q \in \Delta_P$ preserves the pairwise marginals $P(X_1,Y)$ and $P(X_2,Y)$, under which $X_1 = Y_1$ and $X_2 = Y_2$ almost surely; hence under every $Q \in \Delta_P$, $\mathbf{X}$ and Y are in bijection and $I_Q(X_1, X_2;Y) = H(Y) = 2$. Therefore $S = 2 - 2 = 0$. The BROJA unique information $U_1 = \min_{Q \in \Delta_P} I_Q(X_1;Y \mid X_2) = H(Y_1 \mid Y_2) = 1$, and likewise $U_2 = 1$, so $R = I(X_1;Y) - U_1 = 0$. $Z_S = X_1 \oplus X_2$ is independent of X_1 and of X_2 but $I(Z_S;Y) = H(Z_S) = 1 = S + \min(U_1, U_2)$. □

Remark 1 (Scope). Proposition 1 characterizes the idealized constraints, not the trained networks. Part (a) uses exact recoverability, which the Gács–Körner common part makes restrictive for noisy continuous data; the consistency terms $\mathcal{C}_{R,i}$ relax it to contrastive predictability. Part (b) excludes information about raw singletons X_i, whereas $\mathcal{L}_{S\perp}$ repels learned references Z_i with a cosine hinge, which is weaker ($I(Z_S;Z_i) = 0$ does not imply $I(Z_S;X_i) = 0$). Neither part provides a lower bound on how much synergy is preserved. These gaps are why we evaluate the learned factors empirically.

B Experimental Details

Protocol. Experiments run on a single NVIDIA RTX 4090 (24 GB). For each dataset we run seeds 42–46 and report the mean ± sample standard deviation ($\mathrm{ddof} = 1$) over the five runs; the standard deviation includes the variability of both pretraining and probe fitting. We follow the benchmark preprocessing, splits, and linear-probing protocol of CoMM (Dufumier et al., 2025); evaluation freezes the pretrained encoder and trains a linear classifier or regressor on the extracted features.

For the Trifeature factor-probing matrix, all four targets use the same pretrained checkpoint and pretraining dataset for each seed; only the downstream target label and probe change. The validation and probing convention of this matrix follows CoMM. The two-modality V&T CP results for Cross (85.82%) and Cross+Self (86.74%) in Table 5 are our reproductions.

Augmentations. We use the same modality-specific augmentations as CoMM (Dufumier et al., 2025), and the two views are generated by independently sampling transformations from the same input. Raw images (Trifeature, MM-IMDb, Vision&Touch) use the default SimCLR augmentations (Chen et al., 2020) (random resized crop, color jitter, random grayscale, Gaussian blur, and

horizontal flip); tabular data (MIMIC, Vision&Touch) receive Gaussian noise on each component; time-series data (features extracted from videos in MOSI, UR-FUNNY, and MUSTARD, health recordings in MIMIC, and force-torque readings in Vision&Touch) use a random composition of Gaussian noise and random dropping of 0–80% of the sequence; and raw text (MM-IMDb) has 15% of its tokens randomly masked.

Optimization and architecture. URSyn is implemented in PyTorch (Paszke et al., 2019) and PyTorch Lightning with AdamW (Loshchilov & Hutter, 2017), maximum epoch budgets, and early stopping on a validation metric (accuracy, or MSE for V&T EE; patience 25 for two-modality MultiBench tasks and 35 for three-modality tasks and MM-IMDb; on Trifeature, the validation subset is held out from the training pairs). Labels enter only this checkpoint selection, never the pretraining objective, and the same rule is applied to our CoMM reproduction; MM-IMDb uses cosine learning-rate decay with linear warmup (Chen et al., 2020). The backbone follows CoMM (Dufumier et al., 2025): modality-specific encoders and latent converters produce token sequences, which are concatenated and passed to a lightweight Transformer fusion module (Vaswani et al., 2017) with a learnable `[CLS]` token. URSyn adds one unique branch per modality, a shared redundant extractor applied to each singleton view, a synergy branch, and a shared projection head. During training we enumerate all single-modality masks and the full-modality mask for both augmented views. At inference no augmentation is applied, and the factors are computed with the same masks as in training: Z_{U_i} and Z_{R_i} from the singleton view of modality i, and Z_S from the full input (Algorithm 1).

The contrastive temperature is $\tau = 0.1$. The shared projection head is a three-layer MLP, $d \to 512 \to 512 \to 256$, with SyncBatchNorm and ReLU after each of the first two layers. Its output dimension is 256 for every dataset; this differs from the pre-projection factor dimension d, which is 40 for MOSI, UR-FUNNY (two or three modalities), and MUSTARD; 128 for V&T EE; 512 for Trifeature, MIMIC, and V&T CP (two or three modalities); and 768 for MM-IMDb. Downstream probes use pre-projection factors, with input dimension $(M+2)d$ for the full concatenation and d for a single factor.

Normalization and anchoring. All projected reference and branch vectors are ℓ_2-normalized before gathering the batch and computing the losses. Thus the cross-view anchor is a batch mean of normalized inner products, $\langle A, B\rangle = N^{-1}\sum_{n=1}^{N} a_n^\top b_n$, with no additional temperature or dimension scaling. The two directed anchor means are summed, not averaged with each other.

Entropy-gradient surrogate. Both Eqs. (5) and (6) use the same entropy-gradient surrogate on normalized projected features. It is based on the Stein gradient estimator (Li & Turner, 2018), which estimates the score $\nabla_x \log q(x)$ at the sample points; since $\nabla_\theta H = -\mathbb{E}[\nabla_x \log q(x)^\top \partial x/\partial\theta]$, plugging a detached score estimate into a linear term yields a training signal whose gradient approximates that of the entropy. For a batch $X = (x_1, \dots, x_N)^\top$ we use

$$
\begin{aligned}
b &= \max\Big(0.1, \mathrm{kthlargest}_{\lfloor N^2/2\rfloor}\{1 - x_i^\top x_j\}_{i,j=1}^{N}\Big), \qquad \mathbf{K}_{ij} = \exp(x_i^\top x_j/b),\\
G_i &= b^{-1}\Big(\sum_j \mathbf{K}_{ij}\Big)x_i, \qquad \Psi = -(\mathbf{K} + \eta I_N)^{-1} G,\\
\widehat{H}(X) &= -\frac{1}{N}\sum_i \mathrm{stopgrad}(\Psi_i)^\top x_i.
\end{aligned}
\tag{9}
$$

Here Ψ is the Stein gradient estimate of the score with kernel $k(x,y) = \exp(x^\top y/b)$, whose values on unit-norm features equal those of a Gaussian kernel up to a constant factor (the estimator is used as a heuristic on the sphere); $G_i = \sum_j \nabla_{x_j} k(x_i, x_j)$ for this kernel, row i of Ψ estimates the score at x_i, and $\eta = 0.01$ regularizes the kernel matrix. The bandwidth b is the $\lfloor N^2/2\rfloor$-th largest value of $1 - x_i^\top x_j$ over all i, j, including the diagonal, lower-bounded by 0.1 and independent of the InfoNCE temperature. The kernel matrix $\mathbf{K}$ and Ψ are detached from differentiation. $\widehat{H}$ therefore supplies an entropy gradient; its value is not a calibrated entropy in bits or nats, and no entropy values are reported.

Linear probing. We follow the linear-probing protocol of CoMM. For single-label classification, an affine classifier is fitted with full-batch L-BFGS on cross-entropy plus the penalty $\|W\|_F/(2C)$

on the weights (the CoMM convention), without class reweighting; the same probe is applied to all compared representations. The coefficient C is selected on validation features (or, when no validation split exists, on a stratified hold-out of the training features), and the final classifier is refitted on training and validation features together. MM-IMDb uses per-label ℓ_2-regularized logistic regression, and V&T EE uses ridge regression with a validation-selected penalty.

Hyperparameters. Table 9 lists the three regularization coefficients of the separation terms, λ_h (`entropy_weight`), λ_a (`align_weight`), and λ_p (`lambda_repel`), as used in the training configurations, without rescaling, together with the AdamW learning rate of each dataset. The cosine penalties are averaged over both augmented views. With these symbols, the U/R cosine term has coefficient $\lambda_p/(2M^2)$, while the synergy–singleton term has coefficient $\lambda_p/(2M)$; after the outer weights these become $\lambda_{UR}\lambda_p/(2M^2)$ and $\lambda_{S\perp}\lambda_p/(2M)$, respectively. The remaining coefficients are the cosine margin $\epsilon \in [0.19, 0.26]$ (`epsilon`) and four weights that balance the R, UR, S, and $S \perp$ groups against $\mathcal{L}_U$: λ_R (`lambda_R`), λ_{UR} (`lambda_UR_dis`), λ_S (`lambda_S`), and $\lambda_{S\perp}$ (`lambda_S_dis`). Appendix C.6 reports sweeps of all eight loss coefficients on UR-FUNNY and MOSI; they do not establish sensitivity on the remaining datasets. In practice, $\lambda_a = 1$ and $\tau_r = 0.5$ are fixed across all datasets. Coefficients are selected by validation accuracy, following the data splits and validation protocol of CoMM; the test split is used only for final reporting. CoMM and InfMasking use their public configurations, which we adjusted only where our reproduction deviated substantially from the reported results.

Table 9: Regularization coefficients and AdamW learning rates of URSyn. $\tau_r = 0.5$ for all datasets (downstream only).

Dataset	λ_h `entropy_weight`	λ_a `align_weight`	λ_p `lambda_repel`	Learning rate (AdamW)
Trifeature	0.03	1.0	10	3×10^{-4}
MIMIC	0.02	1.0	7	10^{-3}
MOSI	0.1	1.0	10	10^{-3}
UR-FUNNY (2 mod.)	0.1	1.0	20	10^{-3}
MUSTARD	0.05	1.0	20	10^{-3}
V&T CP (2 mod.)	0.001	1.0	20	10^{-4}
V&T EE	0.01	1.0	20	10^{-4}
MM-IMDb	0.01	1.0	2	10^{-3}
UR-FUNNY (3 mod.)	0.002	1.0	15	10^{-3}
V&T CP (3 mod.)	0.003	1.0	2	10^{-3}

Compute. Wall-clock times are for one seed under the maximum 100-epoch budget; early stopping can reduce them, and they vary with GPU load. One Trifeature seed takes about 14 GPU hours. One seed takes about 0.8–1.2 GPU hours on MIMIC, 4–6 on MM-IMDb, and 5–10 on Vision&Touch, depending on task and number of modalities. Table 10 compares the cost of URSyn and CoMM on MUSTARD, MOSI, and two-modality UR-FUNNY under the same backbone and hardware. URSyn needs 4.50–4.90× the FLOPs per training batch, 1.98–2.08× the forward time, and 1.89–2.05× the 100-epoch training time of CoMM. The extra cost comes from the typed branches and the losses computed on them for every singleton and full view; downstream probes are applied to frozen features in both cases.

Table 10: Training cost of URSyn and CoMM (one seed, single RTX 4090). FLOPs are per training batch; forward time is the mean per-batch forward pass; training time is for 100 epochs.

Dataset	Model	FLOPs/batch (G)	Forward (ms)	Training (min)
MUSTARD	CoMM	2.414	11.67	14.85
	URSyn	11.817	23.05	28.14
MOSI	CoMM	2.824	11.99	16.75
	URSyn	12.716	24.95	32.06
UR-FUNNY	CoMM	1.164	11.41	36.73
	URSyn	5.390	22.90	75.48

C Additional Analyses

C.1 Objective ablation on standard benchmarks

Table 11 evaluates the five objective variants on MOSI and two-modality UR-FUNNY. Each entry is the mean and standard deviation over five runs under the same downstream protocol as the main results. The full model and all variants were trained in the same set of runs with the same code and seeds, and all are evaluated with Relative-Diversity Routing. These experiments measure overall task quality; they do not replace the factor-level Trifeature probing analysis. On UR-FUNNY most differences from the full model are within one standard deviation. In the w/o-entropy variant, $\lambda_h = 0$ removes the entropy terms of both $\mathcal{L}_{UR}$ and $\mathcal{L}_{S\perp}$; all other coefficients are unchanged and not retuned.

Table 11: Objective ablation on standard multimodal benchmarks. Accuracy (%) is reported as mean $\pm$ standard deviation over five runs.

Variant	MOSI	UR-FUNNY
Full	67.06±1.20	64.93±0.62
w/o $\mathcal{L}_{S\perp}$	64.96±2.01	64.58±0.84
w/o ρ in $\mathcal{L}_{S\perp}$	66.60±1.55	64.10±1.00
w/o $\mathcal{L}_{UR}$	62.95±3.14	64.14±0.58
w/o $\mathcal{P}_R$	66.70±1.40	64.30±0.90
w/o entropy	61.64±2.81	63.22±0.58

C.2 Component-level comparison with CoMM

Table 12 lists what URSyn inherits from CoMM and what it changes. URSyn does not add its terms on top of the CoMM loss: the CoMM agreement terms are replaced by factor-specific objectives, and new separation terms and a factor-level output interface are added.

Table 12: Components of CoMM and URSyn.

Component	CoMM	URSyn
Encoders, Transformer fusion, masks, augmentations	singleton/full masks, two augmented views	inherited unchanged
Full-view preservation	full–full InfoNCE on Z	$\mathcal{L}_S$ on the dedicated factor Z_S, anchored to Z
Singleton information	singleton–full InfoNCE into Z	$\mathcal{L}_U$ ($Z_i \to Z_{U_i}$) and $\mathcal{L}_R$ (every $Z_i \to Z_{R_j}$, plus agreement among Z_{R_j})
Separation	none	$\mathcal{L}_{UR}$ (U vs. R) and $\mathcal{L}_{S\perp}$ (S vs. singletons), with entropy terms against collapse
Downstream output	one fused Z	$Z_{U_1}, \ldots, Z_{U_M}, Z_R, Z_S$, probed, removed, or routed individually

C.3 CoMM: reproduced and reported results

All CoMM results in the main text were obtained by running its public code following the CoMM evaluation protocol, which URSyn also follows, so that both methods are evaluated under the same protocol. Table 13 lists them next to the numbers reported in the CoMM paper (Dufumier et al., 2025). The reproduced numbers are lower in several settings; relative to the reported numbers, URSyn is better on MIMIC, UR-FUNNY, MUSTARD, the classification average, two-modality V&T contact prediction, three-modality UR-FUNNY, and both MM-IMDb metrics, and worse on V&T EE (higher MSE), MOSI, and three-modality V&T contact prediction. On Trifeature, the CoMM paper reports 99.9/87.8/71.9% (redundancy/uniqueness/synergy, AlexNet backbone) in its appendix, close to our reproduction (99.93/87.23/70.41%, Table 3); the Cross and Cross+Self rows marked † in Table 3 are taken from that appendix table, and the FactorCL row from the main text of the CoMM paper.

Table 13: CoMM results reproduced with its public code following the CoMM evaluation protocol vs. reported in the CoMM paper. MSE is $\times 10^{-4}$.

Setting (metric)	CoMM (reproduced)	CoMM (reported)	URSyn
V&T EE (MSE ↓)	7.65±1.83	4.55±0.30	5.24±1.26
MIMIC (acc.)	66.47±0.40	66.40±0.32	68.00±0.20
MOSI (acc.)	64.46±3.05	67.50±1.30	67.06±1.20
UR-FUNNY, 2 mod. (acc.)	62.82±1.10	63.10±0.65	64.93±0.62
MUSTARD (acc.)	63.35±6.49	63.90±3.01	66.99±2.48
Classification average	64.28	65.22	66.75
V&T CP, 2 mod. (acc.)	86.14±1.28	88.10	88.76±0.74
V&T CP, 3 mod. (acc.)	94.10±0.10	94.20	94.10±0.05
UR-FUNNY, 3 mod. (acc.)	64.76±1.35	64.60	65.58±0.88
MM-IMDb, CLIP (weighted-F1)	60.72±0.37	61.48±0.18	62.62±0.67
MM-IMDb, CLIP (macro-F1)	53.48±0.38	54.63±0.22	55.76±0.67

C.4 Comparison with InfMasking on real-world benchmarks

Table 14 compares URSyn with InfMasking on all real-world settings. InfMasking numbers are taken from its paper (five runs); URSyn numbers are those of Tables 4–6. Because the InfMasking numbers were not rerun in our pipeline and per-seed results are unavailable, small differences should not be over-interpreted and no significance test is performed.

Table 14: URSyn vs. InfMasking on real-world benchmarks. ‡: reported by InfMasking. MSE is $\times 10^{-4}$.

Setting (metric)	InfMasking‡	URSyn
V&T EE (MSE ↓)	**4.23±0.37**	5.24±1.26
MIMIC (acc.)	**68.10±0.42**	68.00±0.20
MOSI (acc.)	**69.00±1.20**	67.06±1.20
UR-FUNNY, 2 mod. (acc.)	64.30±0.90	**64.93±0.62**
MUSTARD (acc.)	66.80±2.50	**66.99±2.48**
Classification average	**67.05**	66.75
V&T CP, 2 mod. (acc.)	88.50±0.33	**88.76±0.74**
V&T CP, 3 mod. (acc.)	**94.10±0.09**	**94.10±0.05**
UR-FUNNY, 3 mod. (acc.)	**65.60±1.15**	65.58±0.88
MM-IMDb (weighted-F1)	62.60	**62.62±0.67**
MM-IMDb (macro-F1)	**55.93**	55.76±0.67

C.5 Routing: standard deviations, temperature, and nonlinear probes

Table 15: Routing on frozen features with standard deviations (five seeds). CoMM views: singleton and full views exported from unchanged CoMM checkpoints.

Features	MOSI	UR-FUNNY	MUSTARD	UR-FUNNY$_3$	Avg.
CoMM views, Raw	65.03±0.92	64.04±1.43	64.60±2.05	64.91±1.21	64.65
CoMM views, +div.	64.62±0.96	63.64±1.38	64.46±3.04	64.98±1.19	64.43
URSyn, Raw	65.86±1.62	64.42±1.23	65.44±1.64	65.46±0.71	65.30
URSyn, +div.	67.06±1.20	64.93±0.62	66.99±2.48	65.58±0.88	66.14

Temperature. The value $\tau_r = 0.5$ used in all main results was fixed before this sweep. For each τ_r we keep pretrained checkpoints and splits fixed, recompute routed representations, and refit the same linear probe; the sweep therefore affects neither pretraining nor encoder parameters. The temperature changes the sharpness of the weights but not the ranking of diversity scores within a sample; results are in Table 16.

Table 16: Sensitivity of URSyn to the downstream routing temperature τ_r.

τ_r	MOSI	UR-FUNNY	MUSTARD	UR-FUNNY$_3$	Avg.
0.10	66.25±2.65	64.16±0.96	66.12±2.87	65.49±1.36	65.51
0.25	66.02±1.04	64.24±1.28	65.97±2.18	65.48±0.77	65.43
0.50	67.06±1.20	64.93±0.62	66.99±2.48	65.58±0.88	66.14
0.75	66.77±1.68	64.07±1.40	67.53±3.41	65.56±1.23	65.98
1.00	65.15±2.20	63.89±1.48	65.21±2.47	65.41±1.61	64.92

Nonlinear probe. With identical frozen checkpoints, a two-layer MLP probe, the same budget and validation, and five seeds, routing improves all four means (Table 17). Absolute MLP accuracy is not higher than linear probing on the three datasets shared with Table 15, so we do not present the MLP as a stronger readout; the purpose of this control is only to test whether the routing gain disappears once the readout is nonlinear.

Table 17: Routing with a two-layer MLP probe on frozen URSyn features.

Features	MOSI	UR-FUNNY	MUSTARD	MIMIC	Avg.
Raw Concat	63.89±2.71	63.60±1.08	64.16±1.68	67.31±0.60	64.74
Relative-Diversity Routing	64.92±2.32	63.72±0.77	65.55±2.94	67.75±0.51	65.48

C.6 Hyperparameter sensitivity

We report sensitivity results on two-modality UR-FUNNY and MOSI. Figure 2 shows the UR-FUNNY λ_h and λ_p sweeps (means only). Table 18 reports the 22 UR-FUNNY settings for the other six coefficients, each summarized over five runs. Table 19 reports the corresponding MOSI sweeps, in which 16 of the 27 non-default settings exceed CoMM (64.46%) and the means range from 62.44% to 67.04%; MOSI, the smaller of the two datasets, is thus noticeably more sensitive to the coefficients than UR-FUNNY. The average over all 27 non-default MOSI settings (64.75%) is 0.29 points above CoMM, well within one standard deviation, and no setting falls more than one CoMM standard deviation (3.05) below CoMM; this spread is comparable to the high seed variance of this small dataset. The reference point of each dataset is its main-result configuration, marked as the default row. For both datasets, the default configuration was rerun with the same code and seeds (42–46) in an unchanged software and hardware environment; the reruns reproduce the main results of Table 4 exactly (64.93±0.62% and 67.06±1.20%). The confirmation run is listed as the default row and is not counted as an additional sweep point. Of the 22 UR-FUNNY settings in Table 18, 21 have means above CoMM (62.82%); the exception is $\lambda_R = 2$ (62.76±1.18%), and the means range from 62.76% to 64.61%. Including the λ_h and λ_p sweeps, 27 of 30 non-default settings exceed CoMM. These comparisons describe observed means, not statistical significance or robustness on other datasets.

The λ_p sweep varies the cosine penalties in both Eq. (5) and Eq. (6), as in the implementation.

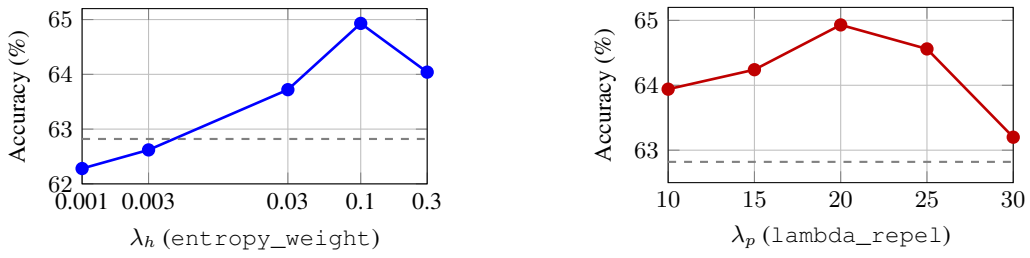

Figure 2: One-at-a-time sensitivity to λ_h and λ_p on UR-FUNNY (two modalities; means over five runs). Dashed line: CoMM (62.82%). The default value was rerun with the same code and seeds (64.93±0.62%).

Table 18: Additional UR-FUNNY sensitivity results (accuracy in %; five runs per measured setting). Standard deviations are in percentage points. Rows marked "default" are the main-result configuration rerun with the same code and seeds (64.93±0.62%) and are not additional sweep runs.

Coefficient	Value	Mean ± std.
ϵ	0.1	63.46 ± 0.64
ϵ	0.15	63.76 ± 1.07
ϵ	0.23 (default)	64.93 ± 0.62
ϵ	0.25	63.85 ± 1.06
ϵ	0.3	63.36 ± 0.82
λ_R	0.5	64.00 ± 0.72
λ_R	0.7 (default)	64.93 ± 0.62
λ_R	1.1	63.83 ± 0.67
λ_R	1.5	63.56 ± 0.76
λ_R	2	62.76 ± 1.18
λ_{UR}	5	63.14 ± 1.24
λ_{UR}	7 (default)	64.93 ± 0.62
λ_{UR}	10	63.75 ± 1.35
λ_{UR}	15	63.65 ± 1.31
λ_{UR}	25	63.20 ± 0.78
λ_S	0.7	64.03 ± 1.17
λ_S	1.6	63.29 ± 0.38
λ_S	2.2	63.57 ± 0.79
λ_S	4 (default)	64.93 ± 0.62
λ_S	5	63.66 ± 0.61
$\lambda_{S\perp}$	1	63.59 ± 0.12
$\lambda_{S\perp}$	2	63.51 ± 1.28
$\lambda_{S\perp}$	3	64.00 ± 0.93
$\lambda_{S\perp}$	5 (default)	64.93 ± 0.62
$\lambda_{S\perp}$	7	63.76 ± 0.48
λ_a	0.5	64.61 ± 0.57
λ_a	1 (default)	64.93 ± 0.62
λ_a	2	63.52 ± 0.96

Table 19: MOSI sensitivity results (accuracy in %; five runs per measured setting). Rows marked "default" are the main-result configuration (67.06±1.20%), rerun with the same code and seeds. CoMM on MOSI: 64.46%.

Coefficient	Value	Mean ± std.
ϵ	0.18	62.92 ± 1.36
ϵ	0.2	64.08 ± 1.62
ϵ	0.22	65.00 ± 1.90
ϵ	0.25 (default)	67.06 ± 1.20
ϵ	0.26	64.20 ± 2.44
λ_h	0.01	65.01 ± 1.45
λ_h	0.03	64.76 ± 2.31
λ_h	0.1 (default)	67.06 ± 1.20
λ_h	0.3	66.96 ± 1.43
λ_a	0.5	64.50 ± 2.65
λ_a	1 (default)	67.06 ± 1.20
λ_a	2	66.08 ± 1.19
λ_p	5	63.52 ± 2.55
λ_p	10 (default)	67.06 ± 1.20
λ_p	20	67.04 ± 3.02
λ_R	0.7	62.44 ± 1.62
λ_R	0.9	62.92 ± 1.87
λ_R	1.1 (default)	67.06 ± 1.20
λ_R	1.3	65.01 ± 1.24
λ_R	1.5	66.13 ± 2.13
λ_{UR}	1	65.99 ± 2.71
λ_{UR}	3	65.20 ± 1.87
λ_{UR}	5 (default)	67.06 ± 1.20
λ_{UR}	7	66.03 ± 2.53
λ_{UR}	10	64.12 ± 3.33
λ_S	0.5	63.82 ± 1.08
λ_S	0.7 (default)	67.06 ± 1.20
λ_S	1	64.20 ± 3.55
λ_S	1.3	64.97 ± 1.37
λ_S	1.5	65.24 ± 0.69
$\lambda_{S\perp}$	1	65.98 ± 2.53
$\lambda_{S\perp}$	2	62.97 ± 1.90
$\lambda_{S\perp}$	3 (default)	67.06 ± 1.20
$\lambda_{S\perp}$	5	65.00 ± 3.84
$\lambda_{S\perp}$	7	64.17 ± 1.32

D Dataset Details

All MultiBench datasets follow MultiBench's standardized preprocessing (Liang et al., 2021).

Trifeature. Trifeature (Hermann & Lampinen, 2020) is a synthetic visual benchmark whose images combine three independent attributes, *shape*, *color*, and *texture*, each with ten categories (1,000 combinations). Following prior multimodal interaction studies (Dufumier et al., 2025), 800 combinations are used for training and 200 for testing, with three instances per training combination under random transformations of shape and texture. Shapes are rendered inside a 128×128 box, rotated by an angle in $[-45°, 45°]$, and placed within a 224×224 canvas. Pairs of images form two-modality samples, giving 10,000 training and 4,096 test pairs. Shape is shared by both images (redundancy), texture is specific to each image (uniqueness), and the binary synergy target depends jointly on texture and color across the two images.

MIMIC. MIMIC (Johnson et al., 2016) contains 53,423 hospital admissions of 38,597 patients (2001–2012). We use 24-hour hourly clinical measurements (12-dimensional time series) and static patient attributes (5-dimensional tabular vectors). The task is binary prediction of ICD-9 group 7 (codes 460–519, respiratory diseases).

MOSI. CMU-MOSI (Zadeh et al., 2016) contains 2,199 opinion video clips with video, audio, and text. Sentiment labels in $[-3, 3]$ are binarized. We use text and video.

UR-FUNNY. UR-FUNNY (Hasan et al., 2019) contains 16,514 segments from 1,866 TED talks with video, audio, and text, for binary humor detection. The two-modality setting uses text and video; the three-modality setting adds audio.

MUSTARD. MUSTARD (Castro et al., 2019) contains 690 balanced sarcastic/non-sarcastic utterances from TV shows with video, audio, and text. We use text and video.

Vision&Touch. Vision&Touch (Lee et al., 2019) contains 150 robotic manipulation trajectories of 1,000 steps with RGB images, depth, force, and end-effector pose/velocity. Tasks are next-step contact prediction (CP, binary) and end-effector position regression (EE, MSE). The two-modality configuration uses vision and proprioception; the three-modality configuration uses vision, force, and proprioception.

MM-IMDb. MM-IMDb (Arevalo et al., 2017), derived from MovieLens 20M (Harper & Konstan, 2015), contains 25,959 movies with posters, plot summaries, and metadata, labeled with one or more of 23 genres. We use posters and plots and train on raw inputs with CLIP encoders rather than MultiBench pre-extracted features. Both CLIP encoders are frozen for CoMM and URSyn alike; only the fusion module and the method-specific heads are trained.

E Pseudocode

For two batches $A = \{a_k\}_{k=1}^N$ and $B = \{b_k\}_{k=1}^N$, the symmetric InfoNCE with temperature τ is

$$\mathcal{L}_{\mathrm{NCE}}(A, B) = -\frac{1}{2N} \sum_{k=1}^{N} \left[\log \frac{\exp(\mathrm{sim}(a_k, b_k)/\tau)}{\sum_{l=1}^{N} \exp(\mathrm{sim}(a_k, b_l)/\tau) + \sum_{l \neq k} \exp(\mathrm{sim}(a_k, a_l)/\tau)} + \log \frac{\exp(\mathrm{sim}(b_k, a_k)/\tau)}{\sum_{l=1}^{N} \exp(\mathrm{sim}(b_k, a_l)/\tau) + \sum_{l \neq k} \exp(\mathrm{sim}(b_k, b_l)/\tau)} \right].$$

Algorithm 1 uses one unique branch per modality, one redundant extractor evaluated on each singleton view, and one synergy branch on the full view. For brevity, it writes z_i, U_i, R_i, and S for the projected features Z_i, Z_{U_i}, Z_{R_i}, and Z_S; here U_i, R_i, and S denote factors, not PID quantities.

Algorithm 1. **URSyn pretraining**

Require: Dataset $\mathcal{D}$ with M modalities, label-preserving transformations $\mathcal{T}^\star$, batch size N, singleton/full masks $\mathcal{M} = \{e_1, \ldots, e_M, \mathbf{1}\}$, fusion encoder E, heads h, temperature τ, margin ϵ, coefficients $\lambda_h, \lambda_a, \lambda_p, \lambda_R, \lambda_{UR}, \lambda_S, \lambda_{S\perp}$.

for *sampled mini-batch* $\{x_k\}_{k=1}^{N}$ **do**

draw two augmented views $X', X'' \sim \mathcal{T}^\star$

$(z'_1, \ldots, z'_M, Z', U'_1, \ldots, U'_M, R'_1, \ldots, R'_M, S') \leftarrow h(E(X', \mathcal{M}))$

$(z''_1, \ldots, z''_M, Z'', U''_1, \ldots, U''_M, R''_1, \ldots, R''_M, S'') \leftarrow h(E(X'', \mathcal{M}))$

ℓ_2-normalize and all-gather all projected reference and branch features

$\mathcal{L}_U \leftarrow \frac{1}{2M} \sum_i [\mathcal{L}_{\mathrm{NCE}}(z'_i, U''_i) + \mathcal{L}_{\mathrm{NCE}}(z''_i, U'_i)]$

$\mathcal{C}_R \leftarrow \{(2M)^{-1} \sum_j [\mathcal{L}_{\mathrm{NCE}}(z'_i, R''_j) + \mathcal{L}_{\mathrm{NCE}}(z''_i, R'_j)]\}_{i=1}^{M}$

$\mathcal{P}_R \leftarrow \{[\mathcal{L}_{\mathrm{NCE}}(R'_i, R'_j) + \mathcal{L}_{\mathrm{NCE}}(R''_i, R''_j)]/2\}_{i<j}$

$\mathcal{L}_R \leftarrow \mathrm{mean}(\mathcal{C}_R \cup \mathcal{P}_R)$

$\mathcal{L}_{UR} \leftarrow \frac{1}{M} \sum_i \{-\lambda_h [\widehat{H}(U'_i) + \widehat{H}(U''_i)] + \frac{\lambda_p}{2M} \sum_j [\rho(U'_i, R'_j) + \rho(U''_i, R''_j)]\}$

$\mathcal{L}_S \leftarrow \mathcal{L}_{\mathrm{NCE}}(S', S'')$

$\mathcal{L}_{S\perp} \leftarrow -\lambda_h [\widehat{H}(S') + \widehat{H}(S'')] - \lambda_a [\langle S', Z'' \rangle + \langle S'', Z' \rangle] + \frac{\lambda_p}{2M} \sum_i [\rho(S', z'_i) + \rho(S'', z''_i)]$

$\mathcal{L}_{URSyn} \leftarrow \mathcal{L}_U + \lambda_R \mathcal{L}_R + \lambda_{UR} \mathcal{L}_{UR} + \lambda_S \mathcal{L}_S + \lambda_{S\perp} \mathcal{L}_{S\perp}$

update E and h to minimize $\mathcal{L}_{URSyn}$

end

return: E, h