RePercENT: Scaling Disentangled Representation Learning Beyond Two Modalities

Vasiliki Rizou
EPFL
vasiliki.rizou@epfl.ch

Pascal Frossard
EPFL
pascal.frossard@epfl.ch

Dorina Thanou
EPFL
dorina.thanou@epfl.ch

Code

Abstract

To leverage the full potential of multimodal data, we need representations that go beyond the state-of-the-art alignment and fusion approaches and exploit all cross-modal interactions without sacrificing modality-specific information. Learning disentangled representations is a principled way to identify the underlying shared and unique factors that are hidden in observational data. However, while multi-modal disentanglement is a compelling paradigm, existing methods are largely confined to the two-modality regime due to its inherent scalability bottleneck. To address this, we propose RePercENT, a self-supervised framework designed to surpass these limitations and *unlocks scalable pairwise disentanglement* beyond two modalities. Through a multimodal 'plug-and-play' architecture, our approach operates directly on pre-extracted embeddings, eliminating the need for extensive joint pre-training while making no assumptions regarding the underlying modalities or foundation model backbones. Moreover, we introduce a joint optimization objective for simultaneously deriving the shared and unique components, and provide formal theoretical guarantees that characterize the optimality of our solution. Spanning modalities from images and videos to clinical reports and omics, and tasks from sentiment analysis to tumor classification, RePercENT successfully recovers disentangled components while maintaining competitive performance and significantly reducing computational complexity.

1 Introduction

A central aim of multimodal machine learning is to endow models with the ability to synthesize and reason over information coming from multiple sources, e.g. images, audio, text [Baltrušaitis et al., 2019, Krones et al., 2024]. Landmark multimodal foundation models, such as CLIP [Radford et al., 2021], ALIGN [Jia et al., 2021], Flamingo [Alayrac et al., 2022], have shown that integrating different modalities gives rise to richer, more semantic and broadly applicable representations, than those derived from a single source alone. Notably, most previous work focuses on cross-modal alignment, exploiting the critical assumption of *multi-view redundancy* [Liang et al., 2023], which suggests that *all* task-relevant information is shared across modalities. However, as demonstrated by Tjandrasuwita et al. [2025], this assumption critically depends on the degree of similarity between the modalities and the balance between redundant and unique information they provide for the task under consideration (see Figure 1, left).

Multimodal fusion provides a better solution for capturing more complex cross-modal interactions, by combining heterogeneous modalities into a joint representation. While effective, this representation often relies on the most "dominant" modality [Kontras et al., 2025], under-representing the seemingly weaker signals and provides very little intuition regarding how information is distributed across them. This motivates *disentangled representation learning*, where the goal is to decompose the multimodal

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

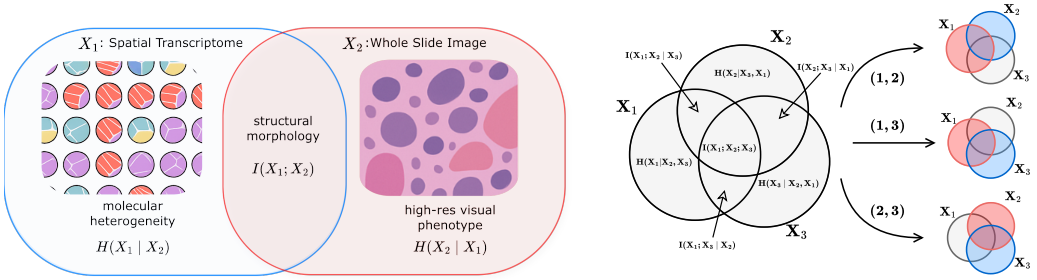


Figure 1: **Left:** Example in oncology when multi-view redundancy is limited. While both modalities capture shared structural morphology, WSI resolves fine-grained cellular features, whereas ST reveals underlying molecular variations, that are invisible in histology. **Right:** Information Venn diagram for three modalities, along with their pairwise shared and unique component visualizations.

information into distinct, meaningful components, separating what is shared across modalities from what remains modality-specific. Identifying the underlying shared and unique factors, reveals which modalities are most task-critical, which contribute unique information, and which exhibit significant overlap. This, in turn, can inform practical decisions such as pruning modalities that are computationally or cost demanding, yet *redundant*. Additionally, such a decomposition can facilitate adaptability to unseen tasks, integration of new heterogeneous modalities and efficient fine-tuning on new datasets or cohorts.

Despite recent progress [Wu et al., 2025, Liang et al., 2023], reliably disentangling multimodal information remains an open challenge. Importantly, the majority of prior work remains confined to two modalities, largely due to a fundamental scalability bottleneck, arising from two main factors. First, the number of possible shared and modality-specific factors grows exponentially with the number of modalities, making them increasingly difficult to model. Second, beyond two modalities, tractable objectives or bounds that explicitly characterize all these interactions remain limited. Moreover, in real-world settings, unique and shared information is often highly entangled, leading to a trade-off between minimizing redundancy across components and preserving as much relevant information as possible [Wang et al., 2025]. Capturing only the Minimal Necessary Information (MNI) [Fischer, 2020] is therefore key to learning informative, non-redundant components without inducing representational collapse. These challenges naturally raise a fundamental question:

"How can we scalably extract the underlying unique and shared factors beyond two modalities?"

We answer this question with **RePercENT**, a novel information-theoretic framework for **Re**duced-complexity **Perc**eiver-based dis**ENT**anglement. Our approach effectively addresses the scalability barrier, through two key ingredients. First, to avoid the exponential growth of components, we model pairwise interactions between modalities: this preserves the information of each modality while capturing all relevant cross-modal interactions–see Figure 1 (right). Second, we propose an architecture that redefines the efficient, modality-agnostic design of the Perceiver encoder [Jaegle et al., 2021] through dedicated structural mechanisms that effectively route the available information into shared and modality-specific representations. Furthermore, by leveraging pre-extracted foundation model embeddings and a self-supervised training regime, our method remains modality- and task-agnostic, while still being robust to missing or incomplete modalities at inference time. Finally, our approach is accompanied with formal theoretical guarantees on the optimality of our solution. Our contributions are as follows:

- **Scalable multimodal disentanglement:** We introduce RePercENT, a novel information-theoretic framework for multimodal disentanglement that efficiently scales beyond the two-modality setting.
- **Plug-and-play architecture:** RePercENT operates directly on pre-extracted embeddings, making it agnostic to modality types and backbone architectures. By design, it accommodates missing data during inference and seamlessly transfers to unseen tasks.
- **Formal optimality guarantees:** We provide formal optimality guarantees both when minimal necessary information (MNI) is attainable and when it is not.
- **Extensive empirical validation:** From synthetic data to figurative language and oncology, we demonstrate the robustness and effectiveness of our framework across nine real-world modalities and five domains.

Overall, we provide a principled framework that bridges information-theoretic foundations and formal guarantees with a practical and efficient implementation, paving the way towards more reliable, interpretable, and mechanistically grounded multimodal representations.

2 Problem formulation

In this section, we formalize the problem of information-theoretic disentangled representation learning in multimodal settings and introduce the necessary notation and definitions.

We denote as $\mathcal{X} = \{X_1, X_2, \ldots, X_M\}$ the set of M modalities, and let $\mathcal{M} = \{1, 2, \ldots, M\}$ be the corresponding index set. Let Z_i be a latent, information-preserving representation of X_i. We associate each subset $A \subseteq \mathcal{M}$ with its corresponding latent representation z_A, which we term *atomic representation* (see Definition 2.1), reflecting the distinct information subspaces between modalities (see Figure 2).

Definition 2.1. Atomic representation. For any non-empty subset $A \subseteq \mathcal{M}$, we define as the atomic representation z_A, the latent representation that captures information shared by all modalities in A and exclusive to that subset, such that

$$I(z_A; z_{A'}) = 0 \qquad \forall\, A' \subseteq \mathcal{M} \text{ with } A \cap A' = \varnothing, \tag{1}$$

where $I(z_A; z_{A'})$ is the mutual information between z_A and $z_{A'}$.

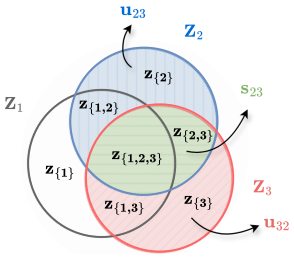


Figure 2: Atomic representation decomposition for three modalities, through a Venn diagram illustration.

Notice that each modality-specific latent representation Z_i can then be interpreted as a composition of atomic representations:

Definition 2.2. Composite representation. For $i \in \mathcal{M}$, let $\mathcal{A}_i = \{A \subseteq \mathcal{M} : i \in A\}$. The composite representation Z_i is the combination of all atomic representations whose subsets contains i: $Z_i = \bigoplus_{A \in \mathcal{A}_i} z_A$. Here, $\bigoplus$ denotes a generic composition operation, e.g., concatenation or any aggregation function.

Subsequently, using the atomic representations as building blocks, we define shared and modality-specific representations for each modality pair (i, j), according to Definition 2.3.

Definition 2.3. Pairwise unique and shared representations. For $i, j \in \mathcal{M}$, let $\mathcal{A}_{i|j} = \mathcal{A}_i \cap \mathcal{A}_j^c$, where $\mathcal{A}_j^c$ denotes the complementary set of $\mathcal{A}_j$, and let $\mathcal{A}_{ij} = \mathcal{A}_i \cap \mathcal{A}_j$. The pairwise unique and shared representations are defined as,

$$\mathbf{u}_{ij} = \bigoplus_{A \in \mathcal{A}_{i|j}} z_A, \qquad \mathbf{s}_{ij} = \bigoplus_{A \in \mathcal{A}_{ij}} z_A.$$

Here, $\mathbf{u}_{ij}$ and $\mathbf{s}_{ij}$ capture information unique to i relative to j, and shared by i and j, respectively.

Finally, we formalize the learning objective as follows:

Problem Statement Given the composite representation Z_i of each modality $i \in \mathcal{M}$, for every $j \in \mathcal{M} \setminus \{i\}$, our goal is to derive the pairwise disentangled latent representations $\mathbf{u}_{ij}$ and $\mathbf{s}_{ij}$:

$$Z_i \longmapsto \{(\mathbf{u}_{ij}, \mathbf{s}_{ij})\}_{j \in \mathcal{M} \setminus \{i\}}.$$

Notice that the above formulation efficiently circumvents the exponential growth in the number of atomic representations[1] reducing the component complexity from $\mathcal{O}(2^M)$ to $\mathcal{O}(M^2)$, while explicitly capturing all pairwise cross-modal interactions. Importantly, no modality information is lost, as for each $j \neq i$, $(\mathbf{u}_{ij}, \mathbf{s}_{ij})$ encapsulate the complete information reflected in Z_i. Consequently, even the higher order components z_A, while not explicitly disentangled are fully retained through the pairwise components, residing in $\mathbf{s}_{ij}$ whenever $i, j \in A$ and in $\mathbf{u}_{ij}$ whenever $i \in A$ and $j \notin A$.

[1] Each composite representation Z_i consists of exactly 2^{M-1} atomic representations.

3 RePercENT: Unlocking scalable disentangled representations

Building on this formulation, we introduce RePercENT. We first present the architectural design, then define the optimization objective leading to tractable training losses, and finally derive theoretical guarantees on the optimality of our solution.

3.1 Architecture overview

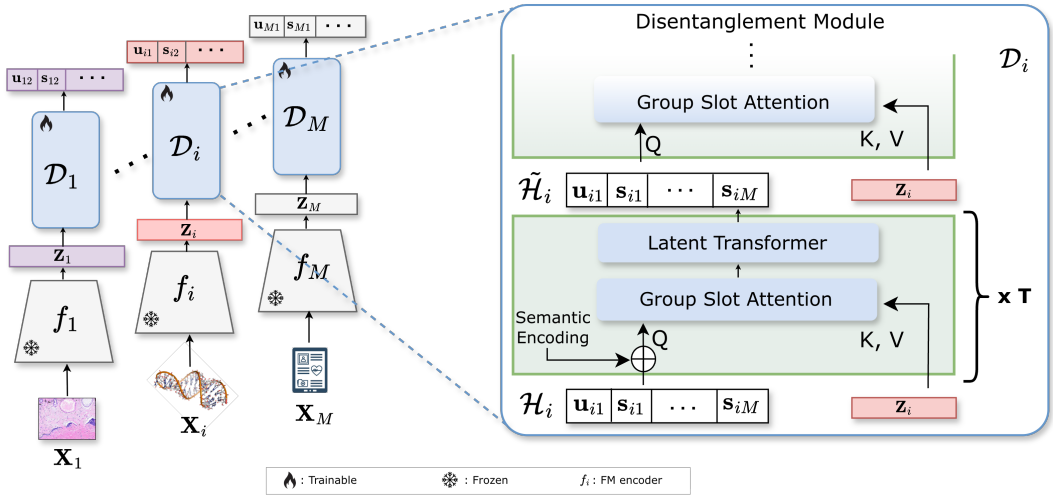


Figure 3: *Model overview*. Each modality X_i is first encoded, using modality-specific FMs. Afterwards, each encoded representation is processed through its dedicated disentanglement module $\mathcal{D}_i$.

We propose a scalable framework for multimodal disentanglement that extracts the desired information-theoretic representations while requiring *a single encoder* per modality, yielding linear scaling in M. Specifically, from each modality i, we leverage pre-trained modality specific foundation models (FMs) to obtain the initial embeddings Z_i. Under the assumption that these encoders are sufficiently expressive and thus the resulting representations preserve most of the relevant modality information, we treat the embeddings Z_i, as composite representations. Subsequently, each representation is processed via a separate disentangling module $\mathcal{D}_i$, that decomposes Z_i into all desired information components involving modality i, i.e., $Z_i \longmapsto \{(\mathbf{u}_{ij}, \mathbf{s}_{ij})\}_{j \in \mathcal{M} \setminus \{i\}}$.

Disentanglement module Within each $\mathcal{D}_i$, we model the disentangled representations using a learnable latent array $\mathcal{H}_i \in \mathbb{R}^{N \times D}$, where $N = 2(M-1)$ denotes the number of latent slots per modality and D is the dimension of each slot. To link pairwise unique and shared components to the latent array, we assign each row $h_{ik} \in \mathbb{R}^D$ of $\mathcal{H}_i$ to exactly one component (unique or shared) of modality i (see Appendix F.1). $\mathcal{H}_i$ is then iteratively refined through a Perceiver-inspired latent attention mechanism. While Jaegle et al. [2021] introduced the Perceiver as a general-purpose architecture for scalable input processing, we leverage its latent attention mechanism for a different objective: learning granular disentangled representations. In our setup, $\mathcal{H}_i$ serves as a query over the encoded representation Z_i, followed by a latent self-attention block, as illustrated in Figure 3. This latent bottleneck is well suited to our setting, as it enables efficient representation encoding without making any modality-specific assumptions. To encourage disentanglement, we introduce two complementary routing mechanisms: **Semantic Encoding**, which assigns a functional role to each latent slot, and **Group Slot Attention**, a novel attention mechanism, that promotes competition among slots.

Semantic encoding (SE) To promote component specialization, each slot h_{ik} is augmented with two learnable encodings: a pair encoding, $e^{\mathrm{p}}_{ij} \in \mathbb{R}^D$, which identifies the modality pair (i, j), and a type encoding, $e^{\mathrm{t}}_{ij} \in \mathbb{R}^D$, indicating the component type, $\mathrm{t} \in \{\mathrm{unique}, \mathrm{shared}\}$. The initialized latent slot is thus defined as: $\tilde{h}_{ik} = h_{ik} + e^{\mathrm{p}}_{ij} + e^{\mathrm{t}}_{ij}$.

Group slot attention (GSA) Rather than compressing modality X_i into a single pooled embedding, we retain the full sequence of embeddings $Z_i \in \mathbb{R}^{S_i \times E_i}$, allowing the latent slots to attend to fine-grained input structure. We then introduce *group slot attention*: a structural routing mechanism tailored to pairwise disentanglement. While motivated by the competitive assignment principle of Slot Attention [Locatello et al., 2020], our mechanism applies competition within modality-pair groups. Specifically, for each pair (i, j), the slots corresponding to $(\mathbf{u}_{ij}, \mathbf{s}_{ij})$ compete over the same input embeddings, enforcing specialization within the pair. This way, each group effectively separates modality i's information into i-specific and i-j shared latent representations (see Appendix F).

Together, these design choices impose an explicit routing bias, enabling each slot to specialize in its assigned representation while keeping the inference modality-local. This offers a central scalability advantage, as all pairwise unique and shared representations are obtained with only one encoder per modality. In contrast, state-of-the-art disentanglement methods [Wu et al., 2025, Liang et al., 2023] rely on separate encoders for different representations, leading to $\mathcal{O}(M^2)$ structural complexity, which quickly becomes prohibitive as M grows. Our method reduces this to $\mathcal{O}(M)$ while retaining the full set of pairwise decompositions. Moreover, by construction, for every $j \in \mathcal{M}$, the representations $(\mathbf{u}_{ij}, \mathbf{s}_{ij})$ are inferred from X_i alone, i.e., $\mathbf{u}_{ij} \sim p(\cdot \mid X_i)$, and $\mathbf{s}_{ij} \sim p(\cdot \mid X_i)$. Equivalently,

$$(\mathbf{u}_{ij}, \mathbf{s}_{ij}) \perp X_j \mid X_i, \qquad (\mathbf{u}_{ji}, \mathbf{s}_{ji}) \perp X_i \mid X_j. \tag{2}$$

Crucially, this decoupling ensures robustness to missing modalities during inference. Since $(\mathbf{u}_{ij}, \mathbf{s}_{ij})$ are inferred directly from X_i, they remain available whenever X_i is observed, regardless of whether X_j is present, bypassing the need for complex imputation or heuristic mapping.

3.2 Information criteria for optimal disentangled representations

In order to learn the desired representations, we cast each pairwise decomposition as a single optimization problem. Unlike recent approaches that follow a sequential strategy, in which the shared representation is learned first and the modality-specific representation is subsequently inferred as a residual component [Wang et al., 2025], we propose to simultaneously optimize all representations $\theta_{ij} = (u_{ij}, u_{ji}, s_{ij}, s_{ji})$, according to the following optimization objective:

Joint optimization objective. For each pair (i, j), we optimize $\theta_{ij} = (u_{ij}, u_{ji}, s_{ij}, s_{ji})$,

$$\hat{\theta}_{ij} \in \arg\max_{\theta_{ij}} \mathcal{J}(\theta_{ij}) = \arg\max_{\theta_{ij}} \left[\alpha(\mathcal{L}_{s_i} + \mathcal{L}_{s_j}) + \mathcal{L}_{u_i} + \mathcal{L}_{u_j}\right], \text{ where,} \tag{3}$$

$$\mathcal{L}_{s_i} = I(s_{ij}; X_j) - \beta I(s_{ij}; X_i \mid X_j), \quad \mathcal{L}_{u_i} = I(u_{ij}, s_{ji}; X_i) - \lambda I(u_{ij}; s_{ji}) \tag{4}$$

$$\mathcal{L}_{s_j} = I(s_{ji}; X_i) - \beta I(s_{ji}; X_j \mid X_i), \quad \mathcal{L}_{u_j} = I(u_{ji}, s_{ij}; X_j) - \lambda I(u_{ji}; s_{ij}). \tag{5}$$

The shared objectives $\mathcal{L}_{s_i}, \mathcal{L}_{s_j}$, encourage each shared representation to capture all *necessary* cross-modal information, while at the same time being *minimal*, i.e. penalizing information that can only be explained by its source modality. Specifically, the mutual information $I(s_{ij}; X_j)$, promotes cross-modal relevance, whereas the conditional mutual information, $I(s_{ij}; X_i \mid X_j)$, suppresses cross-modal leakage. Complementarily, the unique objectives $\mathcal{L}_{u_i}, \mathcal{L}_{u_j}$ preserve source-modality information via $I(u_{ij}, s_{ji}; X_i)$, while enforcing cross-modal disentanglement by penalizing overlap between unique and shared components across modalities through $I(u_{ij}; s_{ji})$. We introduce, the hyperparameter $\alpha > 0$ acting as a weighting constant between the shared and unique objectives, while (β, λ) [Wang et al., 2025] control the trade-off between representation coverage and redundancy.

3.2.1 Tractable training objectives.

For training the framework, we instantiate the information-theoretic components in Eq. (3) using tractable, principled relaxations of the quantities in Eqs. (3) - (5). Specifically, we employ the InfoNCE objective [van den Oord et al., 2018], which variationally lower-bounds mutual information to estimate the terms $I(s_{ij}; X_j)$ and $I(u_{ij}, s_{ji}; X_i)$. For the latter, positive pairs are formed from two independently augmented views of X_i. We model $I(s_{ij}; X_i \mid X_j)$ using a KL-divergence penalty, $\mathcal{L}_{\text{KL}}$, between the conditional distributions $p(s_{ij} \mid X_i)$ and $p(s_{ji} \mid X_j)$, as it provides a variational upper bound on the corresponding conditional mutual information [Federici et al., 2020]. Lastly, we reduce the dependence between u_{ij} and s_{ji} via a cross-covariance penalty, $\mathcal{L}_{\text{xcov}}$, on standardized representations. Concretely, Eq. (4) is approximated as

$$\mathcal{L}_{s_i} \approx \mathcal{L}^{\text{tr}}_{s_i} = -\mathcal{L}^{\text{INCE}}_{s_i} - \beta\, \mathcal{L}_{\text{KL}}, \qquad \mathcal{L}_{u_i} \approx \mathcal{L}^{\text{tr}}_{u_i} = -\mathcal{L}^{\text{INCE}}_{u_i} - \lambda\, \mathcal{L}_{\text{xcov}}. \tag{6}$$

The terms corresponding to Eq. (5) are defined analogously by exchanging i and j. Averaging over all modality pairs, we obtain the full trainable objective

$$\mathcal{L}^{\text{tr}}_{\mathcal{E}} = \frac{1}{|\mathcal{E}|} \sum_{(i,j)\in\mathcal{E}} \left[\alpha \left(\mathcal{L}^{\text{tr}}_{s_i} + \mathcal{L}^{\text{tr}}_{s_j}\right) + \mathcal{L}^{\text{tr}}_{u_i} + \mathcal{L}^{\text{tr}}_{u_j}\right], \qquad \mathcal{E} = \{(i,j) \in \mathcal{M}^2 : i < j\}. \tag{7}$$

Detailed expressions and further discussion on the losses can be found in Appendix F.3.

3.2.2 Optimality guarantees

We now establish theoretical guarantees for our joint objective relative to the optimal information-theoretic decomposition. Specifically, we show that under attainable *Minimum Necessary Information* (MNI) [Fischer, 2020], our proposed objective admits the same global optimum as the ideal optimizer. Moreover, we prove near-optimality, in the more challenging regime, when MNI is unattainable.

Optimal information-theoretic objectives Concretely, considering the ideal sequential optimization problem [Wang et al., 2025], the optimal shared and unique representations, can be derived following:

$$\textbf{shared:} \qquad s^*_{ij} \in \arg\max_{s_{ij}} \mathcal{L}^o_{s_i} = \arg\max_{s_{ij}} I(s_{ij}; X_j) - \beta\, I(s_{ij}; X_i \mid X_j), \tag{8}$$

$$\textbf{unique:} \qquad u^*_{ij} \in \arg\max_{u_{ij}} \mathcal{L}^o_{u_i} = \arg\max_{u_{ij}} I(u_{ij}, X_j; X_i) - \lambda I(u_{ij}; s^*_{ij}).^{2} \tag{9}$$

Notice that the unique representations are inferred from the optimal solutions of the first step.

From sequential to joint optimization For the case when MNI holds, Theorem (1), demonstrates that any step-by-step optimal solution, is also a global maximizer of our joint objective.

Theorem 1. *Exact optimality under attainable MNI:* Let $\theta^*_{ij} = (u^*_{ij}, u^*_{ji}, s^*_{ij}, s^*_{ji})$ be an optimal solution of the step-by-step optimization problem defined in Eqs. (8) and (9). If the MNI criterion is attainable and $\alpha \geq \lambda$, then θ^*_{ij} is a global optimizer of the joint objective defined in Eqs. (3) - (5). Equivalently,

$$\theta^*_{ij} \in \arg\max_{\theta_{ij}} \mathcal{J}(\theta_{ij}).$$

Proof. See Appendix E.

Moreover, our guarantees extend beyond the idealized setting where the MNI criterion holds. This is essential, as real multimodal data rarely admit such a clean separation: shared and modality-specific factors can be inherently entangled. Therefore, when MNI is unattainable, we show that the maximum value of our joint objective is no worse than the optimal value of the ideal problem, up to a tractable constant, as stated in Theorem (2). Further discussion of MNI is provided in Appendix D.

Theorem 2. *Near-optimality under unattainable MNI:* Let $\theta^*_{ij} = (u^*_{ij}, u^*_{ji}, s^*_{ij}, s^*_{ji})$ be an optimal solution of the ideal step-by-step optimizer as stated in Eqs. (8) - (9), and $\hat{\theta}_{ij}$ the corresponding optimal solution of the joint optimization problem, as stated in Eqs. (3) - (5). If MNI is unattainable, any maximizer $\hat{\theta}_{ij}$ of the joint objective satisfies

$$\mathcal{J}(\hat{\theta}_{ij}) \geq \mathcal{J}^o(\theta^*_{ij}) - 2(1+\lambda)\delta_c,$$

where,

$$\mathcal{J}^o(\theta^*_{ij}) = \alpha\mathcal{L}^o_{s_i}(\theta^*_{ij}) + \alpha\mathcal{L}^o_{s_j}(\theta^*_{ij}) + \mathcal{L}^o_{u_i}(\theta^*_{ij}) + \mathcal{L}^o_{u_j}(\theta^*_{ij}),$$

and δ_c denotes the information-gap term introduced in Wang et al. [2025], which quantifies the trade-off between coverage and redundancy in the shared representations.

Proof. See Appendix E.

Together, Theorems (1) and (2) establish that our coupled objective is not merely a practical surrogate, but a formally grounded solution for the ideal step-by-step solution. This is not just a convenience, but rather a prerequisite for scalability beyond two modalities, where a sequential treatment of all shared and unique components becomes computationally prohibitive. By reconciling joint optimization with

[2] The representations s_{ji}, u_{ji} are accordingly derived by exact symmetry.

formal guarantees under both attainable and unattainable MNI, we offer a robust solution that scales without sacrificing theoretical integrity.

4 Experiments

In this section, we evaluate RePercENT across a challenging synthetic setup, and three real-world benchmarks, spanning five datasets and nine modalities in total. Our analysis focuses on three core axes: i) the trade-off between representation recovery and scalability as the number of modalities grows, ii) the quality and downstream utility of the resulting decomposition relative to state-of-the-art alignment-based and disentanglement approaches, and iii) robustness to missing modalities.

4.1 Experimental setup

Datasets For the synthetic experiments, we generate each $Z_i \in \mathbb{R}^E$, by sampling independently from a normal distribution all the relevant atomic representations $z_A \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_{d_\ell}),\ i \in A$. Unlike prior work [Liang et al., 2023, Wang et al., 2025], we further map each Z_i into a sequence representation rather than compressing it into a single pooled vector, mimicking how information is distributed in realistic settings, e.g. across image patches or text tokens. This construction provides direct access to the ground-truth unique and shared representations $\mathbf{u}_{ij}$ and $\mathbf{s}_{ij}$, allowing us to conduct a linear probing experiment. Specifically, we associate each ground-truth $\mathbf{u}_{ij}$ and $\mathbf{s}_{ij}$, with binary labels $y_{u_{ij}}$ and $y_{s_{ij}}$ respectively, via a non-linear deterministic projection. The learned representation $\hat{\mathbf{u}}_{ij}$, should accurately predict $y_{u_{ij}}$, while remaining uninformative about $y_{s_{ij}}$. Analogously, $\hat{\mathbf{s}}_{ij}$ should exhibit the opposite pattern. Δ_{model} averages the absolute deviations of these accuracies from their ideal values (see Appendix G.1 for further details). Regarding real-world datasets, we provide an overview along with the considered downstream tasks in Table 1.

Table 1: Real-world benchmarks. Modalities: T = `Text`, V = `Video`, A = `Audio`, I = `Image`, C = `Caption`, D = `Definition`, WSI = `Whole Slide Image`, Mol = `Molecular`, Path = `Pathology` reports, Clin = `Clinical` records. †Patients, train / test split. ‡Following Yosef et al. [2023]; see Appendix G.3.

Dataset	Downstream Task	Modalities	# Samples	Preprocessing
MOSI	Binary sentiment classification (YouTube videos)	T, V, A	2,199	MultiBench [Liang et al., 2021]
MOSEI	Binary emotion classification (monologue videos)	T, V, A	23,000	MultiBench
UR-FUNNY	Binary humor classification (TED talks)	T, V, A	>16,000	MultiBench
IRFL	Four-way image choice for a figurative phrase (see Figure 7)	I, C, D	3,380	Ours‡
TCGA [Weinstein et al., 2013]	Cancer-type classification (10 cancer types)	WSI, Mol, Clin, Path	4,585 / 1,147†	HONeYBEE [Tripathi et al., 2025]

Baselines For the synthetic experiments, we compare RePercENT, under a shared training regime, against the architectures used by leading disentanglement approaches, each with separate representation-specific encoders, including MLP, gMLP [Liu et al., 2021], and GRU [Chung et al., 2014] variants. For IRFL, we additionally include CLIP baselines, trained with the standard contrastive objective by Radford et al. [2021], offering a direct comparison between alignment-based approaches and disentangled decomposition. As most existing disentanglement methods are limited to two modalities, we compare against DisentangledSSL [Wang et al., 2025], FOCAL [Liu et al., 2023], JointOpt [Pan et al., 2021], and FactorCL [Liang et al., 2023] on the MultiBench Text–Video pairs. In the oncology setting, we use the original HONeYBEE embeddings, assessing whether our decomposition improves the utility of existing biomedical representations. Finally, for missing-modality robustness on WSIs and molecular data, we consider three fusion strategies: late fusion with averaging, early fusion trained with either mean imputation or modality dropout. These are applied both to the raw embeddings and ClipSur, our parameter-matched baseline of per-modality transformer encoders aligned with the standard CLIP contrastive loss.

Table 2: Linear probe accuracy (%) on MultiBench, reported as (mean ± std.) across three random seeds. For each model we present the best result among the latent representations: unique, shared and the concatenation of both. For each dataset, the best result among the reported rows is shown in bold.

Model	MOSI (↑)	MOSEI (↑)	UR-FUNNY (↑)	Params (↓)	FLOPs (↓)
FactorCL	52.87 ± 0.96	75.71 ± 2.83	58.76 ± 1.87	21.98M	3.789B
FOCAL	60.93 ± 0.36	77.83 ± 0.63	62.67 ± 0.94	17.68M	1.674B
JointOpt	61.42 ± 2.56	77.40 ± 0.35	62.04 ± 0.47	17.68M	1.675B
DisentangledSSL	**65.35 ± 1.48**	74.33 ± 0.59	61.56 ± 0.88	17.68M	0.835B
RePercENT	62.59 ± 0.83	**77.99 ± 0.32**	**62.95 ± 0.87**	**7.34M**	**0.052B**

4.2 Results

Scalability and component recovery Figure 4 reports Δ_{model}, w.r.t. parameter count as the number of modalities grows. Multi-encoder baselines scale by adding one encoder per derived representation, with fixed per-component capacity. RePercENT instead adds latent slots and only one disentanglement module per new modality, making it the only method with linear scaling in M while maintaining competitive disentanglement performance. The $M = 2$ confusion matrix further shows strong intra-component prediction and limited cross-component leakage, indicating effective separation of unique and shared components. A detailed breakdown in Appendix H.1 shows that only RePercENT and gMLP recover all components, while GRU and MLP suffer representation collapse as M increases. These results confirm that successful component recovery depends on both the objective and the architecture. Similar conclusions are drawn for MultiBench, when observing Table 2, where RePercENT matches or outperforms the baselines using $\mathbf{58.5 - 66.6}\%$ less parameters and $\mathbf{93.8 - 98.6}\%$ less compute (FLOPs). Notably, in the three-modality regime (Text-Video-Audio), where the other methods are not applicable, RePercENT's parameters are even lower than all baselines at M=2, as shown in Table 14, Appendix H.2. Accordingly on IRFL, Table 18 in Appendix H.3 shows that gMLP requires roughly 2× more parameters and 7× more FLOPs than RePercENT without improving figurative language detection performance.

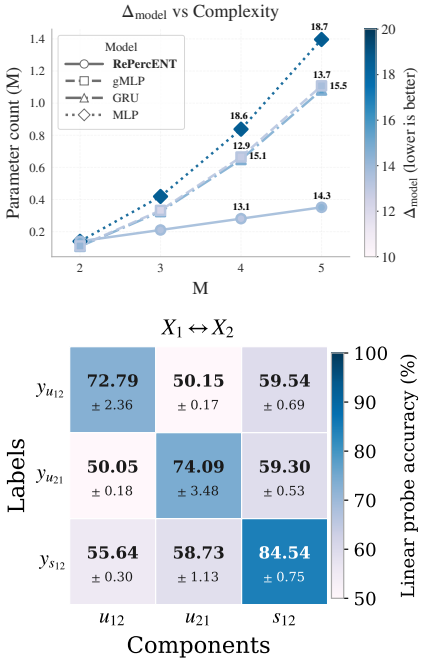


Figure 4: **Top:** Synthetic performance across modality count and parameter budgets. RePercENT achieves competitive scores at substantially lower model complexity. **Bottom:** Linear-probe confusion matrix for $M = 2$.

Downstream performance Beyond scaling, we investigate under what conditions, and to what extent, such a granular decomposition is beneficial for downstream tasks. On IRFL (Table 3), disentanglement generally improves over fine-tuned CLIP variants, and RePercENT achieves the best overall and out-of-distribution (OoD) accuracy on the figurative language task. Comparing our method with alignment approaches, we observe that isolating the shared component yields superior performance, especially for idioms and metaphors, but is not ideal for similes. This is consistent with the linguistic nature of the task, as similes are typically more explicit and structurally constrained, favoring direct cross-modal alignment. By contrast, our principled decomposition appears to benefit idioms and mainly metaphors since they are more abstract and context-dependent. The ablation further supports RePercENT's key structural mechanisms: removing both semantic encodings and group slot attention gives the weakest performance, while enabling either components improves accuracy. Using both results in the best performance, improving overall accuracy from 40.30 to 44.07, with large gains on idioms (28.2 → 33.60) and metaphors (28.87 → 35.56).

Likewise, RePercENT's per-component results on MultiBench (Table 13, Appendix H.2) mirror how label information is distributed across modalities. Unique text is strongest on MOSI/MOSEI, where

Table 3: Top-1 accuracy (%)on the IRFL detection task, reported as (mean ± std) over 5 random seeds.

Model	SE	GSA	Idioms	Simile	Metaphor–OoD	Overall
Baselines						
CLIP-ViT-B/32 (zero-shot)	–	–	16.00	45.49	23.42	29.14
CLIP-ViT-B/32 (projection-only FT)	–	–	20.4 ± 1.0	68.38 ± 1.1	31.59 ± 1.0	41.41 ± 0.7
CLIP-ViT-B/32 (end-to-end FT)	–	–	22.5 ± 1.8	$\mathbf{70.18 \pm 0.6}$	29.61 ± 1.0	41.73 ± 1.0
gMLP	–	–	33.70 ± 2.4	54.37 ± 2.1	28.23 ± 2.4	38.52 ± 0.7
GRU	–	–	$\mathbf{36.10 \pm 3.1}$	48.81 ± 3.0	33.69 ± 1.9	39.46 ± 1.1
Architectural ablations						
RePercENT	✗	✗	28.20 ± 5.0	63.97 ± 2.0	27.87 ± 3.96	40.30 ± 1.7
RePercENT	✗	✓	30.20 ± 4.3	63.11 ± 2.9	30.45 ± 3.38	41.56 ± 1.34
RePercENT	✓	✗	32.60 ± 2.2	61.88 ± 0.9	31.95 ± 5.7	42.35 ± 2.24
RePercENT	✓	✓	33.60 ± 5.3	61.88 ± 3.1	$\mathbf{35.56 \pm 3.4}$	$\mathbf{44.07 \pm 0.8}$

sentiment is conveyed mainly through language, whereas the shared component is strongest on UR-FUNNY, as laughter depends on delivery as well as content. Together, these results reflect the inherent structure of the data and show that task-relevant information is not only shared across modalities.

On TCGA (Table 4) this fine-grained decomposition improves accuracy in cancer type classification over HONeYBEE embeddings for most cancer types, with the largest gains observed on molecular data. Most strikingly, for KIRC and LGG, where raw molecular embeddings are nearly uninformative, it recovers substantially higher accuracy, indicating that cross-modal conditioning exposes predictive structure hidden in the original embedding space. Notably, the full decomposition D_{ij} consistently outperforms both U_{ij} and S_{ij} alone, further challenging the assumption of multi-view redundancy.

Table 4: Cancer type prediction accuracy (%, mean ± std) on TCGA cohorts. HONeYBEE embeddings vs. RePercENT components of modality i relative to j: unique U_{ij}, shared S_{ij}, and complete $D_{ij} = U_{ij} \oplus S_{ij}$. **ΔRePercENT** denotes the absolute gains over HONeYBEE.

Representation	BRCA	COAD	GBM	HNSC	KIRC	LGG	LUAD	LUSC	OV	PRAD
WSI										
Honeybee	95.4	32.2	41.5	18.3	48.6	22.6	26.0	14.3	36.2	29.0
$U_{\text{wsi,mol}}$	79.2 ± 1.8	39.3 ± 3.7	39.8 ± 2.5	28.5 ± 3.2	61.3 ± 4.2	28.8 ± 2.6	24.2 ± 2.5	22.7 ± 1.3	47.8 ± 3.1	37.8 ± 1.9
$S_{\text{wsi,mol}}$	81.6 ± 1.5	37.2 ± 4.8	39.8 ± 2.0	26.7 ± 3.0	61.1 ± 2.3	24.3 ± 2.3	26.0 ± 2.6	20.8 ± 3.1	46.9 ± 2.5	34.8 ± 2.8
$D_{\text{wsi,mol}}$	79.3 ± 1.8	39.1 ± 2.7	40.2 ± 1.5	28.7 ± 2.2	63.4 ± 2.1	31.0 ± 2.8	27.8 ± 2.2	23.3 ± 1.5	51.2 ± 1.4	38.8 ± 2.8
Δ RePercENT	**-13.8**	**+7.1**	**-1.3**	**+10.4**	**+14.8**	**+8.4**	**+1.8**	**+9.0**	**+15.0**	**+9.8**
Molecular										
Honeybee	69.8	21.8	100.0	31.7	2.8	2.9	58.0	51.0	65.5	74.0
$U_{\text{mol,clin}}$	67.9 ± 0.9	32.9 ± 3.7	98.1 ± 0.4	66.7 ± 1.9	60.6 ± 2.9	50.0 ± 2.1	51.0 ± 3.4	65.3 ± 3.8	80.3 ± 1.4	91.0 ± 0.7
$S_{\text{mol,clin}}$	67.5 ± 2.9	33.6 ± 2.7	98.3 ± 0.0	66.0 ± 1.5	60.4 ± 3.1	49.2 ± 1.6	50.8 ± 4.3	66.7 ± 1.7	78.5 ± 2.9	91.8 ± 0.5
$D_{\text{mol,clin}}$	70.0 ± 1.4	37.5 ± 3.1	98.3 ± 0.0	67.9 ± 1.1	62.2 ± 1.6	52.8 ± 1.8	55.4 ± 5.2	72.0 ± 2.8	79.5 ± 1.4	92.0 ± 0.0
Δ RePercENT	**+0.2**	**+15.7**	**-1.7**	**+36.2**	**+59.4**	**+49.9**	**-2.6**	**+21.0**	**+14.8**	**+18.0**

Measuring Disentanglement In real-world settings, downstream performance indicates the utility and semantic information of the learned disentangled representations. Beyond this empirical evaluation, we argue that a faithful decomposition should satisfy three testable properties. (i) **Shared-subspace invariance:** the shared components of both modalities should span a consistent subspace, assessed through similarity metrics such as CKA and mutual-kNN alignment [Huh et al., 2024]. (ii) **Cross-modal exclusion:** a unique component, u_{ij} should contain minimal information about the partner modality j, evaluated through its ability to reconstruct that modality. (iii) **Complementarity:** u_{ij} and s_{ij} should be more informative jointly than individually, measured via the Reconstruction Gain (RG) [Wang et al., 2025]. On MultiBench (Table 5), no model leaks information (all scores ≤ 0). RePercENT significantly outperforms all baselines on both alignment metrics, with similarly good performance in terms of RG.

Table 5: Disentanglement metrics on MOSEI.

	Reconstruction Gain (↑)		Shared Space Alignment (↑)		Cross-Modal Leakage (≤ 0)	
Model	RG-`Text`	RG-`Video`	CKA	mKNN	$U_{T,V} \to X_V$	$U_{V,T} \to X_T$
FOCAL	$0.027_{\pm.001}$	$0.025_{\pm.001}$	$0.117_{\pm.012}$	$0.033_{\pm.002}$	$-0.011_{\pm.001}$	$-0.006_{\pm.002}$
JointOpt	$0.060_{\pm.017}$	$0.034_{\pm.003}$	$0.108_{\pm.012}$	$0.040_{\pm.002}$	$-0.021_{\pm.003}$	$-0.020_{\pm.006}$
DisentangledSSL	$0.015_{\pm.002}$	$0.054_{\pm.010}$	$0.007_{\pm.001}$	$0.019_{\pm.0004}$	$-0.001_{\pm.005}$	$-0.005_{\pm.002}$
RePercENT	$\mathbf{0.076}_{\pm.004}$	$\mathbf{0.076}_{\pm.005}$	$\mathbf{0.838}_{\pm.010}$	$\mathbf{0.356}_{\pm.022}$	$\mathbf{-0.003}_{\pm.003}$	$\mathbf{-0.014}_{\pm.001}$

Robustness to missing modalities Figure 5 examines the robustness of different models in terms of cancer type prediction, as the fraction of available `WSI` samples decreases. All fusion baselines exhibit a clear performance drop, including those explicitly trained with modality dropout. This suggests that simply accounting for missing modalities during training is insufficient when the models still rely on jointly combining modalities at inference. In contrast, RePercENT maintains substantially higher Macro-F1 across all missingness levels. This behavior follows directly from its structure (recall Eq. 2), which allows, both the unique information of the `Molecular` w.r.t. the `WSI` as well as the shared between the two modalities, to remain fully accessible. Notably, whenever `WSI` information is available, RePercENT improves over both single-modality references, further supporting the argument that there is significant complementary information between the two modalities.

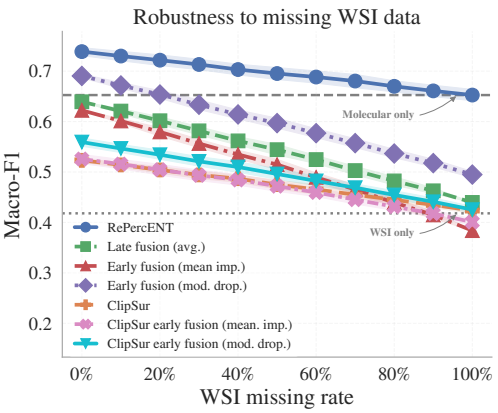


Figure 5: We reduce `WSI` availability while preserving `Molecular` data. Fusion baselines degrade sharply, whereas RePercENT remains robust.

5 Limitations and future directions

RePercENT establishes a scalable and theoretically grounded framework for disentanglement across multiple modalities, inviting further exploration in several promising research directions. Our formulation focuses on pairwise unique and shared components, which provide a provably sufficient granularity for modeling the interactions considered in this work. While this choice comes with favorable scaling and formal guarantees, an important direction for future work is modeling richer, higher order interactions among three or more modalities through more expressive decompositions.

In addition, our practical implementation depends on the quality of the pre-extracted modality embeddings. While foundation models provide strong representations and enable a plug-and-play design, they also introduce a trade-off between expressivity and compression [Tishby et al., 1999]. This motivates future work on end-to-end training, in settings with sufficient data, which could yield richer and more informative representations. Another promising direction is to dynamically balance the objectives across modality pairs, potentially improving optimization efficiency and the quality of the learned disentangled components. Finally, the structured nature of the learned decomposition opens an exciting path toward more interpretable multimodal models, motivating future work to examine how these components encode, exchange, and use information to shape the model's predictions for different downstream tasks.

6 Conclusion

As multimodal intelligence aims to derive representations that capture the world's complex multimodal reality, it becomes essential to decouple the nuanced interactions across diverse modalities. In this work, we introduce RePercENT, and demonstrate that, in settings involving more than two modalities, disentangled representations i) benefit downstream tasks, ii) can be learned computationally efficiently, while iii) remaining theoretically grounded. Through an extensive validation of our framework across synthetic, figurative language, and complex biomedical benchmarks, we demonstrate competitive performance across diverse modalities, foundation model backbones and tasks, while scaling seamlessly as the number of modalities grows. Ultimately, we believe that Re*Perc*ENT, via its flexible structure and granular representations, will help pave the way towards a more comprehensive, multimodal *Perc*eption.

7 Acknowledgements

This work was supported by the Swiss National Science Foundation (SNSF) under Lead Agency grant 200021E_229452. We thank the anonymous reviewers for their insightful feedback and constructive discussion.

References

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

Felix Krones, Umar Marikkar, Guy Parsons, Adam Szmul, and Adam Mahdi. Review of multimodal machine learning approaches in healthcare. *An International Journal on Information Fusion*, 114, 2024.

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 *International Conference on Machine Learning*, 2021.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. 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*, 2021.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millicah, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, 2022.

Paul Pu Liang, Zihao Deng, Martin Q. Ma, James Zou, Louis-Philippe Morency, and Ruslan Salakhutdinov. Factorized contrastive learning: going beyond multi-view redundancy. In *Advances in Neural Information Processing Systems*, 2023.

Megan Tjandrasuwita, Chanakya Ekbote, Liu Ziyin, and Paul Pu Liang. Understanding the emergence of multimodal representation alignment. In *International Conference on Machine Learning*, 2025.

Konstantinos Kontras, Thomas Strypsteen, Christos Chatzichristos, Paul Pu Liang, Matthew Blaschko, and Maarten De Vos. Balancing Multimodal Training Through Game-Theoretic Regularization. In *Advances in Neural Information Processing Systems*, 2025.

Qilong Wu, Yiyang Shao, Jun Wang, and Xiaobo Sun. Learning optimal multimodal information bottleneck representations. In *International Conference on Machine Learning*, 2025.

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

Ian Fischer. The conditional entropy bottleneck. *Entropy*, 2020.

Andrew Jaegle, Felix Gimeno, Andrew Brock, Andrew Zisserman, Oriol Vinyals, and Joao Carreira. Perceiver: General perception with iterative attention. In *International Conference on Machine Learning*, 2021.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. In *Advances in Neural Information Processing Systems*, 2020.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *ArXiv*, abs/1807.03748, 2018.

Marco Federici, Anjan Dutta, Patrick Forré, Nate Kushman, and Zeynep Akata. Learning robust representations via multi-view information bottleneck. In *International Conference on Learning Representations*, 2020.

Ron Yosef, Yonatan Bitton, and Dafna Shahaf. Irfl: Image recognition of figurative language. In *Conference on Empirical Methods in Natural Language Processing*, 2023.

Paul Pu Liang, Yiwei Lyu, Xiang Fan, Zetian Wu, Yun Cheng, Jason Wu, Leslie Chen, Peter Wu, Michelle A. Lee, Yuke Zhu, Ruslan Salakhutdinov, and Louis-Philippe Morency. MultiBench: Multiscale Benchmarks for Multimodal Representation Learning. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M Stuart. The cancer genome atlas pan-cancer analysis project. *Nature Genetics*, 2013.

A. Tripathi, A. Waqas, M. B. Schabath, et al. Honeybee: enabling scalable multimodal ai in oncology through foundation model-driven embeddings. *npj Digital Medicine*, 2025.

Hanxiao Liu, Zihang Dai, David So, and Quoc V Le. Pay attention to MLPs. In *Advances in Neural Information Processing Systems*, 2021.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling, December 2014. arXiv:1412.3555 [cs].

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. In *Advances in Neural Information Processing Systems*, 2023.

Ziqi Pan, Li Niu, Jianfu Zhang, and Liqing Zhang. Disentangled Information Bottleneck. In *AAAI Conference on Artificial Intelligence*, 2021.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. In *International Conference on Machine Learning*, 2024.

Naftali Tishby, Fernando Pereira, and William Bialek. The information bottleneck method. *Allerton Conference on Communication, Control and Computing*, 1999.

Paul Pu Liang, Amir Zadeh, and Louis philippe Morency. Foundations & trends in multimodal machine learning: Principles, challenges, and open questions. *ACM Computing Surveys*, 2022a.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, 2022.

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

Fabian Gröger, Shuo Wen, Huyen Le, and Maria Brbic. With limited data for multimodal alignment, let the STRUCTURE guide you. In *Advances in Neural Information Processing Systems*, 2025.

Noël Vouitsis, Zhaoyan Liu, Satya Krishna Gorti, Valentin Villecroze, Jesse C. Cresswell, Guangwei Yu, Gabriel Loaiza-Ganem, and Maksims Volkovs. Data-efficient multimodal fusion on a single gpu. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, 2019.

Antonio Norelli, Marco Fumero, Valentino Maiorca, Luca Moschella, Emanuele Rodolà, and Francesco Locatello. ASIF: Coupled data turns unimodal models to multimodal without training. In *Advances in Neural Information Processing Systems*, 2023.

Bo Zhou, Liulei Li, Yujia Wang, Huafeng Liu, Yazhou Yao, and Wenguan Wang. Unialign: Scaling multimodal alignment within one unified model. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Konstantin Hemker, Nikola Simidjievski, and Mateja Jamnik. Healnet: multimodal fusion for heterogeneous biomedical data. In *Advances in Neural Information Processing Systems*, 2024.

Richard J. Chen, Tong Ding, Ming Y. Lu, et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 2024.

Yingxue Xu and Hao Chen. Multimodal optimal transport-based co-attention transformer with global structure consistency for survival prediction. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

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. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Paul Pu Liang, Yiwei Lyu, Xiang Fan, Jeffrey Tsaw, Yudong Liu, Shentong Mo, Dani Yogatama, Louis philippe Morency, and Ruslan Salakhutdinov. High-modality multimodal transformer: Quantifying modality & interaction heterogeneity for high-modality representation learning. *Transactions on Machine Learning Research*, 2022b.

Xin Wang, Hong Chen, Si'ao Tang, Zihao Wu, and Wenwu Zhu. Disentangled representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Aniek Eijpe, Soufyan Lakbir, Melis Erdal Cesur, Sara Pires de Oliveira, Sanne Abeln, and Wilson Silva. Disentangled and interpretable multimodal attention fusion for cancer survival prediction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2025.

Lu Gao, Wenlan Chen, Daoyuan Wang, Fei Guo, and Cheng Liang. Disentangled Cross-Modal Representation Learning with Enhanced Mutual Supervision. *Advances in Neural Information Processing Systems*, 2025.

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. *AAAI Conference on Artificial Intelligence*, 2024.

Lucas Robinet, Ahmad Berjaoui, Ziad Kheil, and Elizabeth Cohen-Jonathan Moyal. Drim: Learning disentangled representations from incomplete multimodal healthcare data. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2024.

Min Namgung, Yijun Lin, JangHyeon Lee, and Yao-Yi Chiang. Less is more: Multimodal region representation via pairwise inter-view learning. *arXiv preprint arXiv:2505.18178*, 2025.

Alessandro Achille and Stefano Soatto. Information dropout: Learning optimal representations through noisy computation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? In *Advances in Neural Information Processing Systems*, 2020.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxin Yang, Jingren Zhou, Jingren Zhou, Junyan Lin, Kai Dang, Keqin Bao, Ke-Pei Yang, Le Yu, Li-Chun Deng, Mei Li, Min Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shi-Qiang Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yi-Chao Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025.

Asim Waqas, Aakash Tripathi, Sabeen Ahmed, Ashwin Mukund, Hamza Farooq, Matthew B Schabath, Paul Stewart, Mia Naeini, and Ghulam Rasool. Senmo: A self-normalizing deep learning model for enhanced multi-omics data analysis in oncology. *arXiv preprint arXiv:2405.08226*, 2024.

Appendix Contents

A Useful Notation **15**

B Supplementary Related Work **15**

C Impact Statement **16**

D Information Theory Background **16**

D.1 Basic quantities and identities 16

D.2 Multi-view redundancy 17

D.3 Minimum Necessary Information 17

D.4 Information gap under unattainable MNI 17

E Proofs **18**

E.1 Proof of Theorem 1 18

E.2 Proof of Theorem 2 21

F Framework Implementation Details **24**

F.1 Slot assignment in latent array $\mathcal{H}$ 24

F.2 Group slot attention 25

F.3 Training objectives 26

G Experimental Details **26**

G.1 Synthetic data creation 26

G.2 MultiBench overview 28

G.3 IRFL overview 30

G.4 TCGA dataset extraction 31

G.5 Computational Resources 34

H Additional Experiments **34**

H.1 Synthetic Dataset 34

H.2 MultiBench 42

H.3 IRFL 43

H.4 TCGA 44

A Useful Notation

Table 6: Notation summary.

Symbol	Meaning
Sets and Indices	
M	Total number of modalities
$\mathcal{M}$	Modality index set $\{1, \ldots, M\}$
$\mathcal{E}$	Set of all unique modality pairs $\{(i,j) \in \mathcal{M}^2 : i < j\}$
$A \subseteq \mathcal{M}$	A non-empty subset of modalities
$\mathcal{A}^u_{ij}, \mathcal{A}^s_{ij}$	Index sets for atomic components unique to i (w.r.t j) or shared by $\{i,j\}$
Representations and Components	
X_i	Input modality i
Z_i	Composite latent representation for modality i
z_A	Atomic representation associated with subset A
$\mathbf{u}_{ij}, \mathbf{s}_{ij}$	Unique and shared latent components for modality i relative to modality j
θ_{ij}	Optimization variable as a collection of all pairwise components $(u_{ij}, u_{ji}, s_{ij}, s_{ji})$
Architectural Elements	
$\mathcal{D}_i$	Disentangling module for modality i
$\mathcal{H}_i$	Latent-slot array ($\mathbb{R}^{N \times D_l}$)
h_{ik}	The k-th individual row of $\mathcal{H}_i$
ϕ_i	Mapping function from latent slots to pairwise components for modality i
Hyperparameters	
α	Weighting parameter between unique and shared losses
β	Penalty parameter for shared component coverage
λ	Penalty parameter for redundancy between unique and shared components

B Supplementary Related Work

Multimodal representation learning has revolutionized the way we process and integrate information coming from different modalities [Liang et al., 2022a, Krones et al., 2024]. Recent advances in multimodal foundation models [Team et al., 2023, Li et al., 2022, Singh et al., 2022] reflect a growing trend toward flexible, general-purpose AI systems. However, these models typically require large amounts of paired data and extensive pre-training, making them difficult to apply in domains, such as medicine, where paired multimodal data are limited.

To address this limitation, several approaches have been proposed, such as STRUCTURE [Gröger et al., 2025] and FuseMix [Vouitsis et al., 2023], that align representations into a shared space using frozen foundation models in low-data regimes. Complementary, training-free approaches have also been explored. For instance, Centered Kernel Alignment (CKA) [Kornblith et al., 2019] provides a way to compare representation spaces without additional training, while ASIF [Norelli et al., 2023] constructs a common multimodal space from paired anchor data. On the other hand, UniAlign by Zhou et al. [2025], improves scalability by using a single encoder to align diverse modalities.

While alignment approaches improve efficiency and exploit the expressive power of foundation models, they mainly capture shared cross-modal information. Fusion methods overcome this [Hemker et al., 2024, Chen et al., 2024, Xu and Chen, 2023], by leveraging joint representations that additionally encode complementary information across modalities. ImageBind [Girdhar et al., 2023] is another notable example that learns a joint embedding space for six different modalities, showing that not all modality-pairs need to be observed. Additionally, Liang et al. [2022b] propose HighMMT that learns efficient-joint representations, by scaling multimodal transformers to many modalities and measuring modality and interaction heterogeneity to guide parameter sharing.

Despite their success, fusion methods usually result to less interpretable representations, and face a difficulty in handling missing modalities. A promising, recent direction to address this is through disentanglement [Wang et al., 2024], where the different modality-interactions are explicitly encoded in separate representations. For example, DIMAF [Eijpe et al., 2025] models shared and specific components between spatial transcriptomics and histology, improving downstream performance and interpretability. Gao et al. [2025], DrFuse [Yao et al., 2024], and DRIM [Robinet et al., 2024]

further explore disentangled fusion under incomplete-modality settings, while FactorCL [Liang et al., 2023] formalizes multi-view redundancy and separates shared from unique task-relevant information. CooKIE [Namgung et al., 2025] extends FactorCL beyond two modalities for geospatial region representation, replacing explicit high-order factorization with a pairwise inter-view learning approach that reduces complexity in factorized multimodal region representation learning. However, it leaves the resulting cross-pair redundancy to be handled by the downstream predictor and offers no optimality guarantees for its decomposition. More recently, Wang et al. [2025] propose information-theoretic criteria for controlled two-modality disentanglement.

C Impact Statement

The primary objective of this paper is to advance multimodal representation learning under a flexible framework that extracts disentangled pairwise unique and shared components, with applications spanning figurative language, and biomedical research. Although advances in multimodal representation learning may have both positive and negative societal and ethical implications, particularly in areas like biomedical research, we currently do not foresee any immediate societal concerns associated with the proposed methodology.

D Information Theory Background

In this section we include useful definitions, and key background elements on which we build upon our information theory objectives.

D.1 Basic quantities and identities

We begin with some basic notions from classical information theory as well as some key inequalities that have been essential for the proofs that follow.

Mutual information (MI) Let (X, Y) be a set of continuous random variables over the space $\mathcal{X} \times \mathcal{Y}$. The mutual information $I(X;Y)$ is defined as:

$$I(X;Y) = \iint_{\mathcal{X}\times\mathcal{Y}} p_{X,Y}(x,y) \log \left(\frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)} \right) dx\, dy = D_{KL}(p_{X,Y} \parallel p_X p_Y),$$

where $p_{X,Y}$ is the joint probability density function of X, Y, p_X and p_Y are the marginal probability density functions of X and Y respectively, and D_{KL} is the *Kullback-Leibler divergence*. Intuitively, MI measures how much knowing one of these variables reduces uncertainty about the other.

Conditional mutual information (CMI) Let X, Y and Z be three continuous random variables over the spaces $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$. The conditional mutual information (CMI) $I(X;Y \mid Z)$ is denoted as:

$$\begin{aligned} I(X;Y) &= \iiint_{\mathcal{X}\times\mathcal{Y}\times\mathcal{Z}} p_{X,Y,Z}(x,y,z) \log \left(\frac{p_{X,Y,Z}(x,y,z)p_Z(z)}{p_{X,Z}(x,z)p_{Y,Z}(y,z)} \right) dx\, dy\, dz \\ &= \int_{\mathcal{Z}} D_{KL}(p_{(X,Y)|Z} \,||\, p_{X|Z} \otimes p_{Y|Z})\, dz, \end{aligned}$$

where all the conditional, marginal and joint probability distributions are denoted as p with the corresponding subscript. Similarly to the MI, CMI expresses the mutual information between X and Y when conditioning on a third variable Z.

Useful properties We present below several useful identities & inequalities, used extensively in the following proofs.

- $I(X;Y) \geq 0$, non-negativity of MI
- $I(X;Y \mid Z) \geq 0$, non-negativity of CMI
- $I(X;Y) = I(Y;X)$, symmetry
- $I(X;Y,Z) = I(X;Y \mid Z) + I(X;Z)$, chain rule of MI

- $I(X;Y) - I(X;Z) = I(X;Y \mid Z) - I(X;Z \mid Y)$.

The last one is derived, by applying the chain rule of MI in its two equivalent forms.

D.2 Multi-view redundancy

We extend the notion of multi-view redundancy, proposed by Liang et al. [2023], to the general multimodal setting as follows:

Definition D.1. *Multi-view redundancy*: Let $\mathcal{X} = \{X_1, X_2, \ldots, X_M\}$ be a set of M modalities, and Y a downstream task. Let also, $X_{-i} = \{X_1, \ldots, X_{i-1}, X_{i+1}, \ldots X_M\}$ represent the set of all modalities except X_i. We say that multi-view redundancy holds, if there exists sufficiently small $\epsilon > 0$, such that:

$$I(X_{-i}; Y \mid X_i) < \epsilon, \quad \forall i \in \mathcal{M}.$$

This formulation implies that each X_i captures sufficient information to perform the task Y and any complementary information from the remaining modalities offers insignificant predictive gain.

D.3 Minimum Necessary Information

In the simpler unimodal case, assume a dataset (X, Y) with observations X and target labels Y. Given the Markov structure $Z \leftarrow X \leftrightarrow Y$, we want to derive representations Z that capture useful information from X that is relevant to task Y [Tishby et al., 1999]. As introduced by Fischer [2020], the *Minimum Necessary Information* (MNI) criterion for a representation Z is satisfied, when i) the representation Z attains all the *necessary* information from X to perform Y, and ii) given all the possible representations Z that solve the task Y, the desired Z needs to be *minimal*, i.e. achieve $\inf_{Z \in \mathcal{Z}} I(Z; X, Y)$. In other words, at the MNI point it holds:

$$I(X;Y) = I(X;Z) = I(Y;Z). \tag{10}$$

Following Wang et al. [2025], we extend this idea to the multimodal self-supervised setting by considering observations (X_i, X_j) from two distinct modalities. We would like to derive the representations s_{ij} and s_{ji}, respecting the Markov structures: $s_{ij} \leftarrow X_i \leftrightarrow X_j$ and $s_{ji} \leftarrow X_j \leftrightarrow X_i$, that optimally balance between expressivity and compression. In other words, s_{ij} is desired to capture all the necessary information from X_i that is shared with X_j, while minimizing redundancy with X_i. The same holds for s_{ji}. Formally, when the MNI point is attainable, s_{ij} and s_{ji} satisfy,

$$I(s_{ij}; X_i) = I(s_{ij}; X_j) = I(X_i; X_j), \qquad I(s_{ji}; X_i) = I(s_{ji}; X_j) = I(X_i; X_j).$$

D.4 Information gap under unattainable MNI

In real world scenarios, the MNI point described in section D.3 is not always attainable, and therefore Wang et al. [2025] propose the following definition for extracting the optimal shared representations s_{ij} and s_{ji}:

$$\begin{aligned} s^*_{ij} &\in \arg\min_{s_{ij}} I(s_{ij}; X_i \mid X_j), \quad \text{s.t. } I(X_i; X_j) - I(s_{ij}; X_j) \leq \delta_c, \\ s^*_{ji} &\in \arg\min_{s_{ji}} I(s_{ji}; X_j \mid X_i), \quad \text{s.t. } I(X_i; X_j) - I(s_{ji}; X_i) \leq \delta_c. \end{aligned} \tag{11}$$

Minimizing $I(s_{ij}; X_i \mid X_j)$ penalizes information retained by s_{ij} about X_i that is not explained by X_j. The constraint, on the other hand, ensures that s_{ij} preserves nearly all shared information between X_i and X_j, up to an information gap δ_c. The same interpretation applies symmetrically to s_{ji}. Notice that if MNI was always attainable, we could substitute $I(X_i; X_j) - I(s_{ij}; X_j) \leq \delta_c$ with $I(X_i; X_j) = I(s_{ij}; X_j)$, as suggested by Achille and Soatto [2018]. However, when MNI is unattainable, this leads to a suboptimal solution and therefore the parameter δ_c is introduced, which controls the gap between capturing the complete shared information and having a compressed representation. Using the Lagrangian formulation of constrained problem in Eq. (11), we end up with the first part of the optimal step-by-step optimization objective:

$$\begin{aligned} s^*_{ij} \in \arg\max_{s_{ij}} \mathcal{L}^o_{s_i} &= \arg\max_{s_{ij}} \; I(s_{ij}; X_j) - \beta\, I(s_{ij}; X_i \mid X_j), \\ s^*_{ji} \in \arg\max_{s_{ji}} \mathcal{L}^o_{s_j} &= \arg\max_{s_{ji}} \; I(s_{ji}; X_i) - \beta\, I(s_{ji}; X_j \mid X_i). \end{aligned} \tag{12}$$

Here, $\beta > 0$ controls the trade-off between preserving cross-modal shared information and removing modality-specific information. As shown by Wang et al. [2025], when MNI is attainable, then for any positive β, maximizing the objective $\mathcal{L}_i^o$ achieves MNI, and there exists a bijective mapping from β in $\mathcal{L}_{s_i}^o$ to the value of the information constraint δ_c as defined in Eq. (11).

E Proofs

We first derive the following Lemma 1 under attainable MNI, that will help us with the proof of Theorem 1.

Lemma 1. Let (s_{ij}^*, s_{ji}^*) be an optimal solution of the step-by-step optimization problem as defined in Eq. (8). Under attainable MNI and $u_{ij} \perp (X_j, s_{ij}^*) \mid X_i$,

$$I(u_{ij}; s_{ij}^*) = I(u_{ij}; X_j) = I(u_{ij}; s_{ji}^*),$$

for every feasible u_{ij}.

Proof. Using the standard mutual-information identity, we get

$$I(u_{ij}; X_j) - I(u_{ij}; s_{ij}^*) = I(u_{ij}; X_j \mid s_{ij}^*) - I(u_{ij}; s_{ij}^* \mid X_j).$$

For attainable MNI and under Proposition 1 by Wang et al. [2025]) for s_{ij}^* it holds that,

$$I(s_{ij}^*; X_i \mid X_j) = 0, \qquad I(s_{ij}^*; X_i) = I(X_i; X_j).$$

Now, from the data processing inequality one also gets

$$0 \leq I(u_{ij}; s_{ij}^* \mid X_j) \leq I(X_i; s_{ij}^* \mid X_j) = 0 \implies I(u_{ij}; s_{ij}^* \mid X_j) = 0.$$

For attainable MNI at s_{ij}^* we also have $I(X_i; X_j \mid s_{ij}^*) = 0$, which follows from standard identities and since u_{ij} depends only on X_i, $I(u_{ij}; X_j \mid s_{ij}^*) = 0$, which proves $I(u_{ij}; s_{ij}^*) = I(u_{ij}; X_j)$.

The equality with s_{ji}^* follows symmetrically. □

E.1 Proof of Theorem 1

Theorem 1. *Exact optimality under attainable MNI:* Let $\theta_{ij}^* = (u_{ij}^*, u_{ji}^*, s_{ij}^*, s_{ji}^*)$ be an optimal solution of the step-by-step optimization problem defined in Eqs. (8) and (9). If the MNI criterion is attainable and $\alpha \geq \lambda$, then θ_{ij}^* is a global optimizer of the joint objective defined in Eqs. (3) - (5). Equivalently,

$$\theta_{ij}^* \in \arg\max_{\theta_{ij}} \mathcal{J}(\theta_{ij}).$$

Proof. Let (s_{ij}^*, s_{ji}^*) and (u_{ij}^*, u_{ji}^*) denote the optimal shared and unique representations obtained from the step-by-step optimization in Eqs. (8) and (9) respectively. We proceed by deriving an upper bound on the joint objective and then showing that this bound is attained at θ^* when MNI is attainable.

We begin with the term $\mathcal{L}_{u_i}$ in the joint objective:

$$\begin{aligned}
\mathcal{L}_{u_i} &= I(u_{ij}, s_{ji}; X_i) - \lambda I(u_{ij}; s_{ji}) \\
&= I(u_{ij}, s_{ji}; X_i) - \lambda I(u_{ij}; s_{ji}) \pm \lambda I(u_{ij}; X_j) \\
&\leq I(u_{ij}, X_j; X_i) - \lambda I(u_{ij}; X_j) + \lambda \big[I(u_{ij}; X_j) - I(u_{ij}; s_{ji}) \big]. \qquad (13)
\end{aligned}$$

The inequality follows from

$$I(u_{ij}, s_{ji}; X_i) \leq I(u_{ij}, X_j; X_i),$$

which holds due to the Markov relation $s_{ji} \leftarrow X_j \leftrightarrow X_i$.

Next, we use the identity

$$I(u_{ij}; X_j) - I(u_{ij}; s_{ji}) = I(u_{ij}; X_j \mid s_{ji}) - I(u_{ij}; s_{ji} \mid X_j). \qquad (14)$$

Since the joint distribution factorizes according to the Markov structure

$$s_{ji} \leftarrow X_j \leftrightarrow X_i \rightarrow u_{ij},$$

we have $u_{ij} \perp s_{ji} \mid X_j$, and therefore

$$I(u_{ij}; s_{ji} \mid X_j) = 0.$$

Hence Eq. (14) becomes

$$\begin{aligned} I(u_{ij}; X_j) - I(u_{ij}; s_{ji}) &= I(u_{ij}; X_j \mid s_{ji}) \\ &\leq I(X_i; X_j \mid s_{ji}), && \text{by } u_{ij} \leftarrow X_i \leftrightarrow X_j \\ &= I(X_i; X_j, s_{ji}) - I(X_i; s_{ji}) \\ &= I(X_i; X_j) - I(X_i; s_{ji}), && \text{by } s_{ji} \leftarrow X_j \leftrightarrow X_i. \end{aligned} \tag{15}$$

Substituting Eq. (15) into Eq. (13) yields

$$\begin{aligned} \mathcal{L}_{u_i} &\leq \Big[I(u_{ij}, X_j; X_i) - \lambda I(u_{ij}; X_j) \Big] + \lambda I(X_i; X_j) - \lambda I(X_i; s_{ji}) \\ &= \mathbf{B}_{u_i} - \lambda I(X_i; s_{ji}) + c, \end{aligned} \tag{16}$$

where

$$\mathbf{B}_{u_i} := I(u_{ij}, X_j; X_i) - \lambda I(u_{ij}; X_j), \qquad c := \lambda I(X_i; X_j) \geq 0.$$

By symmetry, the same argument gives

$$\mathcal{L}_{u_j} \leq \mathbf{B}_{u_j} - \lambda I(X_j; s_{ij}) + c, \tag{17}$$

where

$$\mathbf{B}_{u_j} := I(u_{ji}, X_i; X_j) - \lambda I(u_{ji}; X_i).$$

We now combine these bounds in the full joint objective:

$$\begin{aligned} \mathcal{J} &= \alpha(\mathcal{L}_{s_i} + \mathcal{L}_{s_j}) + \mathcal{L}_{u_i} + \mathcal{L}_{u_j} \\ &\leq \alpha(\mathcal{L}_{s_i} + \mathcal{L}_{s_j}) + \mathbf{B}_{u_i} + \mathbf{B}_{u_j} + 2c - \lambda I(X_i; s_{ji}) - \lambda I(X_j; s_{ij}). \end{aligned} \tag{18}$$

Define the modified shared objectives

$$\tilde{\mathcal{L}}_{s_i} := \alpha \mathcal{L}_{s_i} - \lambda I(X_j; s_{ij}), \tag{19}$$

and

$$\tilde{\mathcal{L}}_{s_j} := \alpha \mathcal{L}_{s_j} - \lambda I(X_i; s_{ji}). \tag{20}$$

Using the definition of $\mathcal{L}_{s_i}$, we obtain

$$\begin{aligned} \tilde{\mathcal{L}}_{s_i} &= \alpha \big[I(s_{ij}; X_j) - \beta I(s_{ij}; X_i \mid X_j) \big] - \lambda I(X_j; s_{ij}) \\ &= (\alpha - \lambda) I(s_{ij}; X_j) - \alpha\beta I(s_{ij}; X_i \mid X_j) \\ &\leq (\alpha - \lambda) I(X_i; X_j), \qquad \text{for } \alpha \geq \lambda. \end{aligned} \tag{21}$$

The last step follows from

$$I(s_{ij}; X_j) \leq I(X_i; X_j)$$

by the assumed Markov structure, together with the non-negativity of conditional mutual information.

By symmetry,

$$\tilde{\mathcal{L}}_{s_j} \leq (\alpha - \lambda) I(X_i; X_j). \tag{22}$$

Substituting Eqs. (21) and (22) into Eq. (18) gives

$$\mathcal{J} \leq 2(\alpha - \lambda) I(X_i; X_j) + \mathbf{B}_{u_i} + \mathbf{B}_{u_j} + 2c, \qquad \forall\, u_{ij}, u_{ji}. \tag{23}$$

We now show that this bound is attained at the step-by-step shared optimum when MNI is attainable. Since s^*_{ij} and s^*_{ji} satisfy the MNI criterion, we have

$$I(s^*_{ij}; X_j) = I(s^*_{ij}; X_i) = I(X_i; X_j), \qquad I(s^*_{ji}; X_i) = I(s^*_{ji}; X_j) = I(X_i; X_j),$$

and therefore

$$I(s^*_{ij}; X_i \mid X_j) = 0, \qquad I(s^*_{ji}; X_j \mid X_i) = 0$$

as well as,

$$I(u_{ij}; X_j \mid s^*_{ji}) = I(X_i; X_j \mid s^*_{ji}), \qquad I(u_{ji}; X_i \mid s^*_{ij}) = I(X_j; X_i \mid s^*_{ij}).$$

Moreover, in the attainable-MNI regime, Proposition 3 by Wang et al. [2025] implies that

$$I(u_{ij}, s^*_{ji}; X_i) = I(u_{ij}, X_j; X_i), \qquad I(u_{ji}, s^*_{ij}; X_j) = I(u_{ji}, X_i; X_j),$$

while Lemma (1) suggests that $I(u_{ij}; s^*_{ij}) = I(u_{ij}; X_j) = I(u_{ij}; s^*_{ji})$ for every feasible u_{ij}.

It follows that all inequalities above become equalities when $(s_{ij}, s_{ji}) = (s^*_{ij}, s^*_{ji})$. Hence

$$\begin{aligned} \mathcal{J} &= 2(\alpha - \lambda) I(X_i; X_j) + \mathbf{B}_{u_i} + \mathbf{B}_{u_j} + 2c \\ &= 2\alpha I(X_i; X_j) + \mathbf{B}_{u_i} + \mathbf{B}_{u_j}, \qquad c = \lambda I(X_i; X_j). \end{aligned} \tag{24}$$

Since the term $2\alpha I(X_i; X_j)$ is constant with respect to u_{ij} and u_{ji}, maximizing $\mathcal{J}$ is equivalent to maximizing

$$\mathbf{B}_{u_i} + \mathbf{B}_{u_j}.$$

Furthermore, $\mathbf{B}_{u_i}$ depends only on u_{ij}, while $\mathbf{B}_{u_j}$ depends only on u_{ji}. Therefore the maximization decouples as

$$\arg \max_{u_{ij}, u_{ji}} \left(\mathbf{B}_{u_i} + \mathbf{B}_{u_j} \right) = \arg \max_{u_{ij}} \mathbf{B}_{u_i} \times \arg \max_{u_{ji}} \mathbf{B}_{u_j}.$$

By construction, this is exactly the optimization problem for the unique components in Eq. (9) under attainable MNI. Therefore the step-by-step optimizer (u^*_{ij}, u^*_{ji}) also maximizes the joint surrogate objective. Consequently,

$$\theta^*_{ij} \in \arg \max_{\theta_{ij}} \mathcal{J}(\theta_{ij}),$$

which proves the claim. □

E.2 Proof of Theorem 2

Theorem 2. *Near-optimality under unattainable MNI:* Let $\theta^*_{ij} = (u^*_{ij}, u^*_{ji}, s^*_{ij}, s^*_{ji})$ be an optimal solution of the ideal step-by-step optimizer as stated in Eqs. (8) - (9), and $\hat{\theta}_{ij}$ the corresponding optimal solution of the joint optimization problem, as stated in Eqs. (3) - (5). If MNI is unattainable, any maximizer $\hat{\theta}_{ij}$ of the joint objective satisfies

$$\mathcal{J}(\hat{\theta}_{ij}) \geq \mathcal{J}^o(\theta^*_{ij}) - 2(1+\lambda)\delta_c,$$

where,

$$\mathcal{J}^o(\theta^*_{ij}) = \alpha\mathcal{L}^o_{s_i}(\theta^*_{ij}) + \alpha\mathcal{L}^o_{s_j}(\theta^*_{ij}) + \mathcal{L}^o_{u_i}(\theta^*_{ij}) + \mathcal{L}^o_{u_j}(\theta^*_{ij}),$$

and δ_c denotes the information-gap term introduced in Wang et al. [2025], which quantifies the trade-off between coverage and redundancy in the shared representations.

Proof. Assume now that MNI is unattainable. Let

$$\theta^*_{ij} = (u^*_{ij}, u^*_{ji}, s^*_{ij}, s^*_{ji})$$

denote the solution of the ideal step-by-step optimization problem defined in Eqs. (8) and (9), i.e.,

$$\begin{aligned}
s^*_{ij} \in \arg\max_{s_{ij}} \mathcal{L}^o_{s_i} &= \arg\max_{s_{ij}} \Big[I(s_{ij}; X_j) - \beta\, I(s_{ij}; X_i \mid X_j) \Big], \\
s^*_{ji} \in \arg\max_{s_{ji}} \mathcal{L}^o_{s_j} &= \arg\max_{s_{ji}} \Big[I(s_{ji}; X_i) - \beta\, I(s_{ji}; X_j \mid X_i) \Big], \\
u^*_{ij} \in \arg\max_{u_{ij}} \mathcal{L}^o_{u_i} &= \arg\max_{u_{ij}} \Big[I(u_{ij}, X_j; X_i) - \lambda I(u_{ij}; s^*_{ij}) \Big], \\
u^*_{ji} \in \arg\max_{u_{ji}} \mathcal{L}^o_{u_j} &= \arg\max_{u_{ji}} \Big[I(u_{ji}, X_i; X_j) - \lambda I(u_{ji}; s^*_{ji}) \Big].
\end{aligned}$$

We define the corresponding ideal reweighted value at θ^*_{ij} as

$$\begin{aligned}
\mathcal{J}^o(\theta^*_{ij}) = {} & \alpha I(s^*_{ij}; X_j) - \alpha\beta I(s^*_{ij}; X_i \mid X_j) \\
& + \alpha I(s^*_{ji}; X_i) - \alpha\beta I(s^*_{ji}; X_j \mid X_i) \\
& + I(u^*_{ij}, X_j; X_i) - \lambda I(u^*_{ij}; s^*_{ij}) \\
& + I(u^*_{ji}, X_i; X_j) - \lambda I(u^*_{ji}; s^*_{ji}).
\end{aligned}$$

Equivalently,

$$\mathcal{J}^o(\theta^*_{ij}) = \alpha\mathcal{L}^o_{s_i}(\theta^*_{ij}) + \alpha\mathcal{L}^o_{s_j}(\theta^*_{ij}) + \mathcal{L}^o_{u_i}(\theta^*_{ij}) + \mathcal{L}^o_{u_j}(\theta^*_{ij}).$$

Next, for an arbitrary feasible point

$$\theta_{ij} = (u_{ij}, u_{ji}, s_{ij}, s_{ji}),$$

the joint objective is

$$\begin{aligned}
\mathcal{J}(\theta_{ij}) = {} & \alpha I(s_{ij}; X_j) - \alpha\beta I(s_{ij}; X_i \mid X_j) \\
& + \alpha I(s_{ji}; X_i) - \alpha\beta I(s_{ji}; X_j \mid X_i) \\
& + I(u_{ij}, s_{ji}; X_i) - \lambda I(u_{ij}; s_{ji}) \\
& + I(u_{ji}, s_{ij}; X_j) - \lambda I(u_{ji}; s_{ij}).
\end{aligned}$$

Let

$$\hat{\theta}_{ij} \in \arg\max_{\theta_{ij}} \mathcal{J}(\theta_{ij}).$$

By definition of $\hat{\theta}_{ij}$, we have

$$\mathcal{J}(\hat{\theta}_{ij}) \geq \mathcal{J}(\theta^*_{ij}), \qquad \hat{\theta}_{ij} \in \arg\max_{\theta_{ij}} \mathcal{J}(\theta_{ij}). \tag{25}$$

Evaluating the joint objective at θ^*_{ij}, we obtain

$$\begin{aligned}
\mathcal{J}(\theta^*_{ij}) = {} & \alpha I(s^*_{ij}; X_j) - \alpha\beta I(s^*_{ij}; X_i \mid X_j) \\
& + \alpha I(s^*_{ji}; X_i) - \alpha\beta I(s^*_{ji}; X_j \mid X_i) \\
& + I(u^*_{ij}, s^*_{ji}; X_i) - \lambda I(u^*_{ij}; s^*_{ji}) \\
& + I(u^*_{ji}, s^*_{ij}; X_j) - \lambda I(u^*_{ji}; s^*_{ij}).
\end{aligned}$$

We now consider the difference between the ideal reweighted value and the joint objective at θ^*_{ij}:

$$\begin{aligned}\mathcal{J}^o(\theta^*_{ij}) - \mathcal{J}(\theta^*_{ij}) = &\Big[I(u^*_{ij}, X_j; X_i) - I(u^*_{ij}, s^*_{ji}; X_i)\Big] \\ &+ \Big[I(u^*_{ji}, X_i; X_j) - I(u^*_{ji}, s^*_{ij}; X_j)\Big] \\ &+ \lambda\Big[I(u^*_{ij}; s^*_{ji}) - I(u^*_{ij}; s^*_{ij})\Big] \\ &+ \lambda\Big[I(u^*_{ji}; s^*_{ij}) - I(u^*_{ji}; s^*_{ji})\Big].\end{aligned}$$

For brevity, we define

$$\Delta_{u_i} := I(u^*_{ij}, X_j; X_i) - I(u^*_{ij}, s^*_{ji}; X_i), \qquad \Lambda_{u_i} := I(u^*_{ij}; s^*_{ji}) - I(u^*_{ij}; s^*_{ij}),$$

and similarly

$$\Delta_{u_j} := I(u^*_{ji}, X_i; X_j) - I(u^*_{ji}, s^*_{ij}; X_j), \qquad \Lambda_{u_j} := I(u^*_{ji}; s^*_{ij}) - I(u^*_{ji}; s^*_{ji}).$$

Then

$$\mathcal{J}^o(\theta^*_{ij}) - \mathcal{J}(\theta^*_{ij}) = \Delta_{u_i} + \Delta_{u_j} + \lambda(\Lambda_{u_i} + \Lambda_{u_j}).$$

We next bound the terms Λ_{u_i} and Λ_{u_j}. From the Markov relation

$$s_{ji} \leftarrow X_j \leftrightarrow X_i \rightarrow u_{ij},$$

we have

$$I(u^*_{ij}; s^*_{ji}) \le I(u^*_{ij}; X_j), \tag{26}$$

and therefore

$$I(u^*_{ij}; s^*_{ji}) - I(u^*_{ij}; s^*_{ij}) \le I(u^*_{ij}; X_j) - I(u^*_{ij}; s^*_{ij}). \tag{27}$$

Hence

$$\begin{aligned}\Lambda_{u_i} &\le I(u^*_{ij}; X_j) - I(u^*_{ij}; s^*_{ij}) \\ &= I(u^*_{ij}; X_j \mid s^*_{ij}) - I(u^*_{ij}; s^*_{ij} \mid X_j) \\ &\le I(u^*_{ij}; X_j \mid s^*_{ij}) \\ &\le I(X_i; X_j \mid s^*_{ij}), \qquad \text{by } u_{ij} \leftarrow X_i \leftrightarrow X_j \\ &= I(X_j; X_i, s^*_{ij}) - I(X_j; s^*_{ij}) \\ &= I(X_j; X_i) - I(X_j; s^*_{ij}) \\ &\le \delta_c.\end{aligned} \tag{28}$$

The last inequality follows from the definition of the ideal shared optimization problem: since s^*_{ij} is an optimal solution of Eq. (8), it satisfies

$$I(X_j; X_i) - I(X_j; s^*_{ij}) \le \delta_c.$$

By the same argument, we also obtain

$$\Lambda_{u_j} \le \delta_c. \tag{29}$$

Furthermore, by Proposition 4 of Wang et al. [2025], the terms Δ_{u_i} and Δ_{u_j} are each upper bounded by δ_c. Therefore,

$$\mathcal{J}^o(\theta^*_{ij}) - \mathcal{J}(\theta^*_{ij}) \le 2\delta_c + 2\lambda\delta_c. \tag{30}$$

Combining Eqs. (25) and (30), we conclude that

$$\mathcal{J}(\hat{\theta}_{ij}) \ge \mathcal{J}^o(\theta^*_{ij}) - 2(1+\lambda)\delta_c.$$

proving near-optimality of the joint optimization objective with respect to the ideal step-by-step optimization problem. □

Stronger form of near-optimality When MNI is unattainable, we further state the following assumption:

Assumption 1. At the optimal solution $\hat{\theta}_{ij}$ of the joint optimizer, we define d to quantify the residual redundancy between the unique and shared components when MNI is unattainable, and for which it holds:

$$I(\hat{u}_{ij}; \hat{s}_{ij}) \leq d, \qquad I(\hat{u}_{ji}; \hat{s}_{ji}) \leq d.$$

Here, d is the disentanglement analogue of the coverage gap δ_c and is zero under attainable MNI (follows from Theorem 1).

Under Assumption 1, we can additionally derive Propositions 1 and 2.

Proposition 1. Let $\theta^*_{ij} = (u^*_{ij}, u^*_{ji}, s^*_{ij}, s^*_{ji})$ be an optimal solution of the step-by-step optimizer as stated in Eqs. (8) - (9), and $\hat{\theta}_{ij}$ the corresponding optimal solution of the joint optimization problem, as stated in Eqs. (3) - (5). When MNI is unattainable, under Assumption 1 it holds,

$$\mid \mathcal{J}(\hat{\theta}_{ij}) - \mathcal{J}(\theta^*_{ij}) \mid \leq 2(1+\lambda)\delta_c + 2\lambda d.$$

Proof. By definition of $\hat{\theta}_{ij}$, we have

$$\mathcal{J}(\hat{\theta}_{ij}) - \mathcal{J}(\theta^*_{ij}) \geq 0, \qquad \hat{\theta}_{ij} \in \arg\max_{\theta_{ij}} \mathcal{J}(\theta_{ij}).$$

Moreover, the difference can we rewritten as follows:

$$\mathcal{J}(\hat{\theta}_{ij}) - \mathcal{J}(\theta^*_{ij}) = \underbrace{[\mathcal{J}(\hat{\theta}_{ij}) - \mathcal{J}^o(\hat{\theta}_{ij})]}_{A} + \underbrace{[\mathcal{J}^o(\hat{\theta}_{ij}) - \mathcal{J}^o(\theta^*_{ij})]}_{B} + \underbrace{[\mathcal{J}^o(\theta^*_{ij}) - \mathcal{J}(\theta^*_{ij})]}_{C},$$

where,

$$\mathcal{J}^o(\theta_{ij}) = \alpha\mathcal{L}^o_{s_i}(\theta_{ij}) + \alpha\mathcal{L}^o_{s_j}(\theta_{ij}) + \mathcal{L}^o_{u_i}(\theta_{ij}) + \mathcal{L}^o_{u_j}(\theta_{ij}).$$

For A we have:

$$\mathcal{J}(\hat{\theta}_{ij}) - \mathcal{J}^o(\hat{\theta}_{ij}) = -\hat{\Delta}_{u_i} - \hat{\Delta}_{u_j} - \lambda\hat{\Lambda}_{u_i} - \lambda\hat{\Lambda}_{u_j},$$

where,

$$\hat{\Delta}_{u_i} := I(\hat{u}_{ij}, X_j; X_i) - I(\hat{u}_{ij}, \hat{s}_{ji}; X_i), \qquad \hat{\Lambda}_{u_i} := I(\hat{u}_{ij}; \hat{s}_{ji}) - I(\hat{u}_{ij}; \hat{s}_{ij}),$$

and similarly,

$$\hat{\Delta}_{u_j} := I(\hat{u}_{ji}, X_i; X_j) - I(\hat{u}_{ji}, \hat{s}_{ij}; X_j), \qquad \hat{\Lambda}_{u_j} := I(\hat{u}_{ji}; \hat{s}_{ij}) - I(\hat{u}_{ji}; \hat{s}_{ji}).$$

Notice that $\hat{\Delta}_{u_i}, \hat{\Delta}_{u_j}, \hat{\Lambda}_{u_i}, \hat{\Lambda}_{u_j}$ are the same expressions defined in Theorem 2 evaluated at $\hat{u}_{ij}, \hat{s}ij, \hat{u}_{ji}, \hat{s}_{ji}$. It holds,

$$I(\hat{u}_{ij}, \hat{s}_{ji}; X_i) \leq I(\hat{u}_{ij}, X_j; X_i) \implies -\hat{\Delta}_{u_i} \leq 0,$$

from data processing, since $X_j \to s_{ji}$. By exact symmetry, we obtain for $-\hat{\Delta}_{u_j} \leq 0$.

Moreover, we have,

$$I(\hat{u}_{ij}; \hat{s}_{ij}) - I(\hat{u}_{ij}; \hat{s}_{ji}) \leq I(\hat{u}_{ij}; \hat{s}_{ij}) \implies -\Lambda_{u_i} \leq d,$$

where the first inequality follows from the non-negativity of mutual information and the latter holds under Assumption 1. Similarly, we have under exact symmetry that $-\Lambda_{u_j} \leq d$.

Overall we obtain,

$$\mathcal{J}(\hat{\theta}_{ij}) - \mathcal{J}^o(\hat{\theta}_{ij}) \leq 2\lambda d.$$

For B we have:

$$\mathcal{J}^o(\hat{\theta}_{ij}) - \mathcal{J}^o(\theta^*_{ij}) \leq 0$$

since $\theta^*_{ij} \in \arg\max \mathcal{J}^o$.

For C it holds:

$$\mathcal{J}^o(\theta^*_{ij}) - \mathcal{J}(\theta^*_{ij}) \leq 2(1+\lambda)\delta_c$$

which follows from Theorem 2, Eq. (30). Thus we have,

$$\mid \mathcal{J}(\hat{\theta}_{ij}) - \mathcal{J}(\theta^*_{ij}) \mid \leq 2(1+\lambda)\delta_c + 2\lambda d.$$

which concludes the proof. □

Moreover, when MNI is unattainable we can also derive the following:

Proposition 2. Let $\theta^*_{ij} = (u^*_{ij}, u^*_{ji}, s^*_{ij}, s^*_{ji})$ be an optimal solution of the step-by-step optimizer as stated in Eqs. (8) - (9), and $\hat{\theta}_{ij}$ the corresponding optimal solution of the joint optimization problem, as stated in Eqs. (3) - (5). When MNI is unattainable, under Assumption 1 it holds,

$$-2(1+\lambda)\delta_c \leq \mathcal{J}(\hat{\theta}_{ij}) - \mathcal{J}^o(\theta^*_{ij}) \leq 2\lambda d,$$

where,

$$\mathcal{J}^o(\theta^*_{ij}) = \alpha\mathcal{L}^o_{s_i}(\theta^*_{ij}) + \alpha\mathcal{L}^o_{s_j}(\theta^*_{ij}) + \mathcal{L}^o_{u_i}(\theta^*_{ij}) + \mathcal{L}^o_{u_j}(\theta^*_{ij}).$$

Proof. By Theorem 2 it holds,

$$\mathcal{J}(\hat{\theta}_{ij}) - \mathcal{J}^o(\theta^*_{ij}) \geq -2(1+\lambda)\delta_c.$$

Moreover, the difference can be rewritten as follows:

$$\mathcal{J}(\hat{\theta}_{ij}) - \mathcal{J}^o(\theta^*_{ij}) = \underbrace{[\mathcal{J}(\hat{\theta}_{ij}) - \mathcal{J}^o(\hat{\theta}_{ij})]}_{A} + \underbrace{[\mathcal{J}^o(\hat{\theta}_{ij}) - \mathcal{J}^o(\theta^*_{ij})]}_{B},$$

and similarly to the Proposition 1 for the terms A and B, under Assumption 1 we obtain:

$$\mathcal{J}(\hat{\theta}_{ij}) - \mathcal{J}^o(\theta^*_{ij}) \leq 2\lambda d.$$

Thus, in total we have,

$$-2(1+\lambda)\delta_c \leq \mathcal{J}(\hat{\theta}_{ij}) - \mathcal{J}^o(\theta^*_{ij}) \leq 2\lambda d,$$

which concludes the proof. □

Overall, through Theorem 2 we show that for the maximization problem, the value of our optimizer $\mathcal{J}(\hat{\theta}_{ij})$ is lower bounded by $\mathcal{J}^o(\theta^*_{ij})$ minus a tractable constant $2(1+\lambda)\delta_c$ and under Assumption 1 and following Proposition 2 we show that the difference $\mathcal{J}(\hat{\theta}_{ij}) - \mathcal{J}^o(\theta^*_{ij})$ is also upper bounded by $2\lambda d$. Hence, under Assumption 1, $\mathcal{J}(\hat{\theta}_{ij})$ can exceed the ideal optimum by at most $2\lambda d$. The gap is governed by the constants δ_c, d and it closes completely under attainable MNI.

F Framework Implementation Details

F.1 Slot assignment in latent array $\mathcal{H}$

The flexibility of our framework permits any bijective mapping ϕ_i between the different rows of each $\mathcal{H}_i$ and the corresponding components. This mapping may vary across modalities, as each disentangling module operates independently. Nevertheless, for each modality, ϕ_i should cover exactly once all the expected pairwise interactions, excluding self-interactions, and remain consistent throughout both training and inference. For our implementation, we adopt the following definition:

Definition F.1. We define the bijective mapping ϕ_i as:

$$\phi_i : \mathcal{H}_i \to \{u_{ij}, s_{ij}\}_{j \neq i}, \quad \forall i \in \mathcal{M}.$$

Let k denote the row index of $\mathcal{H}_i$ corresponding to the k-th slot of $\mathcal{H}_i$, h_{ik}. For every modality j, we associate each h_{ik}, with either a unique component u_{ij} or a shared component s_{ij} involving modality i. We also set $j = \lceil k/2 \rceil$, as follows.

- If k is odd:

$$\phi_i(h_{ik}) = \begin{cases} u_{ij} & j < i \\ u_{i,j+1} & j > i. \end{cases}$$

- If k is even:

$$\phi_i(h_{ik}) = \begin{cases} s_{ij} & j < i \\ s_{i,j+1} & j \geq i. \end{cases}$$

This construction systematically indexes all shared and unique components involving modality i while excluding self-interactions.

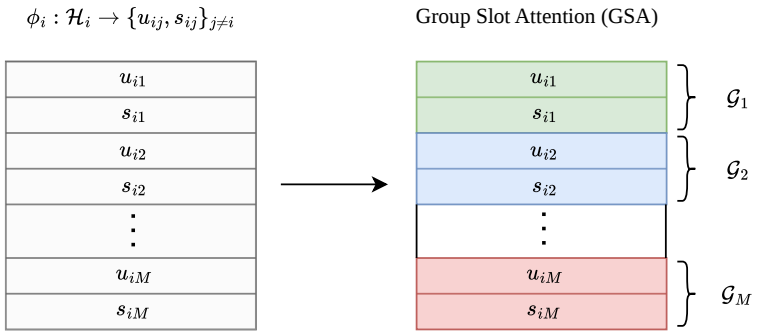


Figure 6: Illustration of the bijective mapping ϕ_i and Group Slot Attention. **Left:** Visualization of the mapping defined in Definition F.1, where each latent slot, h_{ik}, is assigned to a specific unique or shared component. **Right:** the grouping mechanism is observed, where each pair (u_{ij}, s_{ij}) belongs in a separate group $\mathcal{G}_j$.

F.2 Group slot attention

In grouped slot attention, slots are partitioned into disjoint groups, and competition is enforced only within each group rather than across all slots. In our setting, we select the group size equal to 2, resulting in exactly $M - 1$ groups, such that every group contains exactly two slots, e.g. the pair (u_{ij}, s_{ij}) corresponding to the unique and shared components for the modality pair (i, j). For each input token, attention scores are normalized across the two slots in the group, so that the two slots compete to explain that particular token. As a result, each token distributes its information only between the paired slots, encouraging them to specialize into complementary roles while remaining independent of other groups.

Formally, for each cross-attention mechanism at iteration t, we compute

$$Q = \tilde{\mathcal{H}}_i^t W_Q^T \in \mathbb{R}^{N \times D_h}, \qquad K = Z_i W_K^T \in \mathbb{R}^{D_h \times D}, \qquad V = Z_i W_V^T \in \mathbb{R}^{D_h \times D},$$

representing the query, key and value respectively, where $W_Q \in \mathbb{R}^{D_h \times D_h}$, and $W_K, W_v \in \mathbb{R}^{T \times D_h}$ are learnable parameter matrices. Given the mapping defined in F.1, the queries are arranged into contiguous pairs of size two. We compute the standard query-key similarities, but normalize the attention only over slots within the same group. If $\mathcal{G}_m(n)$ denotes the m-th group containing slot n, then

$$\mathbf{A}_{n,m} = \text{softmax}_{n' \in \mathcal{G}_m(n)} \frac{Q_{n'} K_m^T}{\sqrt{D_h}}.$$

Afterwards, we re-normalize across the token dimension,

$$\hat{\mathbf{A}}_{n,m} = \frac{\mathbf{A}_{n,m}}{\sum_{j=1}^{T} \mathbf{A}_{n,j} + \epsilon}, \qquad \epsilon > 0.$$

Lastly, slot representations are updated via the standard attention readout

$$\tilde{\mathcal{H}}_i^{t+1} = \hat{\mathbf{A}} V.$$

Both the bijective mapping as well as the Group Slot Attention are illustrated in Figure 6 for additional intuition.

F.3 Training objectives

During training, in practice we optimize the losses in Eqs. (6), (7). We model the $s_{ij} \sim p(\cdot \mid X_i)$ and $s_{ji} \sim p(\cdot \mid X_j)$ as von Mises-Fisher (vMF) distributions, i.e. $s_{ij} \sim \text{vMF}(\mu(X_i), \kappa)$ and $s_{ji} \sim \text{vMF}(\mu(X_j), \kappa)$. The vMF distribution represents data on a hypersphere, where μ is the mean direction and κ controls the concentration around that direction. For $I(s_{ij}; X_j)$ we use the InfoNCE objective [van den Oord et al., 2018] as follows:

$$\mathcal{L}_{s_i}^{\text{INCE}} = \mathbb{E}_{s_{ij},\, s_{ji}^+,\, \{s_{ji,k}^-\}_{k=1}^N} \left[-\log \frac{\exp\left(s_{ij}^\top s_{ji}^+ / \tau\right)}{\exp\left(s_{ij}^\top s_{ji}^+ / \tau\right) + \sum_{k=1}^N \exp\left(s_{ij}^\top s_{ji,k}^- / \tau\right)} \right],$$

where τ is the temperature parameter, s_{ji}^+ is the positive anchor, $s_{ji,k}^-$ the negatives and N is the total number of negative samples. A similar formulation follows by symmetry for the case of $\mathcal{L}_{s_{ji}}^{INCE}$. For the KL-divergence terms, as we have modeled the representations through vMF, the quantity $\mathcal{L}_{KL}$ obtains the exact form:

$$\mathcal{L}_{KL} = \mathbb{E}_{x_i, x_j} \left[\mu(X_i)^T \mu(X_j) \right]$$

and accordingly for the case of s_{ji}. Analogously, for $\mathcal{L}_{u_i}^{\text{INCE}}$, we employ an InfoNCE objective using augmented views of X_i and X_j. Specifically, we concatenate the unique representation u_{ij} with the shared representation s_{ji},

$$\tilde{z} = u_{ij} \oplus s_{ji},$$

and contrast it against the representation $\tilde{z}_{\text{a}}^+$ obtained from the augmented views of the same sample:

$$\mathcal{L}_{u_i}^{\text{INCE}} = \mathbb{E}_{\tilde{z},\, \tilde{z}_{\text{a}}^+,\, \{\tilde{z}_k^-\}_{k=1}^N} \left[-\log \frac{\exp\left(\tilde{z}^\top \tilde{z}_{\text{a}}^+ / \tau\right)}{\exp\left(\tilde{z}^\top \tilde{z}_{\text{a}}^+ / \tau\right) + \sum_{k=1}^N \exp\left(\tilde{z}^\top \tilde{z}_k^- / \tau\right)} \right].$$

Here, N denotes the number of negative samples, $\tilde{z}_{\text{a}}^+$ is the positive counterpart obtained from the augmented views of the same sample, and $\tilde{z}_k^-$ denotes the k-th negative sample. Note that as discussed by Tian et al. [2020], the augmentation design within this setting is an important practical consideration and should be carefully chosen, so that all the desired information components are preserved within the augmented views. Lastly, the leakage between the unique and shared components is regularized through a cross-covariance orthogonality penalty. Concretely, let $\bar{U}_{ij} \in \mathbb{R}^{B \times D}$ and $\bar{S}_{ji} \in \mathbb{R}^{B \times D}$ denote the row-wise ℓ_2-normalized and feature-wise standardized batch representations of u_{ij} and s_{ji}, respectively. We define

$$\mathcal{L}_{\text{xcov}}(u_{ij}, s_{ji}) = \frac{1}{D} \left\| \frac{1}{B} \bar{U}_{ij}^\top \bar{S}_{ji} \right\|_F,$$

where B is the number of samples, D is the representation dimension, and $\|\cdot\|_F$ denotes the Frobenius norm.

G Experimental Details

G.1 Synthetic data creation

For each atomic representation as stated in Definition 2.1, we begin by sampling N instances for all $2^M - 1$ latent factors independently from a normal distribution, i.e. $z_A \sim \mathcal{N}(0, \sigma^2 I_{d_\ell})$. We then concatenate the sampled factors on a per-modality basis, according to Definition 2.2, to construct Z_i for all $i \in \mathcal{M}$. To avoid introducing structural biases related to dimensionality or distribution, we use the same σ and d_ℓ for all sampled atomic representations. In addition, since we conduct experiments across varying numbers of modalities, we set $d_\ell = \frac{E}{2^{M-1}}$, so that $E = \dim Z_i$ remains constant regardless of M.

Having obtained $Z_i \in \mathbb{R}^E$, we next augment this vector representation to mimic sequence-style embeddings, yielding $Z_i \in \mathbb{R}^{S \times E}$, as in practice foundation models typically produce patch- or token-level representations. To this end, for each latent factor we additionally sample a base transform $U_A \sim \mathcal{N}(0, \sigma_U^2 I_{d_\ell})$, together with a small sequence-specific variation $\Delta_{A,s} \sim \mathcal{N}(0, \sigma_\Delta^2 I)$, where

$\sigma_\Delta \ll \sigma_U$. We then place these transformation matrices in block-diagonal form over all $s \in S$, defining $U_{i,s} = \text{blockdiag}(U_A + \Delta_{A,s})$ for all $i \in \mathcal{M}$.

Moreover, we sample binary masks $\mathbf{M}_{i,s} = [m_1^s, \ldots, m_{2^{M-1}}^s]$, with $m_j^s \in \{0,1\}$, so that not all atomic representations are accessible at every sequence position. Intuitively, this mimics the fact that, for example, not all image patches contain all information components of an image. As a result, $\mathbf{M}_i \in \{0,1\}^{S \times 2^{M-1}}$. We apply this mask pointwise to the constructed transformations, yielding $\tilde{U}_{i,s} = \mathbf{M}_{i,s} \odot U_{i,s}$. Finally, we define modality-specific rotations R_i for all $i \in \mathcal{M}$, and set $W_{i,s} = R_i \tilde{U}_{i,s}$. Each modality's final representation is then given by $\mathbf{Z}_{i,s} = \tanh(\gamma W_{i,s} Z_i)$.

Algorithm 1 Synthetic generation of multimodal factorized representations

Require: number of modalities M, sequence length S, embedding dimension E, variance parameters $\sigma, \sigma_U, \sigma_\Delta$, scaling factor γ

Let $\mathcal{A}$ denote the set of all non-empty modality subsets, $|\mathcal{A}| = 2^M - 1$

Set latent dimension per atomic factor:

$$d_\ell = \frac{E}{2^{M-1}}$$

for each atomic factor $A \in \mathcal{A}$ **do**

  Sample atomic latent representation:

$$z_A \sim \mathcal{N}(0, \sigma^2 I_{d_\ell})$$

end for

for each modality $i \in \mathcal{M}$ **do**

  Construct modality representation by concatenating relevant atomic factors:

$$Z_i = \text{concat}\{z_A \mid A \ni i\}, \quad Z_i \in \mathbb{R}^E$$

  for each sequence position $s = 1, \ldots, S$ **do**

    for each atomic factor $A \ni i$ **do**

      Sample base transform:

$$U_A \sim \mathcal{N}(0, \sigma_U^2 I_{d_\ell})$$

      Sample small sequence-specific variation:

$$\Delta_{A,s} \sim \mathcal{N}(0, \sigma_\Delta^2 I_{d_\ell}), \quad \sigma_\Delta \ll \sigma_U$$

    end for

    Construct block-diagonal transformation:

$$U_{i,s} = \text{blockdiag}(U_A + \Delta_{A,s})$$

    Sample binary accessibility mask:

$$\mathbf{M}_{i,s} \in \{0,1\}^{2^{M-1}}$$

    Apply mask:

$$\tilde{U}_{i,s} = \mathbf{M}_{i,s} \odot U_{i,s}$$

    Apply modality-specific rotation:

$$W_{i,s} = R_i \tilde{U}_{i,s}$$

    Generate sequence embedding:

$$Z_{i,s} = \tanh(\gamma W_{i,s} Z_i)$$

  end for

end for

return $\{Z_i\}_{i \in \mathcal{M}}$, where $Z_i \in \mathbb{R}^{S \times E}$

Evaluation details We evaluate disentanglement performance by measuring whether each learned component $(\hat{\mathbf{u}}_{ij}, \hat{\mathbf{s}}_{ij})$ predicts the label associated with its corresponding ground-truth component $(\mathbf{u}_{ij}, \mathbf{s}_{ij})$, while remaining uninformative about labels associated with other cross components. To assess this behavior, we report confusion matrices for each model across all values of M.

For a more holistic metric, we validate each model's performance based on its deviation from the ideal performance using Δ_{model}, which is defined as follows:

$$\Delta_{\text{model}} = \frac{1}{4}\Big(\underbrace{|A_{\hat{\mathbf{u}}_{ij}\to y_{u_{ij}}} - 100| + |A_{\hat{\mathbf{s}}_{ij}\to y_{s_{ij}}} - 100|}_{\text{intra-component prediction}} + \underbrace{|A_{\hat{\mathbf{u}}_{ij}\to y_{s_{ij}}} - 50| + |A_{\hat{\mathbf{s}}_{ij}\to y_{u_{ij}}} - 50|}_{\text{cross-component leakage}}\Big),$$

where $A_{\mathbf{u}_{ij}\to y_{u_{ij}}} := \mathbb{E}[Acc(\mathbf{u}_{ij} \to y_{u_{ij}})]$.

Training and Hyperparameter specifications For the experiments, we test for the number of modalities $M = \{2, 3, 4, 5\}$. For every case, we generate $N = 150.000$ training samples per modality and we set $E = 64$ and $S = 8$. We also choose $\sigma_U^2 = 0.5$ and $\sigma_\Delta = 0.005 \ll \sigma_U$. For the scaling factor of the non-linearity we, select $\gamma = 0.3$.

The architectural specifications of each model are demonstrated in Table 7 (a). In order to perform a fair scaling experiment, as M grows, the encoder backbone of each disentanglement module for RePercENT, changes only in terms of the number of the required latent slots in $\mathcal{H}$. Specifically, there are 2 dedicated slots for the case of $M = 2$, 4 for the case of $M = 3$ etc. For the baselines, the capacity per separate encoder remains the same.

For the training specifications, see Table 7 (b). For every M the generated dataset is split into training, validation, and test sets with a ratio of $0.7/0.1/0.2$, respectively. For the disentanglement loss, we keep α and β fixed throughout training. The coefficient λ is annealed using an exponential scheduler and updated once per training iteration. Lastly, each model is trained for 150 epochs to ensure convergence. We select the best checkpoint based on the lowest validation loss observed during training and report its performance on the test set.

All models are trained on three different train/test splits, using two random seeds for each split. We aggregate the results across all runs and report the average detection accuracy as mean $\pm$ standard deviation.

Augmentations During training, we augment the synthetic inputs using a simple stochastic augmentation module. Given an input tensor Z, the module returns an augmented tensor Z_{a} by applying one of the following transformations: Gaussian noise injection with scale 10^{-3}, feature swapping, or random feature dropping. For random feature dropping, entries are masked according to a drop scale of 10.

G.2 MultiBench overview

MultiBench [Liang et al., 2021] is a large-scale benchmark for multimodal representation learning that unifies 15 datasets spanning 10 modalities, 20 prediction tasks, and 6 research areas under standardized data loading, preprocessing, and evaluation protocols. We use three of its four affective computing datasets, all of which share the `Text`, `Video`, and `Audio` modalities, allowing us to evaluate both two- and three-modality settings. MOSI contains 2,199 opinion video segments from YouTube vlogs, annotated with sentiment intensity in $[-3, 3]$. MOSEI is its larger counterpart, with 22,777 segments annotated for both sentiment and emotion. UR-FUNNY contains 16,514 segments from TED talks, labeled for the presence of humor. We formulate all three tasks as binary classification. In each task, the relevant cues are distributed across what is said, how it is said, and the speaker's facial expressions, so the three modalities carry complementary information, which makes these datasets well suited to our setting. All three datasets are provided as pre-extracted feature sequences, such as GloVe word embeddings for text, Facet facial features for video, and COVAREP acoustic features for audio. We use the same train/validation/test splits as Liang et al. [2023].

Training and hyperparameter specifications. For all baselines (FactorCL, FOCAL, DisentangledSSL, and JointOpt), we follow the authors' implementations and use the same 5-layer Transformer encoder as Liang et al. [2023]. In contrast, RePercENT operates directly on sequences of embeddings and therefore requires no backbone encoder. All models use the same representation dimensionality

Table 7: Architecture and training specifications used for the synthetic dataset experiments. For each model, table (a) reports the single encoder parameters, which remain fixed across the different values of M. The training specifications are the same across all models.

(a) Model architecture

Model $\sim$ Hyperparameter	Value
RePercENT	
Embedding input dim	64
Embedding sequence length	8
Depth in $\mathcal{D}_i$	3
Cross-attention heads	2
Transformer latent heads	2
Latent dim D_l	32
gMLP	
Embedding input dim	64
Embedding sequence length	8
Feed-forward hidden dim	64
Projection head dim	32
Number of layers	2
Activation	GELU
GRU	
Embedding input dim	64
Hidden dim	64
Latent dim	32
Number of layers	1
Activation	ReLU
MLP	
Input dim	512
Hidden dim	64
Output dim	32
Activation	ReLU

(b) Training specification

Component	Value
Disentanglement loss	
α	2.0
β	0.1
λ start value	0.1
λ end value	1.0
λ annealing iterations	8000
λ start iteration	800
Optimizer	
Optimizer	Adam
Learning rate	8×10^{-4}
Weight decay	10^{-4}
Training	
Number of epochs	150
Batch size	1024

of 300 across the three datasets, and model selection is based on validation performance. Table 8 reports the training hyperparameters for RePercENT, which we tuned on the validation set. For the baselines, we adopt the loss-specific hyperparameters reported by their authors, while using the same optimizer and training schedule as RePercENT.

Table 8: Architecture and training specifications used for RePercENT on the MultiBench sentiment and humor detection.

(a) Model architecture

	Encoder		
Hyperparameter	`Image`	`Text`	`Audio`
RePercENT			
Embedding input dim	20 / 35 / 371[†]	300	5 / 74 / 81[†]
Embedding sequence length	50	50	50
Depth in $\mathcal{D}_i$	3	3	3
Cross-attention heads	2	2	2
Transformer latent heads	2	2	2
Latent dim D_l	300	300	300

[†] MOSI / MOSEI / UR-FUNNY, respectively.

(b) Training specification

Component	Value
Disentanglement loss	
α	1.0
β	1.0
λ value	0.001
Optimizer	
Optimizer	Adam
Learning rate	1×10^{-4}
Weight decay	10^{-5}
Training	
Number of epochs	100
Batch size	512

G.3 IRFL overview

Dataset creation The Image Recognition of Figurative Language (IRFL) dataset [Yosef et al., 2023]. IRFL is a multimodal benchmark that comprises approximately 6,690 instances designed to test the comprehension of figurative language, including idioms, metaphors, and similes. Each text phrase is paired with images categorized by their relationship to the text. The latter falls in one of the categories: figurative, literal, partial literal, or figurative+literal. We extract the complete dataset including idioms, metaphors, and similes, and keep only the rows that correspond to figurative images and text. For instances where no definition is provided, we manually complete the missing annotations. Afterwards, each image is encoded using the OpenAI CLIP-ViT-B/32 vision transformer, while captions and definitions are encoded using the corresponding CLIP text transformer. We preserve the sequence-like structure of both modalities rather than using a single pooled embedding. For images, we retain the patch-level embeddings before the final projection layer. For text, we retain the contextual token embeddings from the final transformer layer instead of using only the pooled end-of-text (EOT) representation. We split the data into 2,594 training and 786 test samples, ensuring that no image or caption is shared across the two sets.

Detection task overview The figurative detection task evaluates Vision and Language Pre-Trained Models' (VL-PTMs) ability to choose the image that best visualizes the meaning of a figurative expression. Figure 7 provides two examples of the task, where even though there is shared information with most of the candidate images, only one describes best the figurative phrase. It is a very interesting task, as state-of-the-art VL models, such as CLIP-VIT-B/32, CLIP-RN50x64 [Radford et al., 2021], BLIP [Li et al., 2022] performed substantially worse than humans.



Figure 7: Adapted from Yosef et al. [2023]. Examples of the multimodal figurative language detection task for idiom, metaphor, and simile. The input is a figurative phrase and four candidate images (for idiom, we also show the definition). The correct answer is marked with an orange square.

We evaluate all models in a zero-shot manner. Let z_{text} denote the text query representation, derived from either the `Caption` alone or the concatenated `Caption` $\oplus$ `Definition`. For a given query, we compute the pairwise similarity between z_{text} and the target figurative image representation z_{image}, as well as the representations of the three distractor images $z_{\text{dist},i}$. A successful prediction requires the target similarity to strictly exceed all distractor similarities:

$$\text{sim}(z_{\text{text}}, z_{\text{image}}) > \max_{i \in \{1,2,3\}} \text{sim}(z_{\text{text}}, z_{\text{dist},i}).$$

Training and Hyperparameter specifications For the alignment approach, we fine-tuned OpenAI CLIP ViT-B/32 in two settings: projection-only and end-to-end. In the projection-only setting, the CLIP image and text encoders remain frozen, and only the final visual and text projection layers were trained using a learning rate of 2×10^{-4}. In the end-to-end setting, all CLIP parameters are trainable and fine-tuned using a learning rate of 10^{-5} to avoid overfitting. Training in both cases uses

the standard symmetric CLIP contrastive loss, AdamW optimization, cosine learning-rate decay with 10% warmup, and gradient clipping.

Similarly to the synthetic setup, all disentanglement baselines are trained under the same training protocol as RePercENT. The separate encoder architectures as well as the training parameters are reported in Table 9. Note that the two text modalities, `Caption` and `Definition`, share the same encoder architecture but are encoded by separate encoder instances with independent parameters.

All models are trained with five independent random seeds, and we report detection accuracy as the mean $\pm$ standard deviation across runs.

Table 9: Architecture and training specifications used of the disentanglement models for the IRFL Detection task experiments.

(a) Model architecture

Model $\sim$ Hyperparameter	`Image` **Encoder**	`Text` **Encoder**
RePercENT		
Embedding input dim	768	512
Embedding sequence length	49	77
Depth in $\mathcal{D}_i$	2	2
Cross-attention heads	2	2
Transformer latent heads	1	1
Latent dim D_l	256	256
gMLP		
Embedding input dim	768	512
Embedding sequence length	49	77
Feed-forward hidden dim	352	352
Projection head dim	256	256
Number of layers	2	2
Activation	GELU	GELU
GRU		
Embedding input dim	768	512
Hidden dim	256	256
Latent dim	256	256
Number of layers	1	1
Activation	ReLU	ReLU

(b) Training specification

Component	**Value**
Disentanglement loss	
α	6.0
β	2.0
λ value	4.0
Optimizer	
Optimizer	Adam
Learning rate	8×10^{-4}
Weight decay	10^{-4}
Training	
Number of epochs	40
Batch size	256

Augmentations To facilitate training, and avoid the inference of the CLIP model, we additionally pre-compute the augmentations used for training. Specifically, we augment training images using a fixed set of visually conservative transformations: horizontal flip, vertical flip, Gaussian blur, Gaussian noise, and posterization. Flip-based augmentations are combined with a mild random resized crop using the scale range $[0.95, 1.0]$ and bicubic interpolation to avoid removing the main object. Horizontal and vertical flips are applied with probability 0.95. Gaussian blur uses a kernel size of 27, Gaussian noise uses mean 0 and standard deviation 0.1 with clipping, and posterization uses 3 bits with probability 0.95. We do not apply hue or saturation jitter, since color information can be essential for figurative language. Figure 8 visualizes the result of two augmentations using Gaussian noise and posterization. For the text augmentations, we only add minor formatting and neutral variation in order to fully preserve the context. During training, for each image or text, we sample one of the pre-computed embedding representations of their corresponding augmented views.

G.4 TCGA dataset extraction

We utilize a pre-processed version of The Cancer Genome Atlas (TCGA) multimodal oncology cohort [Weinstein et al., 2013] by Tripathi et al. [2025] containing in total 10.857 patient records. For our experiments, we consider the four modalities and FM embeddings reported in Table 10. We group for each modality all the embeddings on a patient level, and preserve only the patients that have all four modalities available. Lastly, guided by the cancer-type distribution shown in Figure 9,

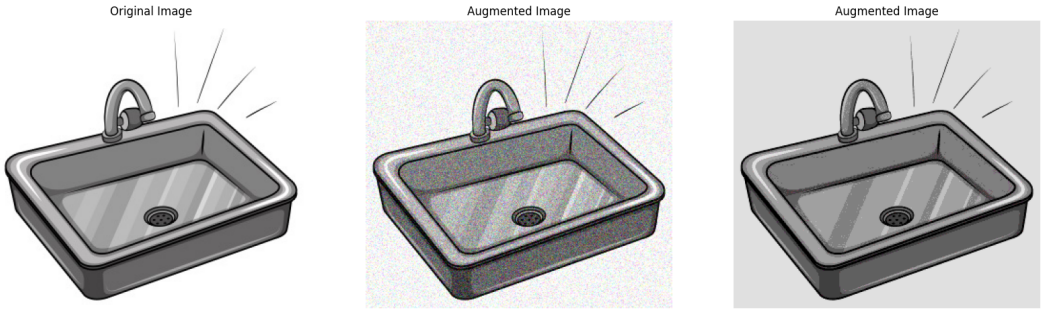


Figure 8: Example of two image augmentation variants used for the figurative language task. The original image from the IRFL dataset is depicted on the left and two additional augmented images are generated.

we restrict our analysis to cancer types with sufficient sample representation, and retain the 10 most prevalent cancer types to ensure reliable evaluation. This results in a final set of 5,732 patients. The selected cancer types, together with brief descriptions of their clinical relevance, are summarized in Table 11.

Table 10: HONeYBEE TCGA modality embeddings used in the oncology experiments.

Modality	Dim.	Encoder	Description
`Clinical`	1024	Qwen3 [Yang et al., 2025]	Patient-level clinical information, including structured and unstructured records such as demographics, laboratory values, medications, and clinical narratives.
`Pathology`	1024	Qwen3 [Yang et al., 2025]	Free-text pathology reports describing diagnostic and histopathological findings, processed as clinical text.
`WSI`	1024	UNI [Chen et al., 2024]	Whole-slide histopathology images processed through tissue detection, stain normalization, patch extraction, and patch-level feature extraction.
`Molecular`	48	SeNMo [Waqas et al., 2024]	Multi-omics molecular profiles, including gene expression, DNA methylation, somatic mutations, miRNA, and protein expression.

Training and Hyperparameter specifications We use all four modalities during training and extract both unique and shared representations for every modality pair. The specifications of the four RePercENT encoders are summarized in Table 12 (a), while the corresponding training details are provided in Table 12 (b). After training, for each extracted representation $\mathbf{u}_{ij}$ and $\mathbf{s}_{ij}$, as well as their concatenation, which together constitute the decomposition of modality i with respect to modality j, we train linear probes and report cancer-type classification accuracy. Finally, we compare against linear probes trained on the original HONeYBEE embeddings, which represent the full modality-specific representations proposed by Tripathi et al. [2025].

Table 11: Selected TCGA cancer types utilized in the Honeybee experimental evaluation.

TCGA Code	Cancer Type Description
GBM	Glioblastoma multiforme
LGG	Brain Lower Grade Glioma
HNSC	Head and Neck squamous cell carcinoma
LUAD	Lung adenocarcinoma
LUSC	Lung squamous cell carcinoma
BRCA	Breast invasive carcinoma
OV	Ovarian serous cystadenocarcinoma
PRAD	Prostate adenocarcinoma
COAD	Colon adenocarcinoma
KIRC	Kidney renal clear cell carcinoma

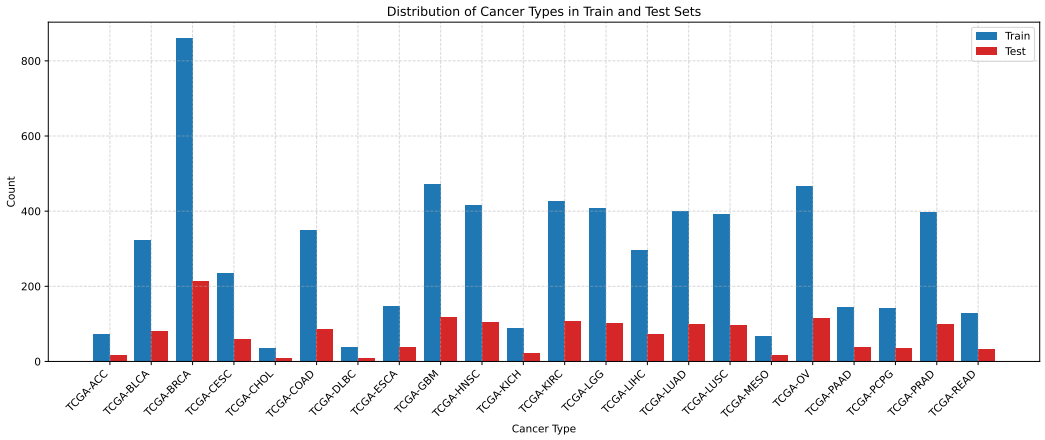


Figure 9: Cancer type patient count of the HONeYBEE extracted embeddings.

Table 12: Architecture and training specifications used for the disentanglement models in the TCGA cohort experiments.

(a) Model architecture

Hyperparameters	`Clinical` **Encoder**	`Pathology` **Encoder**	`WSI` **Encoder**	`Molecular` **Encoder**
RePercENT				
Embedding input dim	1024	1024	1024	48
Input sequence length	1	3	16	4
Depth in $\mathcal{D}_i$	8	8	8	8
Cross-attention heads	2	2	2	2
Transformer latent heads	1	1	1	1
Latent dimension D_l	256	256	256	256

(b) Training specification

Component	**Value**
Disentanglement loss	
α	5.0
β	0.5
λ start value	1.0
λ end value	3.0
λ annealing iterations	600
λ start iteration	200
Optimizer	
Optimizer	Adam
Learning rate	2×10^{-5}
Weight decay	5×10^{-4}
Training	
Number of epochs	50
Batch size	256

Missing modality robustness setup For this experiment, we simulate missingness for the `WSI-Molecular` modality pair. At test time, we keep the `Molecular` data fully available and randomly drop WSI at varying missingness rates.

For RePercENT, we train separate linear probes on the decomposition of WSI w.r.t. Molecular, i.e., $D_{\text{wsi,mol}}$ and the extracted decomposition of Molecular w.r.t. WSI, i.e., $D_{\text{mol,wsi}}$. At test time, we employ the simplest form of late fusion, where the predictions from the available modalities are combined by averaging their classifier scores.

We compare against the following three baselines trained on the original HONeYBEE modality embeddings:

- **Late fusion + averaging:** We train separate linear probes on the raw WSI and molecular embeddings, and average the prediction scores of the modalities available at test time.
- **Early fusion + mean imputation:** We train a single probe on the concatenated embeddings of the two modalities. During evaluation, we replace the WSI missing instances, with the corresponding train-set modality mean.
- **Early fusion + modality dropout:** We use the same masked representation, but in this case the training set is augmented with randomly dropped modalities to expose the classifier to missing-modality patterns during training

Each method is applied both to the original HONeYBEE embeddings and to the representations of a matched-capacity baseline we term **ClipSur**, which controls for the additional capacity and training that RePercENT introduces. ClipSur consists of two 6-layer Transformer encoders for the `Mol-WSI` pair, with a combined parameter count within 10% of RePercENT's, trained in a self-supervised manner with the standard CLIP contrastive loss [Radford et al., 2021]. All methods use class-balanced logistic-regression probes. We report results across 10 uniformly spaced `WSI` availability levels, ranging from fully available (100%) to fully missing (0%). At each availability level, performance is averaged over 10 independently sampled stochastic missingness masks.

G.5 Computational Resources

All experiments were run on a single NVIDIA H200 GPU with 140GB of GPU memory, 96 physical CPU cores, 192 logical threads, 1.5TB of RAM, and 7.2TB of disk storage. The software environment used Python 3.10.12. The most computationally demanding setting was the synthetic five-modality experiment, which required approximately 3 hours under the training regime described in Table 7. Experiments on the real-world datasets were substantially cheaper, with all runs completing in under 1 hour.

H Additional Experiments

H.1 Synthetic Dataset

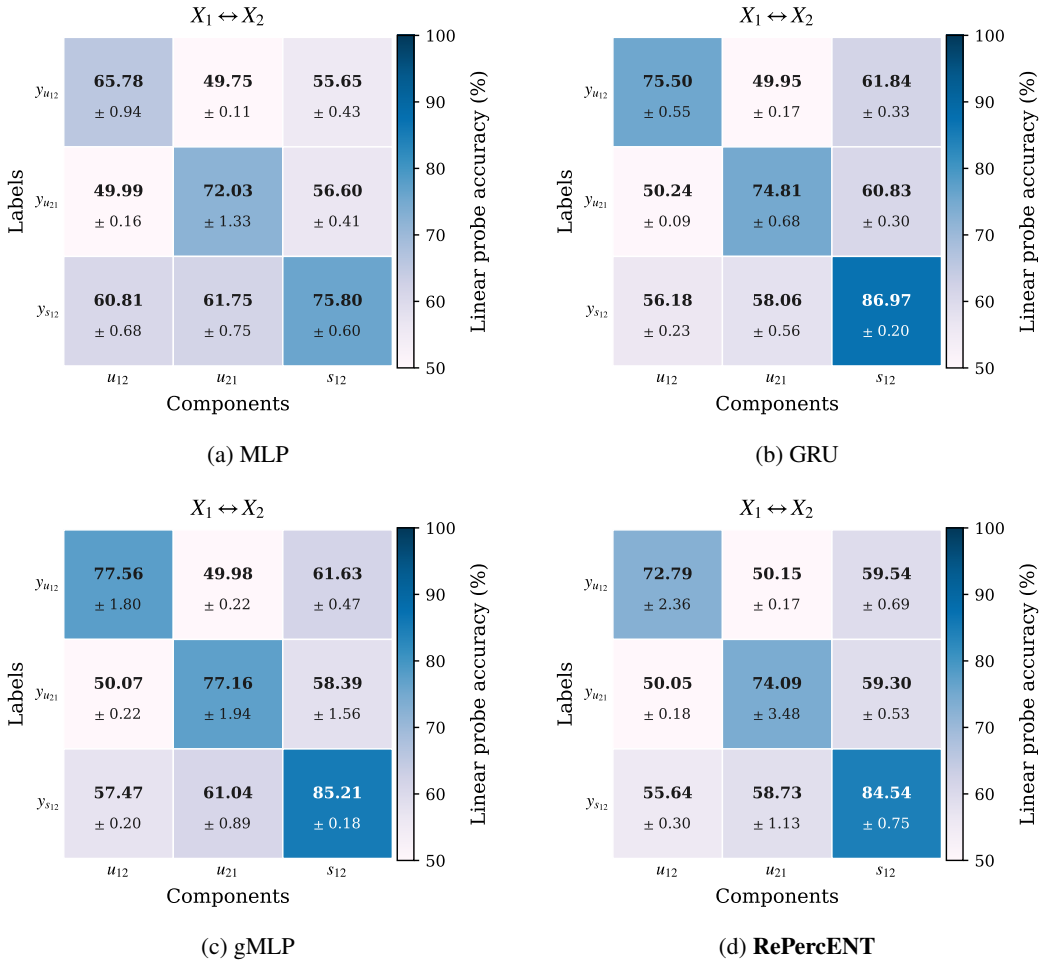
In this section, we provide detailed linear probe results for all the baselines on the synthetic dataset, for each value of $M \in \{2, 3, 4, 5\}$.

For the case of $M = 2$ (Figure 10), most architectures demonstrate a baseline ability to isolate components, characterized by high intra-component prediction and minimal cross-component leakage. However, architectural behavior diverges significantly as complexity increases.

MLP and GRU The MLP exhibits the weakest specialization, with significantly lower diagonal accuracy even in the two-modality regime. Conversely, while the GRU improves the recovery of intended factors—particularly shared representations—its stronger diagonal performance is often offset by elevated off-diagonal values, indicating persistent entanglement. As M increases (Figures 12a, 14b), the GRU undergoes a fundamental component collapse, successfully recovering only the shared factors, while the unique components lose all discriminative power.

gMLP and RePercENT These models demonstrate the most robust disentanglement performance as M grows. Both consistently achieve the disentanglement objective where baselines fail, maintaining high diagonal dominance and near-chance off-diagonal structures well beyond the two-modality setting.

Scalability of RePercENT Notably, RePercENT achieves performance parity with the gMLP while strictly adhering to the proposed pairwise slot-based design. The preservation of a clean, diagonal-dominant structure across all modality pairs demonstrates that RePercENT successfully overcomes the scalability bottleneck, maintaining granular disentanglement in high-dimensional multimodal settings. Moreover, our results indicate that performance at $M = 2$ is an unreliable predictor of architectural robustness in higher-dimensional spaces. As M increases, the information-theoretic complexity of interactions grows exponentially, introducing "interference" and potential "component collapse" risks that are absent in simpler settings.



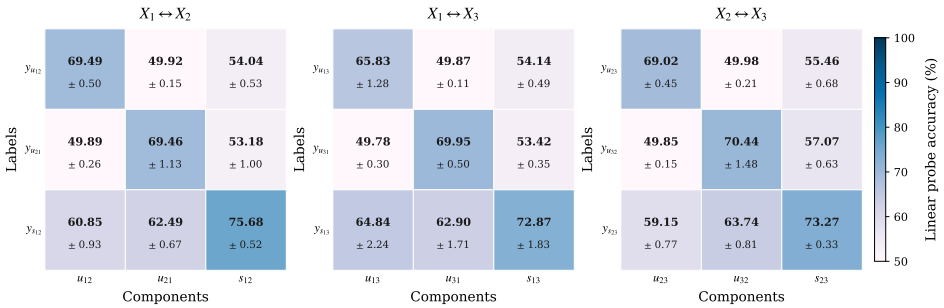
(a) MLP

(b) GRU

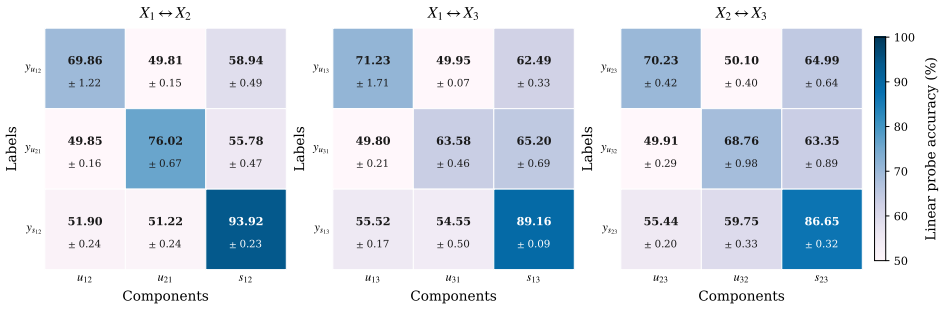
(c) gMLP

(d) **RePercENT**

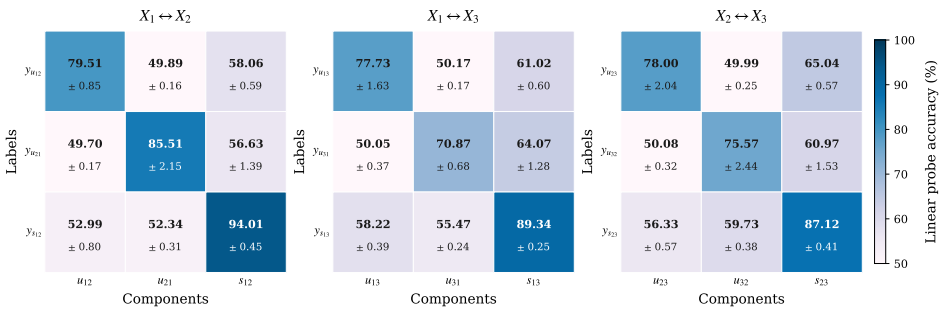
Figure 10: ($\mathbf{M = 2}$) Pairwise confusion matrices for the synthetic setting with two modalities, shown for MLP, GRU, gMLP, and **RePercENT**. While all models largely separate unique and shared information, sequence-aware models yield stronger intra-component accuracy, as reflected by higher main-diagonal values.



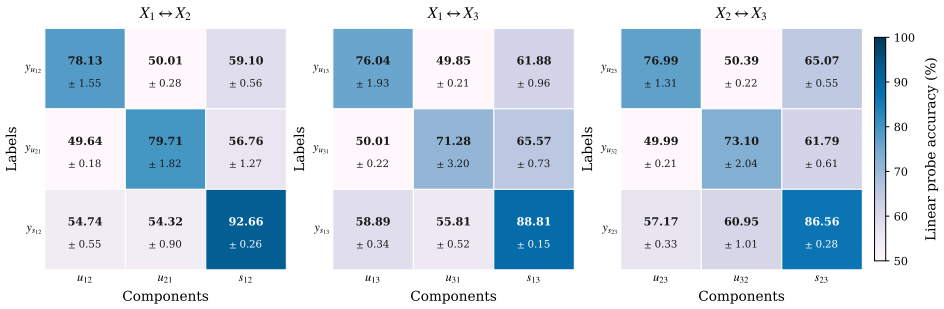
(a) MLP



(b) GRU

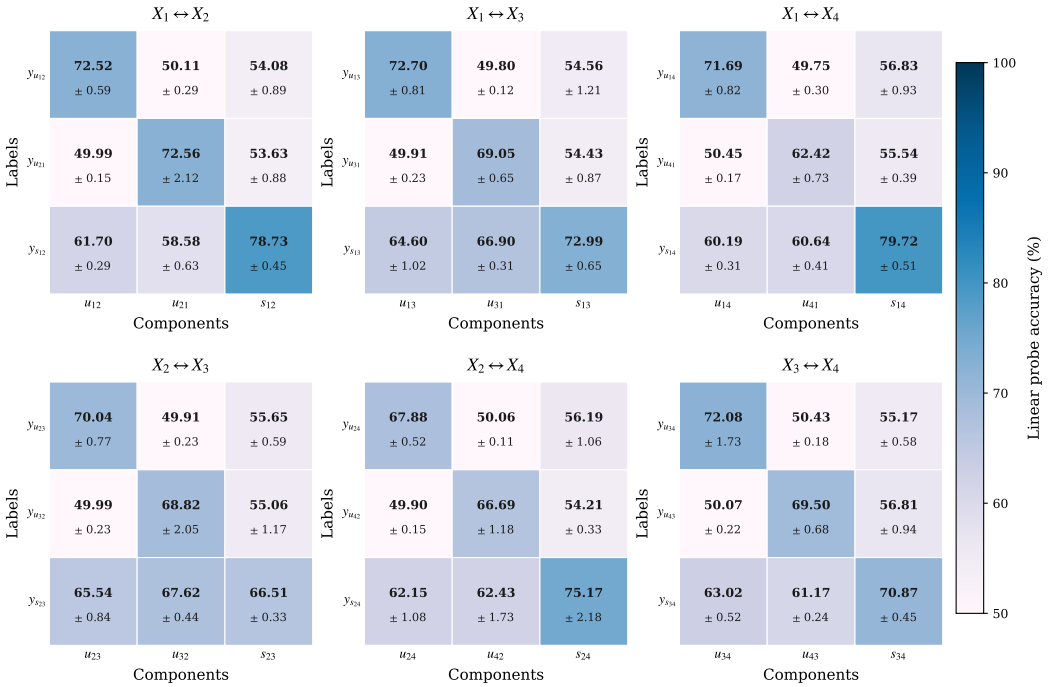


(c) gMLP

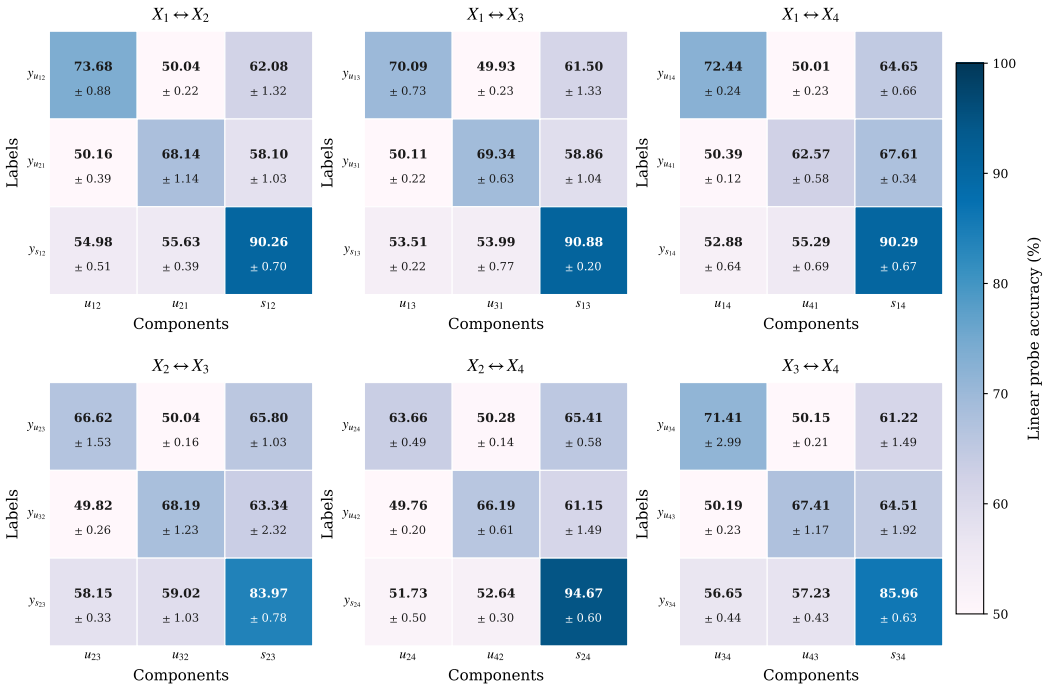


(d) **RePercENT**

Figure 11: ($\mathbf{M = 3}$) Pairwise confusion matrices for the synthetic setting with three modalities, shown for MLP, GRU, gMLP, and **RePercENT**. The performance of GRU and especially MLP degrades, while the gMLP and RePercENT present robust disentanglement performance.

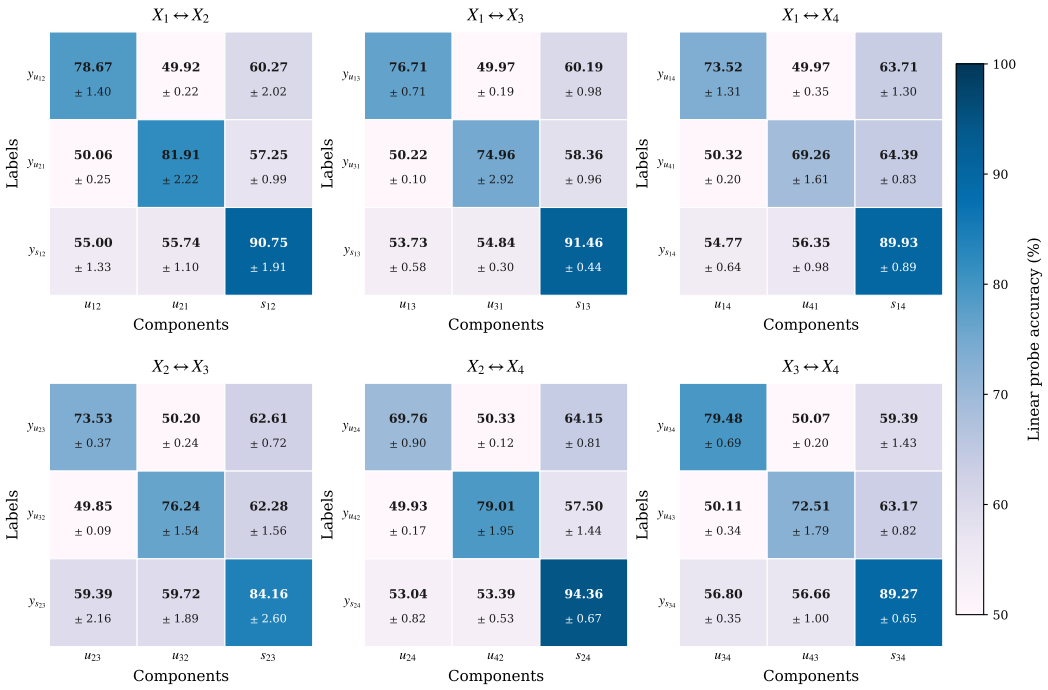


(a) MLP

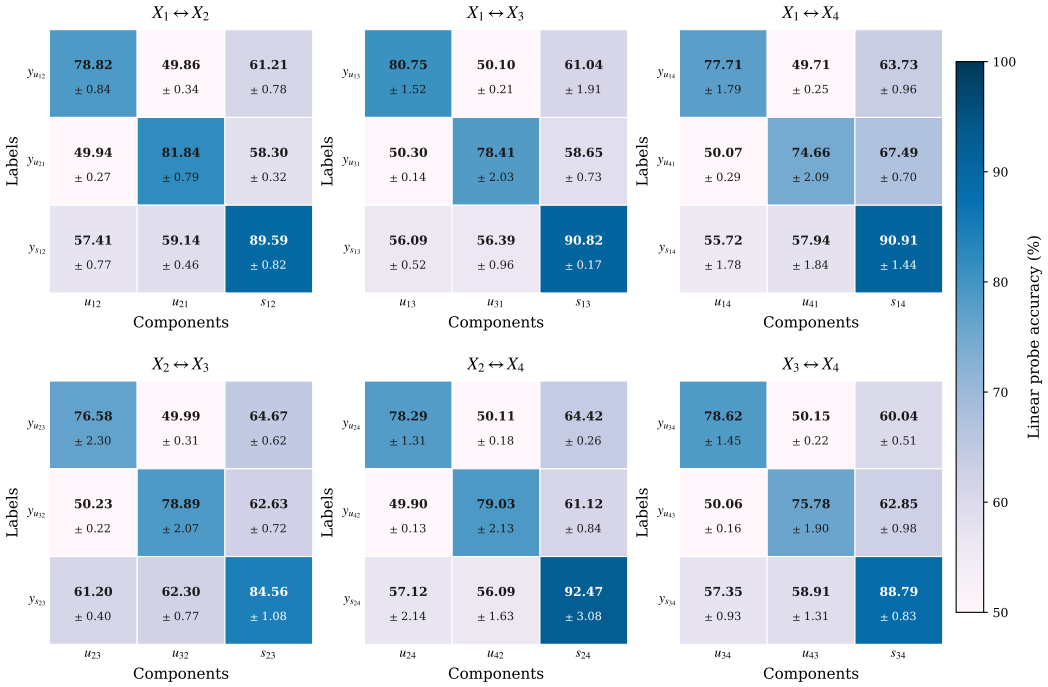


(b) GRU

Figure 12: ($\mathbf{M = 4}$) Pairwise confusion matrices for the synthetic setting with four modalities, shown for MLP and GRU. The MLP is unable to recover the desired representations, as it exhibits substantial cross-component leakage and weak intra-component prediction, while the GRU preserves strong shared representations but weakly encodes unique components.

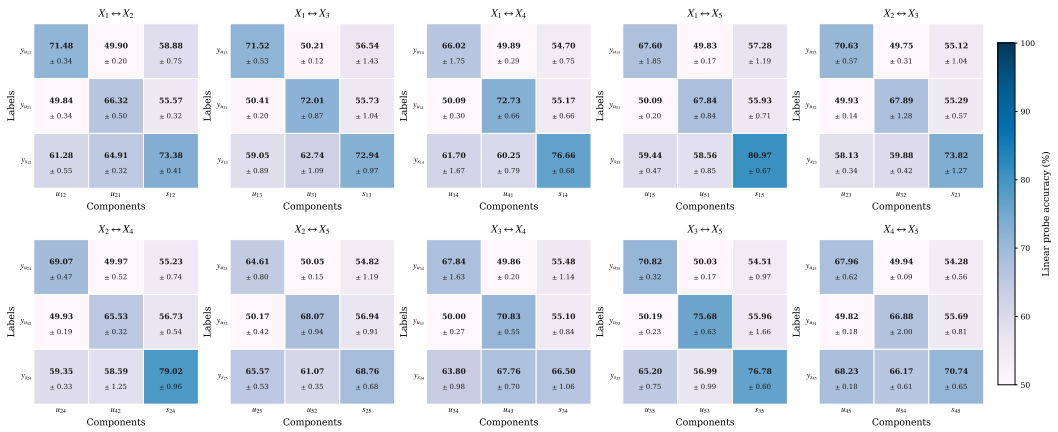


(a) gMLP

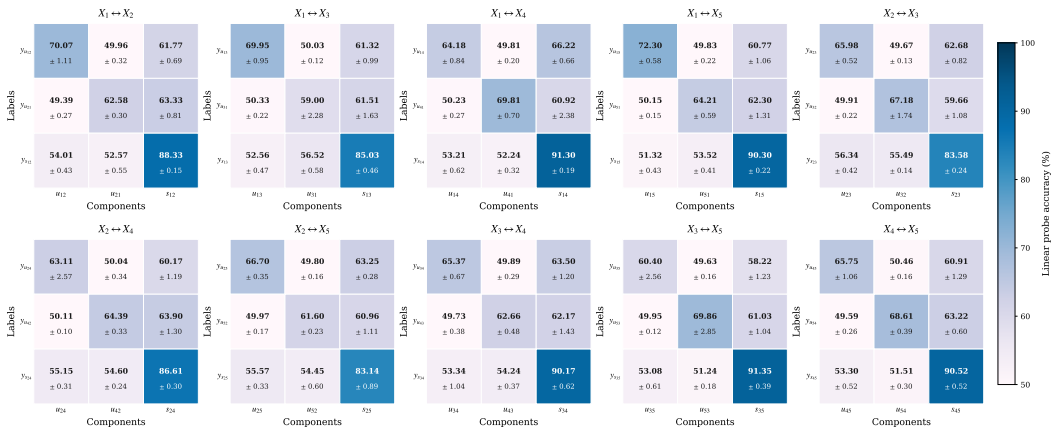


(b) **RePercENT**

Figure 13: ($\mathbf{M = 4}$) Pairwise confusion matrices for the synthetic setting with four modalities, shown for gMLP, and **RePercENT**. Both models achieve strong disentanglement, with RePercENT yielding slightly higher intra-component accuracy especially for the unique components, while gMLP exhibits marginally lower cross-component leakage.

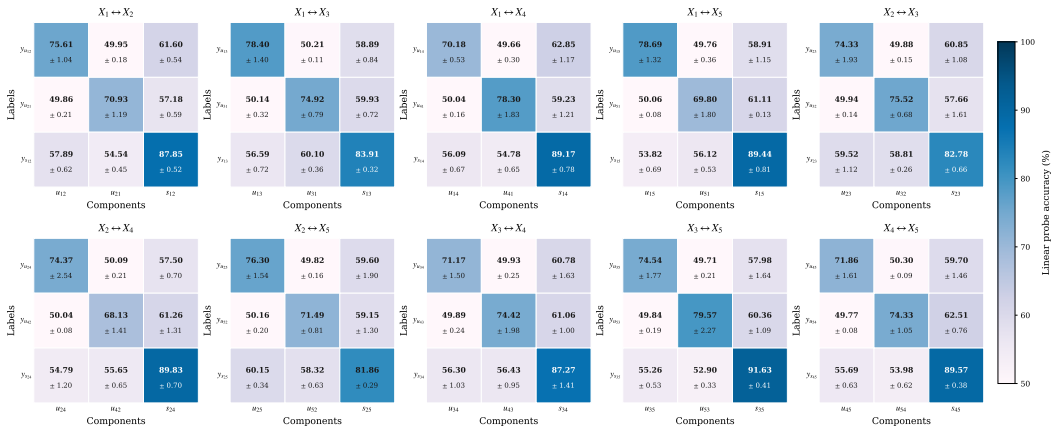


(a) MLP

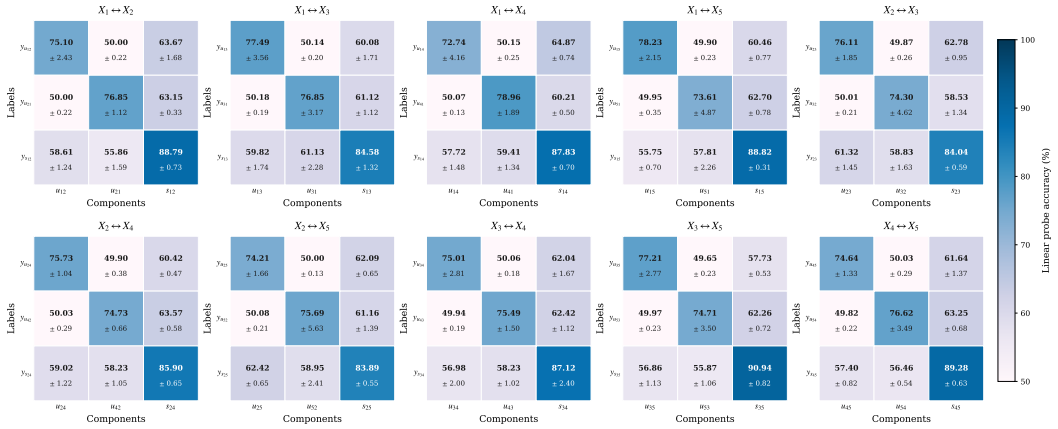


(b) GRU

Figure 14: ($\mathbf{M = 5}$) Pairwise confusion matrices for the synthetic setting with five modalities, shown for MLP and GRU. Similarly to the case of $M = 4$, the MLP is fails to recover the desired representations, while the GRU only captures the shared components successfully.



(a) gMLP



(b) **RePercENT**

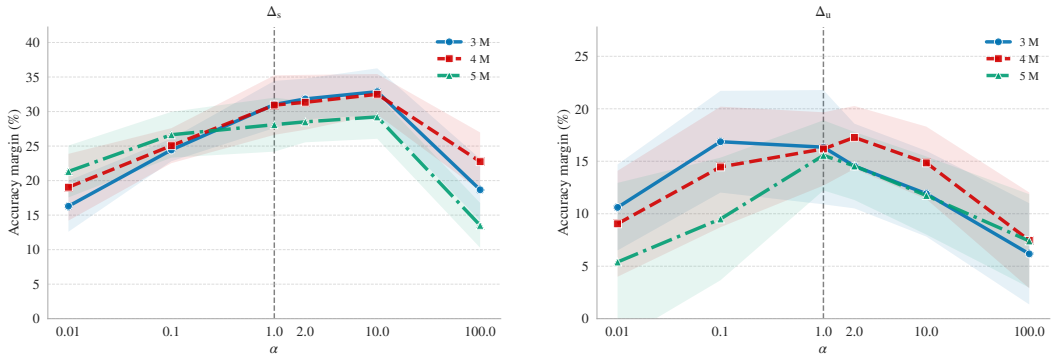
Figure 15: ($\mathbf{M = 5}$) Pairwise confusion matrices for the synthetic setting with five modalities, shown for gMLP, and **RePercENT**. Despite the increased number of modality pairs, both models successfully encode pairwise unique and shared representations, yielding similar disentanglement performance.

Hyperparameter α Figure 16 illustrates the impact of the balancing parameter α, as we keep the remaining parameters constant, for different numbers of modalities, M. Importantly, for this experiment, we set $\lambda = 1$ and $\beta = 1$. We evaluate the disentanglement performance of the unique and shared components separately, according to the quantities:

$$\Delta_s = \mathbb{E}\left[Acc_{\hat{\mathbf{s}}_{ij} \to y_{s_{ij}}}\right] - \mathbb{E}\left[Acc_{\hat{\mathbf{s}}_{ij} \to y_{u_{ij}}}\right], \qquad \Delta_u = \mathbb{E}\left[Acc_{\hat{\mathbf{u}}_{ij} \to y_{u_{ij}}}\right] - \mathbb{E}\left[Acc_{\hat{\mathbf{u}}_{ij} \to y_{s_{ij}}}\right]. \quad (31)$$

Notice that these Δ metrics, Eq. (31), measure the margin between correct intra-component prediction and inter-component leakage. A higher value indicates that the latent factors are not only task-relevant but also successfully isolated from one another. Observing both the unique and shared component performance, there is a similar emerging pattern. Concretely, for lower values of α the representation quality is quite low. The performance gradually increases as α increases, with the best behavior demonstrated within the range $[1, 10]$.

This empirical trend aligns precisely with the requirements of Theorem (1), which states that α **must be greater than or equal to** λ for stable disentanglement. This suggests that during joint optimization, prioritizing the recovery of shared components is more critical than over-penalizing the overlap between shared and unique subspaces. However, extreme values of α yield a negative effect, indicating that excessive imbalance between objectives eventually degrades both the discriminative power of the unique representations and the purity of the shared components.



(a) Δ_s (b) Δ_u

Figure 16: Sweep of parameter α across different M, when λ is fixed to 1.

H.2 MultiBench

In this section, we provide additional results on the MultiBench datasets. Table 13 reports the full per-component breakdown. RePercENT matches or outperforms the baselines, and all disentanglement methods agree on the relative utility of each component: U_T is the most predictive component for sentiment analysis (MOSI and MOSEI), followed by the shared component, while this order is reversed for humor detection (UR-FUNNY). Table 14 extends these results to the three-modality setting, where `Audio` is added and only RePercENT is directly applicable. Notably, even at $M = 3$, RePercENT requires fewer parameters than any other baseline at $M = 2$. In terms of downstream utility, `Text` clearly outperforms `Audio` in the `Text`–`Audio` pair, while in the `Audio`–`Video` pair the two modalities perform on par. An intuitive explanation is that both audio and video mainly carry paralinguistic cues, such as prosody and facial expressions, which are weaker indicators of sentiment and humor than the spoken content itself.

Table 13: Linear probe accuracy (%) across MOSI, MOSEI, and UR-FUNNY. We report (mean $\pm$ std.) over three seeds. For each dataset, the best result among the reported rows is shown in bold.

Model	Component	MOSI (↑)	MOSEI (↑)	UR-FUNNY (↑)	Params (↓)	FLOPs (↓)
FactorCL	–	52.87 ± 0.96	75.65 ± 2.79	58.76 ± 1.87	21.98M	3.789B
FOCAL	S	60.93 ± 0.36	77.20 ± 0.21	62.67 ± 0.94	17.68M	1.674B
	U_V	51.46 ± 2.18	69.59 ± 0.59	57.21 ± 0.83		
	U_T	60.59 ± 2.73	77.83 ± 0.63	60.87 ± 0.89		
JointOpt	S	60.11 ± 0.96	77.40 ± 0.35	62.04 ± 0.47	17.68M	1.675B
	U_V	49.51 ± 2.79	69.35 ± 0.42	56.58 ± 1.74		
	U_T	61.42 ± 2.56	77.33 ± 0.59	60.33 ± 1.16		
DisentangledSSL	S	64.48 ± 1.13	74.18 ± 1.06	61.56 ± 0.88	17.68M	0.835B
	U_V	50.83 ± 1.63	70.90 ± 0.14	57.78 ± 0.52		
	U_t	$\mathbf{65.35 \pm 1.48}$	74.33 ± 0.59	61.03 ± 0.39		
RePercENT	S	59.96 ± 1.10	76.65 ± 0.10	$\mathbf{62.95 \pm 0.87}$	**7.34M**	**0.052B**
	U_V	53.30 ± 0.69	68.96 ± 0.68	57.25 ± 0.77		
	U_T	62.59 ± 0.83	$\mathbf{77.99 \pm 0.32}$	62.41 ± 2.08		

Table 14: Linear probe accuracy across MOSI, MOSEI, and UR-FUNNY. We report (mean $\pm$ std.) over three seeds for the full `Video-Audio-Text` setting.

Model	Component	MOSI ↑	MOSEI ↑	UR-FUNNY ↑	Params ↓	FLOPs ↓
RePercENT (`Video-Audio`)	S	54.95 ± 2.18	69.04 ± 0.35	58.10 ± 1.55	**15.29M**	**0.94B**
	U_V	53.55 ± 0.75	69.10 ± 0.46	56.33 ± 1.57		
	U_A	53.79 ± 1.01	70.76 ± 0.21	55.92 ± 1.37		
RePercENT (`Text-Audio`)	S	61.61 ± 2.27	76.69 ± 0.53	60.40 ± 1.02		
	U_T	62.30 ± 1.49	77.21 ± 0.56	62.85 ± 1.05		
	U_A	55.88 ± 1.85	70.93 ± 0.28	55.23 ± 1.67		

Additional results on disentanglement metrics Tables 15–17 report the disentanglement metrics on the remaining MultiBench datasets. Consistent with the MOSEI results, RePercENT achieves the highest alignment between the shared components and the largest reconstruction gain, indicating that its shared and modality-specific representations are complementary. In terms of cross-modal leakage on UR-FUNNY, no model's $U_{T,V}$ carries measurable information about the `Video` modality, while for $U_{V,T}$, RePercENT exhibits the smallest leakage. Together, these results showcase RePercENT's state-of-the-art disentanglement quality and consistency across datasets.

Table 15: Cross-modal alignment between the shared components $S_{V,T}$ and $S_{T,V}$, measured by CKA and mKNN; higher is better for both. We report (mean± std.) over three seeds on MOSI and UR-FUNNY.

	MOSI		UR-FUNNY	
Model	CKA	mKNN	CKA	mKNN
FOCAL	$0.0326_{\pm.0014}$	$0.0471_{\pm.0023}$	$0.1130_{\pm.0137}$	$0.0307_{\pm.0009}$
JointOpt	$0.0362_{\pm.0034}$	$0.0488_{\pm.0019}$	$0.0707_{\pm.0061}$	$0.0287_{\pm.0003}$
DisentangledSSL	$0.0195_{\pm.0055}$	$0.0443_{\pm.0012}$	$0.0813_{\pm.0168}$	$0.0242_{\pm.0005}$
RePercENT	$\mathbf{0.7226}_{\pm.0381}$	$\mathbf{0.1794}_{\pm.0139}$	$\mathbf{0.8753}_{\pm.0155}$	$\mathbf{0.3278}_{\pm.0149}$

Table 16: Cross leakage from the unique components: R^2 of reconstructing each modality from the *other* modality's unique representation, under ridge-regularized linear probes (lower is better). We report (mean ± std.) over three seeds on UR-FUNNY

Model	$U_{T,V} \to X_V$	$U_{V,T} \to X_T$
FOCAL	$-0.0157_{\pm.0016}$	$0.0523_{\pm.0264}$
JointOpt	$-0.0166_{\pm.0020}$	$0.0602_{\pm.0118}$
DisentangledSSL	$0.0040_{\pm.0078}$	$0.0545_{\pm.0168}$
RePercENT	$\mathbf{-0.0036}_{\pm.0010}$	$\mathbf{0.0395}_{\pm.0359}$

Table 17: Reconstruction gain under ridge-regularized linear probes: reconstruction from the combined shared+specific components minus the mean of the shared-only and specific-only reconstructions, per modality (higher is better). We report (mean ± std.) over three seeds on UR-FUNNY.

Model	RG-`Text`	RG-`Video`
FOCAL	$0.0067_{\pm.0012}$	$0.0206_{\pm.0008}$
JointOpt	$0.0068_{\pm.0007}$	$0.0203_{\pm.0008}$
DisentangledSSL	$0.0045_{\pm.0009}$	$0.0206_{\pm.0016}$
RePercENT	$\mathbf{0.0366}_{\pm.0029}$	$\mathbf{0.0452}_{\pm.0035}$

H.3 IRFL

Scaling complexity Table 18 illustrates the scaling behavior of RePercENT compared to its strongest baseline, i.e., gMLP, as the number of modalities increases from two to three. For the two-modality calculation, we simulate the corresponding setting by excluding the definition encoders. Additionally, the inference latency is measured using a batch size of 32 and averaged over 100 forward passes.

Our proposed framework demonstrates exceptional parameter efficiency. The gMLP requires nearly a $3\times$ increase in parameter count to accommodate the third modality, whereas RePercENT scales by a factor of only ~ 1.5. This demonstrates that our method maintains a substantially lower computational footprint, both in parameter count and floating point operations (FLOPs), when transitioning to higher-dimensional multimodal settings. These gains are also reflected in the average inference time, where our method requires a ~ 2.2 increase, while the gMLP rises to ~ 2.9 with the presence of a third modality.

Enriching text with definition When the definitions are added all the models benefit naturally, with our framework attaining competitive performance.

Table 18: Scaling complexity for RePercENT and strongest disentanglement counterpart gMLP, as we go from two to three modalities for the IRFL detection task. Growth denotes the ratio 3-modality/2-modality. The last column reports the relative reduction of RePercENT compared to gMLP at $M = 3$.

Metric ∼ Model	**M = 2** (reference)	**M = 3**	**Growth**	**Rel. diff. @ M = 3**
Parameter count (M) (↓)				
gMLP	6.12	17.17	2.81×	reference
RePercENT	6.06	**8.96**	**1.48**×	**47.8**% lower
Floating-point operations, FLOPs (B) (↓)				
gMLP	20.89	63.16	3.02×	reference
RePercENT	5.67	**9.50**	**1.68**×	**85.0**% lower
Inference latency (ms) (↓)				
gMLP	15.32	44.12	2.88×	reference
RePercENT	16.89	**37.10**	**2.20**×	**15.9**% lower

Table 19: IRFL figurative language detection results averaged over 5 random seeds. As a fixed reference, we provide zero-shot CLIP alongside two fine-tuned variants, including projection-only and end-to-end fine-tuning. Top-1 accuracy (%) is reported (mean ± std).

Model	**Idioms**	**Simile**	**Metaphor–OoD**	**Overall**
Image vs Caption ⊕ Definition				
CLIP-ViT-B/32 (zero-shot)	26.00	48.38	36.04	38.78
CLIP-ViT-B/32 (finetune-proj)	36.00 ± 1.2	66.5 ± 1.6	41.44 ± 2.0	48.67 ± 1.3
CLIP-ViT-B/32 (finetune-all)	38.1 ± 1.1	70.4 ± 1.5	42.16 ± 3.2	50.81 ± 0.7
gMLP	43.80 ± 1.6	64.55 ± 2.9	44.92 ± 3.0	51.36 ± 1.7
GRU	46.40 ± 1.8	57.40 ± 2.4	44.92 ± 4.1	49.56 ± 1.7
RePercENT	42.50 ± 3.4	65.49 ± 2.4	44.98 ± 2.1	**51.38 ± 0.8**

H.4 TCGA

Additional Baselines On TCGA, we complement our four-modality results by comparing RePercENT against leading disentanglement approaches: FOCAL, DisentangledSSL, and JointOpt. Since none of these baselines inherently supports training on all four modalities, we restrict this experiment to the `Molecular` and `WSI` modalities and retrain RePercENT on this pair. For each baseline, we evaluate two variants: (i) with a separate 5-layer Transformer backbone per modality, matching RePercENT's capacity (B), and (ii) without a backbone, using 3-layer MLP encoders (NB). All methods use a latent dimensionality of 300. Tables 20 and 21 report the per-component performance on cancer type prediction. RePercENT achieves the highest overall accuracy on both pairs and outperforms all baselines on nine out of the ten cancer types for `WSI-Mol` and on half of them for `Mol-WSI`. Finally, the two-modality `Mol-WSI` results closely match the four-modality results in Table 4 and any small differences likely stem from optimization variance and minor architectural changes, namely the slightly larger latent dimensionality and four fewer latent slots per RePercENT encoder.

Table 20: Linear-probe accuracy (%) of the Mol–WSI representations across TCGA cohorts (mean ± std). The best result in each column is shown in **bold**. B and NB denote models with and without a backbone, respectively. †GBM is saturated, with most methods reaching 100%, so no result is highlighted.

Model	Comp.	Overall	BRCA	COAD	GBM†	HNSC	KIRC	LGG	LUAD	LUSC	OV	PRAD
RePercENT	$D_{\mathrm{mol,wsi}}$	**65.68 ± 1.0**	68.37 ± 4.4	**27.36 ± 3.2**	98.14 ± 0.4	62.31 ± 2.0	**58.32 ± 2.9**	**44.71 ± 1.5**	46.00 ± 2.2	62.65 ± 2.8	81.38 ± 1.8	**92.20 ± 0.5**
	$S_{\mathrm{mol,wsi}}$	61.24 ± 0.6	68.19 ± 4.1	19.77 ± 2.9	97.63 ± 0.7	43.46 ± 4.2	53.83 ± 2.7	37.65 ± 3.7	38.60 ± 3.7	54.29 ± 1.5	**85.17 ± 0.7**	91.60 ± 0.6
	$U_{\mathrm{mol,wsi}}$	63.52 ± 0.7	67.16 ± 4.1	25.06 ± 4.1	97.97 ± 0.5	53.08 ± 3.6	56.82 ± 2.0	40.59 ± 2.0	44.00 ± 3.2	60.82 ± 2.6	81.03 ± 1.7	91.80 ± 0.8
DisentangledSSL (B)	$D_{\mathrm{mol,wsi}}$	46.50 ± 3.5	64.65 ± 4.3	2.99 ± 5.5	100.00 ± 0.0	12.31 ± 6.7	35.33 ± 8.2	3.14 ± 1.8	32.60 ± 4.0	36.33 ± 21.3	66.55 ± 7.3	74.60 ± 10.0
	$S_{\mathrm{mol,wsi}}$	42.98 ± 3.2	70.98 ± 6.0	2.30 ± 5.1	100.00 ± 0.0	8.85 ± 7.0	33.46 ± 8.6	0.78 ± 1.3	31.00 ± 3.8	34.29 ± 19.9	61.03 ± 14.9	39.20 ± 23.8
	$U_{\mathrm{mol,wsi}}$	44.93 ± 3.3	70.14 ± 7.6	2.76 ± 5.6	100.00 ± 0.0	10.96 ± 6.1	34.21 ± 8.2	1.57 ± 1.1	31.40 ± 4.5	36.53 ± 21.0	64.48 ± 9.7	52.60 ± 19.1
DisentangledSSL (NB)	$D_{\mathrm{mol,wsi}}$	50.78 ± 0.2	61.02 ± 0.4	5.75 ± 1.2	100.00 ± 0.0	35.19 ± 1.6	25.05 ± 0.4	8.82 ± 3.0	52.20 ± 1.1	55.51 ± 1.7	69.83 ± 0.0	68.20 ± 0.8
	$S_{\mathrm{mol,wsi}}$	48.61 ± 0.2	64.28 ± 0.8	8.28 ± 1.5	100.00 ± 0.0	23.85 ± 2.8	23.55 ± 1.0	1.57 ± 2.6	48.40 ± 2.1	49.59 ± 2.2	71.03 ± 0.8	63.20 ± 1.3
	$U_{\mathrm{mol,wsi}}$	49.29 ± 0.3	62.23 ± 0.2	7.59 ± 0.6	100.00 ± 0.0	29.81 ± 2.5	24.49 ± 0.4	3.14 ± 1.5	49.00 ± 3.4	48.16 ± 2.7	70.52 ± 0.4	68.60 ± 0.9
FOCAL (B)	$D_{\mathrm{mol,wsi}}$	47.52 ± 1.6	60.56 ± 1.7	5.29 ± 5.8	98.31 ± 0.0	22.88 ± 2.9	40.19 ± 10.7	2.94 ± 2.0	26.40 ± 6.1	52.24 ± 6.2	57.76 ± 5.5	79.80 ± 2.6
	$S_{\mathrm{mol,wsi}}$	46.63 ± 1.4	60.65 ± 1.7	4.14 ± 5.2	99.15 ± 0.0	22.69 ± 3.2	39.07 ± 12.1	0.39 ± 0.9	24.40 ± 4.7	53.27 ± 5.7	53.62 ± 1.9	79.20 ± 2.6
	$U_{\mathrm{mol,wsi}}$	47.10 ± 1.6	60.84 ± 2.2	4.83 ± 5.7	98.47 ± 0.4	22.88 ± 2.7	39.81 ± 11.1	1.76 ± 1.3	25.20 ± 5.4	51.63 ± 6.4	56.90 ± 5.0	79.00 ± 2.0
FOCAL (NB)	$D_{\mathrm{mol,wsi}}$	53.36 ± 0.5	59.72 ± 1.5	13.56 ± 2.1	100.00 ± 0.0	43.85 ± 2.2	25.79 ± 0.5	10.98 ± 0.8	**53.40 ± 1.1**	54.90 ± 0.9	70.34 ± 0.5	80.60 ± 0.6
	$S_{\mathrm{mol,wsi}}$	48.95 ± 0.3	62.14 ± 0.7	6.21 ± 1.0	100.00 ± 0.0	25.96 ± 4.5	24.49 ± 0.4	2.55 ± 1.6	48.40 ± 3.0	53.06 ± 1.3	69.83 ± 0.9	67.20 ± 1.1
	$U_{\mathrm{mol,wsi}}$	48.74 ± 0.7	**75.16 ± 1.3**	0.46 ± 0.6	99.83 ± 0.4	**76.15 ± 2.8**	26.92 ± 0.4	2.16 ± 2.1	42.60 ± 4.0	**68.98 ± 6.4**	41.03 ± 3.4	11.20 ± 4.6
JointOpt (B)	$D_{\mathrm{mol,wsi}}$	54.73 ± 1.5	61.40 ± 4.1	9.20 ± 3.5	100.00 ± 0.0	36.73 ± 3.2	47.29 ± 1.7	17.25 ± 5.7	39.00 ± 1.6	56.73 ± 1.7	66.90 ± 2.9	91.20 ± 2.6
	$S_{\mathrm{mol,wsi}}$	52.48 ± 1.2	62.23 ± 4.1	6.67 ± 3.0	100.00 ± 0.0	30.96 ± 4.0	47.29 ± 1.6	9.02 ± 4.1	35.60 ± 2.2	55.31 ± 1.7	64.48 ± 8.2	87.80 ± 4.0
	$U_{\mathrm{mol,wsi}}$	51.67 ± 1.4	62.88 ± 3.5	6.44 ± 2.8	100.00 ± 0.0	25.38 ± 7.6	46.54 ± 2.8	6.86 ± 2.9	32.60 ± 3.4	55.31 ± 1.5	63.45 ± 6.3	90.20 ± 2.4
JointOpt (NB)	$D_{\mathrm{mol,wsi}}$	51.00 ± 0.6	60.74 ± 0.3	5.75 ± 1.4	100.00 ± 0.0	36.73 ± 2.2	24.86 ± 0.5	10.00 ± 1.9	49.80 ± 1.8	56.33 ± 1.1	69.83 ± 0.6	70.40 ± 1.5
	$S_{\mathrm{mol,wsi}}$	48.33 ± 1.4	66.60 ± 7.0	6.90 ± 0.8	100.00 ± 0.0	25.58 ± 2.1	23.18 ± 1.0	2.55 ± 2.3	46.00 ± 1.6	51.43 ± 3.4	70.69 ± 0.6	54.80 ± 25.6
	$U_{\mathrm{mol,wsi}}$	48.23 ± 0.6	64.65 ± 1.7	7.59 ± 3.1	100.00 ± 0.0	25.58 ± 4.1	23.18 ± 1.7	1.37 ± 1.9	41.40 ± 3.2	48.57 ± 1.9	70.86 ± 1.1	65.60 ± 5.0

Table 21: Linear-probe accuracy (%) of the WSI–Mol representations across TCGA cohorts (mean ± std). The best result in each column is shown in **bold**. B and NB denote models with and without a backbone, respectively.

Model	Comp.	Overall	BRCA	COAD	GBM	HNSC	KIRC	LGG	LUAD	LUSC	OV	PRAD
RePercENT	$D_{\mathrm{wsi,mol}}$	41.62 ± 1.0	76.28 ± 2.5	32.64 ± 1.0	36.44 ± 4.0	23.46 ± 2.5	60.37 ± 2.1	**25.88 ± 2.3**	24.20 ± 1.5	22.45 ± 1.4	**42.24 ± 3.6**	**31.40 ± 1.1**
	$S_{\mathrm{wsi,mol}}$	**42.07 ± 0.4**	80.19 ± 0.8	**33.56 ± 3.8**	**40.51 ± 1.9**	**24.62 ± 2.4**	**61.31 ± 1.1**	20.20 ± 2.0	**24.40 ± 1.3**	21.63 ± 2.2	39.83 ± 1.7	29.60 ± 2.5
	$U_{\mathrm{wsi,mol}}$	40.49 ± 0.6	75.44 ± 2.4	29.89 ± 3.0	37.29 ± 0.9	20.38 ± 2.2	60.00 ± 2.3	24.31 ± 2.1	22.20 ± 1.8	**23.27 ± 2.2**	40.34 ± 2.6	30.20 ± 2.3
DisentangledSSL (B)	$D_{\mathrm{wsi,mol}}$	30.32 ± 3.1	93.58 ± 1.2	16.09 ± 7.6	31.36 ± 5.3	7.50 ± 4.7	39.81 ± 13.8	2.35 ± 2.1	16.80 ± 6.3	6.53 ± 4.5	16.55 ± 3.7	0.40 ± 0.5
	$S_{\mathrm{wsi,mol}}$	29.24 ± 3.3	93.77 ± 1.3	15.63 ± 9.0	28.31 ± 5.7	5.77 ± 3.9	38.88 ± 17.6	1.18 ± 1.3	14.20 ± 6.6	5.51 ± 4.3	15.86 ± 5.8	0.00 ± 0.0
	$U_{\mathrm{wsi,mol}}$	29.59 ± 3.0	93.95 ± 1.4	15.86 ± 9.2	28.81 ± 6.7	6.54 ± 4.4	38.69 ± 14.1	1.18 ± 1.6	15.80 ± 6.3	5.31 ± 3.9	16.38 ± 6.4	0.20 ± 0.4
DisentangledSSL (NB)	$D_{\mathrm{wsi,mol}}$	38.74 ± 0.7	95.07 ± 1.0	28.74 ± 1.6	39.83 ± 1.7	15.77 ± 4.1	48.41 ± 1.2	10.00 ± 2.3	23.60 ± 2.3	11.22 ± 2.0	32.76 ± 1.1	17.00 ± 1.9
	$S_{\mathrm{wsi,mol}}$	35.69 ± 0.2	95.07 ± 0.2	23.68 ± 1.3	37.29 ± 2.2	11.15 ± 2.3	45.79 ± 1.3	4.12 ± 0.8	22.00 ± 1.0	9.59 ± 1.7	30.52 ± 2.8	8.80 ± 1.6
	$U_{\mathrm{wsi,mol}}$	35.78 ± 0.3	94.79 ± 0.6	24.14 ± 2.7	39.49 ± 1.8	10.96 ± 2.2	45.61 ± 1.2	4.90 ± 1.2	22.40 ± 2.3	8.98 ± 2.2	28.10 ± 1.6	10.00 ± 1.6
FOCAL (B)	$D_{\mathrm{wsi,mol}}$	28.98 ± 0.4	89.12 ± 2.1	16.78 ± 4.8	25.42 ± 1.2	6.15 ± 4.7	37.57 ± 1.8	2.94 ± 2.5	18.40 ± 2.8	8.16 ± 2.0	13.10 ± 4.9	5.00 ± 2.1
	$S_{\mathrm{wsi,mol}}$	28.35 ± 0.7	89.40 ± 1.4	16.55 ± 3.0	25.93 ± 2.7	5.96 ± 4.8	34.21 ± 2.4	1.57 ± 2.3	17.20 ± 4.9	8.57 ± 0.9	11.21 ± 3.3	5.00 ± 1.6
	$U_{\mathrm{wsi,mol}}$	27.81 ± 0.7	88.84 ± 2.2	15.17 ± 3.0	24.07 ± 4.3	5.77 ± 5.6	35.33 ± 3.2	1.96 ± 1.8	18.40 ± 2.2	6.73 ± 2.0	9.31 ± 4.2	4.80 ± 2.1
FOCAL (NB)	$D_{\mathrm{wsi,mol}}$	39.67 ± 0.4	94.33 ± 0.9	31.49 ± 1.7	37.80 ± 1.0	22.69 ± 1.8	47.48 ± 1.5	11.37 ± 2.6	19.80 ± 3.5	13.88 ± 1.4	33.97 ± 1.2	21.40 ± 1.1
	$S_{\mathrm{wsi,mol}}$	36.09 ± 0.3	94.98 ± 0.4	27.36 ± 1.3	37.12 ± 0.9	13.46 ± 1.9	45.42 ± 1.9	5.69 ± 2.1	20.40 ± 1.8	7.96 ± 1.1	29.31 ± 1.6	11.60 ± 1.5
	$U_{\mathrm{wsi,mol}}$	37.31 ± 0.2	94.98 ± 0.5	24.60 ± 1.0	36.78 ± 1.9	19.42 ± 2.8	43.18 ± 2.4	8.43 ± 0.5	21.80 ± 2.2	11.84 ± 2.1	30.69 ± 2.9	15.00 ± 2.2
JointOpt (B)	$D_{\mathrm{wsi,mol}}$	32.29 ± 1.2	91.72 ± 0.8	18.16 ± 1.7	33.90 ± 1.6	16.35 ± 2.0	37.94 ± 3.5	3.73 ± 2.1	19.80 ± 1.3	9.80 ± 3.3	18.10 ± 1.8	5.60 ± 5.9
	$S_{\mathrm{wsi,mol}}$	30.95 ± 1.6	91.63 ± 0.9	16.78 ± 1.7	32.37 ± 4.0	13.85 ± 3.4	34.58 ± 6.1	2.35 ± 2.0	21.00 ± 2.2	9.18 ± 3.0	16.03 ± 1.4	2.80 ± 4.7
	$U_{\mathrm{wsi,mol}}$	30.97 ± 0.8	92.47 ± 0.8	16.32 ± 2.1	31.36 ± 2.9	12.12 ± 1.9	36.07 ± 4.3	1.76 ± 0.4	19.80 ± 3.1	9.59 ± 3.0	18.45 ± 1.6	1.60 ± 1.5
JointOpt (NB)	$D_{\mathrm{wsi,mol}}$	38.22 ± 0.4	94.98 ± 0.6	29.43 ± 2.6	36.95 ± 1.7	16.73 ± 0.5	47.10 ± 1.9	9.61 ± 1.9	22.40 ± 2.2	11.22 ± 1.4	31.55 ± 1.4	17.40 ± 1.8
	$S_{\mathrm{wsi,mol}}$	35.57 ± 0.4	**95.35 ± 0.8**	25.52 ± 2.9	35.25 ± 1.0	11.92 ± 0.9	45.23 ± 1.4	4.12 ± 1.3	23.20 ± 1.9	9.59 ± 2.0	27.93 ± 1.0	9.20 ± 1.8
	$U_{\mathrm{wsi,mol}}$	34.72 ± 0.5	94.70 ± 0.8	23.45 ± 1.9	34.92 ± 2.2	9.04 ± 0.9	44.11 ± 0.8	5.69 ± 0.8	21.20 ± 2.3	8.98 ± 1.7	27.59 ± 1.6	8.60 ± 2.7

Towards interpretability Beyond serving as a geometric diagnostic of the learned shared spaces, the heatmaps in Figure 17 suggest that the learned shared representations capture biologically plausible cancer-type structure. For example, the lung cancer subtypes LUAD and LUSC often appear closer to one another than to unrelated cancer types, which is consistent with their shared tissue of origin. More broadly, the presence of structured off-diagonal similarities suggests that the shared representations are not only encoding cancer-type separability, but also capturing relationships between diseases that may arise from common tissue, histology, or molecular programs. Importantly, the observed geometry suggests that RePercENT may expose meaningful cross-cancer relationships, motivating future work that integrates domain-specific annotations, such as mutation status, molecular subtypes, pathway activity, immune profiles, to systematically validate and characterize the biological factors driving these patterns.

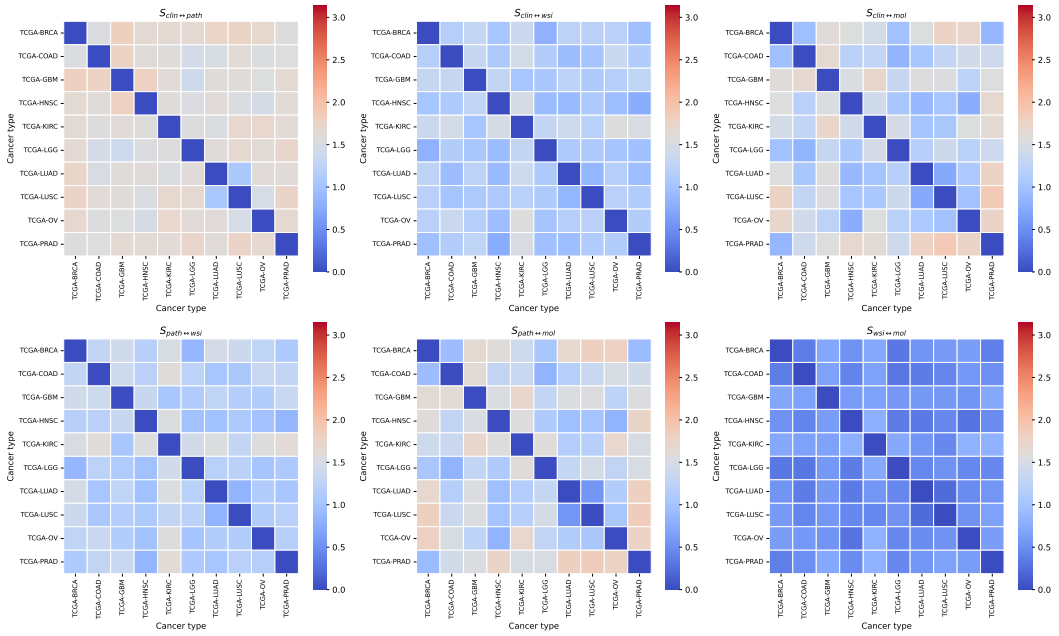


Figure 17: Evaluation of the angular distance between the studied cancer types, across the extracted shared components. The shared components, as they are modeled with the use of von-Mises Fisher distribution, they lie on a hyper-sphere. We cluster the derived shared embeddings across modalities, based on their cancer type and calculate their centroid. The heatmap reflects the angular distance between those centroids within each shared component, where values close to zero (blue) denote proximity between those cancer types, while values closer to π (red) imply that the embeddings between these two matrices are further apart.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: All the claims found in the abstract and introduction are supported by the results and theorems discussed in the paper.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss all the limitations and future work in the dedicated section 5.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: The paper provides a set of detailed definitions and theorems that motivate the choice of the developed model. We provide all the necessary information-theory background in Appendix D and all the corresponding formal proofs are also presented and can be found in Appendix E.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The datasets we use are publicly available and we provide a step-by-step algorithm on how to generate the synthetic data in Appendix G.1. To make our experiments reproducible, the set of all hyperparameters of the models used along with the training specifications are provided in detail in the Appendix.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Yes, the full code with detailed instructions for reproducibility have been made public and can be found here: https://github.com/Vascorn/RePercENT.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide detailed descriptions of all the training parameters, procedure and hyperparameters used for every model that we train in the Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We provide all our results in terms of downstream task performance in the format "mean $\pm$ standard deviation from the mean".

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide detailed information about the used resources in the Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: We preserve anonymity in our submission and our submission abides by the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss potential positive and negative societal impacts in the Appendix C.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The release of our data and models upon acceptance, does not pose a direct risk of misuse.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We provide citations for all sources of code and/or data used. For the datasets, we reference the original open-source work they are based on, in full accordance with their licensing terms.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: This paper does not release any new assets beyond code, which is adequately documented in order to reproduce all results from the paper.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: No crowdsourcing or human subjects were involved in the experiments conducted for this paper.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No crowdsourcing or human subjects were involved in the experiments conducted for this paper.

    Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not a core component of this paper in terms of methods.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.