STRUCTURED LATENT MODELING FOR SUPERVISED MULTIMODAL INFORMATION DECOMPOSITION

Wanting Huang
Department of Computer Science
University of Iowa
wanting-huang@uiowa.edu

Sanvesh Srivastava
Department of Statistics
University of Iowa
sanvesh-srivastava@uiowa.edu

Weiran Wang
Department of Computer Science
University of Iowa
weiran-wang@uiowa.edu

ABSTRACT

Multimodal prediction relies on diverse forms of evidence: information repeated across modalities, cues specific to a single source, and complex cross-modal dependencies that emerge only when inputs are considered together. While recent methods promote richer interactions, they lack a principled way to isolate these target-relative contributions within learned continuous representations. We introduce a framework that applies contrastive or masked objectives at intermediate layers, coupled with source-wise invertible normalizing flows and a supervised, low-rank latent variable model. This architecture explicitly factorizes the joint distribution into shared task-relevant variation, modality-specific predictive variation, and task-irrelevant dependence. Drawing connections to prior multimodal learning assumptions, our approach evaluates how modalities independently and jointly contribute to the target. Ultimately, this framework unites intermediate representation learning with structured likelihood-based guidance, offering a practical latent-variable lens for characterizing continuous multimodal interactions. Empirically, we demonstrate the effectiveness of our approach across diverse multimodal benchmarks, showing robust improvements in predictive performance.

1 INTRODUCTION

Supervised multimodal fusion aims to learn representations that reflect the heterogeneity and interconnections between modalities (Liang et al., 2024a), capturing both overlapping (redundant) and modality-specific (unique) information to maximize predictive performance. Traditionally, multiview learning relies on the redundancy assumption—presuming all task-relevant information is shared across modalities (Tosh et al., 2021; Federici et al., 2020). However, this often fails in real-world supervised settings where modalities carry unique predictive signals. When applied to complex multimodal data, standard late fusion architectures frequently suffer from unimodal bias (also known as modality competition or imbalance (Huang et al., 2022; Zhang et al., 2024b)). During joint training, the fusion network tends to disproportionately rely on a dominant, easily accessible modality, inadvertently suppressing the learning of weaker modalities, and occasionally resulting in worse performance than their single-modality counterparts. A prominent example is visual question answering, where the vision modality is often ignored because the text already correlates strongly with the answer (Goyal et al., 2017; Cadene et al., 2019). By analyzing the learning dynamics of deep fusion networks, recent studies Zhang et al. (2024b); Kontras et al. (2025) suggest that inter-modality correlation is the fundamental driver of this unimodal bias.

Recent approaches have attempted to address these challenges by quantifying or isolating multimodal interactions. Unsupervised representation learning methods (Liang et al., 2023b; Dufumier et al., 2025; Wen et al., 2026) rely on handcrafted, label-preserving perturbations and contrastive learning to extract either a single representation per view, or a fused representation that captures

both shared and unique predictive information. Meanwhile, supervised approaches use labels to filter out task-irrelevant information and mitigate modality competition through mechanisms such as gradient balancing, prototype-based modality rebalancing, and game-theoretic regularization (Wang et al., 2020; Fan et al., 2023; Kontras et al., 2025). However, they do not learn explicitly disentangled latent variables. These developments motivate the important research question: *how can a structured model of the source–target distribution support representation learning and mitigate unimodal bias?* Such a model must explicitly disentangle private and shared predictive information, accommodating nonlinear features while distinguishing source–target relationships from other cross-source associations.

To address these limitations, we propose a supervised latent variable model (LVM) framework centered on the *Supervised Structured Multimodal Latent Variable Model* (S2MLVM). This novel formulation generalizes both unsupervised (Bach & Jordan, 2005; Lock et al., 2013) and existing supervised (Palzer et al., 2022) LVMs. Its key design divides each view's latent space into distinct factors: a shared predictive component, a unique predictive component, and a shared nuisance (non-predictive) component. Marginalizing these factors yields a low-rank-plus-diagonal covariance, imposing a compact dependence structure that accommodates shared source variation without requiring all of it to be directly predictive. Explicitly separating shared and private predictive signals from nuisance variations, our framework enables classifiers to optimally weight each component. Just as single-view disentanglement improves interpretability and sample efficiency (Bengio et al., 2013; Locatello et al., 2020), this multimodal factorization directly empowers supervised fusion, boosting generalization and mitigating unimodal bias. In summary, our main contributions are:

- We integrate the structured S2MLVM covariance model into representation learning, using neural encoders as initial feature extractors coupled with dimension-preserving normalizing flows (Dinh et al., 2017; Papamakarios et al., 2021). To accommodate high-dimensional, nonlinear multimodal data and optimize the statistical model jointly with the representations, we develop two instantiations (S2MLVM-MASK and S2MLVM-CONTRAST). These optimize a three-term objective combining a task loss, the joint Flow–S2MLVM density criterion, and an auxiliary representation objective to maintain input information while reducing dimensionality. Both maintain modality-specific encoding without early cross-modal fusion: S2MLVM-MASK reconstructs masked clean features within each source, whereas S2MLVM-CONTRAST contrasts augmented multimodal observations.

- We theoretically analyze our framework's connection to common assumptions in multimodal learning. By drawing connections to Partial Information Decomposition (PID, Williams & Beer, 2010; Bertschinger et al., 2014), we elucidate the mathematical mechanisms—such as cooperative nuisance suppression and V-structure exploitation—that generate synergistic information under our LVM.

- We demonstrate the efficacy of S2MLVM across synthetic and real-world benchmarks. In controlled settings, it accurately recovers task-relevant subspaces. Across diverse real-world datasets, both instantiations mitigate unimodal bias and outperform strong representation-learning baselines in two- and three-modality scenarios.

2 METHOD

We develop S2MLVM, a supervised multimodal learning framework that integrates representation learning with structured probabilistic modeling. Each modality is encoded via a source-specific pathway and transformed by an invertible normalizing flow. The resulting representations are jointly modeled with the target using a structured latent variable model that explicitly disentangles shared predictive, modality-specific predictive, and target-irrelevant variation. This joint model provides likelihood-based generative guidance during training. While the framework naturally extends to an arbitrary number of modalities (by incorporating additional source pathways and block-structured LVM components), we present the formulation for two sources $x^{(1)}$ and $x^{(2)}$ with target y for clarity.

2.1 STRUCTURED LATENT VARIABLE MODELING

Consider paired observations $\mathcal{D} = \{(x_i^{(1)}, x_i^{(2)}, y_i)\}_{i=1}^N$. Each source $x_i^{(m)}$ is mapped to a continuous representation through a source-specific encoder g_{θ_m} followed by an invertible flow f_{ϕ_m}:

$$h^{(m)} = g_{\theta_m}(x^{(m)}) \in \mathbb{R}^{d_m}, \qquad u^{(m)} = f_{\phi_m}(h^{(m)}) \in \mathbb{R}^{d_m}, \quad m \in \{1, 2\}, \tag{1}$$

and these representations do not depend on the target y. We also extract continuous embedding $\tilde{y} \in \mathbb{R}^{d_y}$ of the target y if it is categorical.

Our framework, the Supervised Structured Multimodal Latent Variable Model (S2MLVM), defines a joint generative process over the concatenated features $w = [u^{(1)}; u^{(2)}; \tilde{y}] \in \mathbb{R}^d$, where $d = d_1 + d_2 + d_y$. We model w using four independent latent factors $z = [z^{12}; z^1; z^2; z^c]$ of dimension $K = k_{12} + k_1 + k_2 + k_c$. The generative model $w = Az + \epsilon$ follows the structured linear mapping:

$$\begin{bmatrix} u^{(1)} \\ u^{(2)} \\ \tilde{y} \end{bmatrix} = \begin{bmatrix} \Lambda_1^{12} & \Lambda_1^1 & 0 & \Lambda_1^c \\ \Lambda_2^{12} & 0 & \Lambda_2^2 & \Lambda_2^c \\ B^{12} & B^1 & B^2 & 0 \end{bmatrix} \begin{bmatrix} z^{12} \\ z^1 \\ z^2 \\ z^c \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_y \end{bmatrix}, \qquad z \sim \mathcal{N}(0, I_K), \tag{2}$$

where Λ_m^{12}, Λ_m^m, and Λ_m^c represent the shared predictive, modality-specific predictive, and target-irrelevant components, respectively, for modality m. Loading matrices B^{12} and B^m characterize the effects of shared and modality m-specific latent variables on $\tilde{y}$. The observation noise $\epsilon = [\epsilon_1; \epsilon_2; \epsilon_y]$ is drawn from $\mathcal{N}(0, \Psi)$, with a positive diagonal covariance $\Psi = \mathrm{diag}(\Psi_1, \Psi_2, \Psi_y)$.

This block-structured loading pattern explicitly disentangles the latent space into distinct semantic components. The factor z^{12} captures shared predictive information that influences both sources and the target, while z^1 and z^2 isolate unique predictive information specific to each source. Crucially, the target $\tilde{y}$ does not depend on the shared nuisance factor z^c. This structure ensures that task-irrelevant cross-modal correlations are absorbed by z^c rather than confounding the predictive representations. Marginalizing over the latent factors gives

$$w \sim \mathcal{N}(0, \Omega), \qquad \Omega = AA^\top + \Psi. \tag{3}$$

This formulation provides a unified perspective on multi-view latent variable models. In the unsupervised setting (where the target y is absent), it is impossible to distinguish predictive shared variation (z^{12}) from nuisance shared variation (z^c), so they collapse into a single shared factor. Under this restriction, the model reduces to probabilistic Canonical Correlation Analysis (CCA) (Bach & Jordan, 2005) if we retain only this shared factor, and to Joint and Individual Variation Explained (JIVE) (Lock et al., 2013) if we also include the unique factors z^1 and z^2. Conversely, the full supervised structure generalizes recent predictive LVMs like supervised JIVE (Palzer et al., 2022), which lack the nuisance factor z^c. By introducing z^c, S2MLVM explicitly separates shared predictive variation from task-irrelevant cross-modal correlations.

2.2 FLOW-BASED DISTRIBUTION MODELING

To apply the Gaussian S2MLVM to complex multimodal data, we must bridge the gap between the structured Gaussian base distribution and the highly non-Gaussian neural representations $h^{(m)}$. We achieve this using source-wise normalizing flows f_{ϕ_m}, which provide invertible, dimension-preserving transformations with tractable Jacobian determinants (Dinh et al., 2017; Papamakarios et al., 2021). Although each flow processes a single modality independently, their outputs $u^{(m)}$ are modeled jointly with the target $\tilde{y}$ under the S2MLVM base distribution to retain all source–source and source–target dependencies. This unified construction enables a joint maximum likelihood framework: we can perform maximum likelihood estimation of the structured loading matrix A and residual covariance Ψ while simultaneously learning the flow parameters $\phi = (\phi_1, \phi_2)$ to map the initial encoder outputs into Gaussian-distributed variables.

For fixed initial encoders, the joint log-density of this feature-space model is given by

$$\log p_{\phi,A,\Psi}(h^{(1)}, h^{(2)}, \tilde{y}) = \log \mathcal{N}(w; 0, \Omega) + \sum_{m=1}^{2} \log \left| \det J_{f_{\phi_m}}(h^{(m)}) \right|, \tag{4}$$

where $\Omega = AA^\top + \Psi$. The corresponding fitting objective for an observation is the negative log-likelihood, $\mathcal{L}_{\text{density}} = -\log p_{\phi,A,\Psi}(h^{(1)}, h^{(2)}, \tilde{y})$. The Gaussian term fits the structured joint relationships, while the Jacobian terms account for volume changes induced by the flows.

While finite-capacity flows practically optimize the Kullback-Leibler divergence between the empirical distribution and the Gaussian LVM, normalizing flows theoretically possess universal approximation capabilities for continuous probability distributions (Papamakarios et al., 2021). Because our flows operate independently on each modality, this provides a strong asymptotic guarantee that, given sufficient capacity, they can perfectly transform the source representations to match the Gaussian marginals required by the joint LVM.

2.3 Auxiliary Representation Learning

Because the initial encoders perform dimension reduction on the raw inputs, an auxiliary objective is necessary to prevent representation degeneracy before the representations enter the structured LVM. Although the subsequent normalizing flows are invertible, this auxiliary loss ensures the encoders extract and retain sufficient information. The objective $\mathcal{L}_{\text{rep}}$ is computed directly on the encoder representations $h^{(m)}$ and is instantiated by either masked reconstruction ($\mathcal{L}_{\text{rec}}$) or contrastive learning ($\mathcal{L}_{\text{con}}$). Only one of these alternatives is used in each variant; they are not added together.

Masked reconstruction. Following recent masked representation learning methods (He et al., 2022; Baevski et al., 2022), let $H^{(m)}$ denote clean features at an intermediate layer, $\widehat{H}^{(m)}$ the reconstructed features, and M_m the set of masked indices for modality m. We define

$$\mathcal{L}_{\text{rec}} = \frac{1}{2} \sum_{m=1}^{2} \ell\Big(\widehat{H}^{(m)}[M_m], \text{sg}(H^{(m)}[M_m])\Big), \tag{5}$$

where ℓ is a distance metric (e.g., Smooth L1 loss) averaged over the masked entries, and sg stops gradients through the target argument. The objective encourages recovery of source features from partial observations.

Contrastive learning. Let v and v^+ be normalized projections of the representations $h^{(m)}$ extracted from two independently augmented versions of the same modality. The InfoNCE objective (Oord et al., 2018; Chen et al., 2020) for this observation is:

$$\mathcal{L}_{\text{con}} = -\log \frac{\exp(v^\top v^+/\tau)}{\exp(v^\top v^+/\tau) + \sum_{v^- \in \mathcal{N}(v)} \exp(v^\top v^-/\tau)}, \tag{6}$$

where $\tau > 0$ is the temperature, and $\mathcal{N}(v)$ contains negative samples from other observations. In practice, this loss is symmetrized across both anchor directions. Crucially, positive pairs are formed from the same joint observation rather than across modalities, leaving the extraction of cross-modal shared information to the S2MLVM.

2.4 Total Representation Learning Objective

Because our framework explicitly isolates predictive components, we construct the final representation by computing and concatenating the posterior means of the predictive latent factors (z^{12}, z^1, and z^2) given $(u^{(1)}, u^{(2)})$ from the S2MLVM. To provide direct discriminative supervision and accommodate standard downstream evaluations, this representation is passed to a multilayer perceptron (MLP) optimized with a standard task loss $\mathcal{L}_{\text{MLP}}$ (e.g., cross-entropy for classification or mean squared error for regression) against the ground-truth targets.

The overall training objective minimizes a weighted combination of this discriminative task loss, the generative Flow–S2MLVM maximum-likelihood loss ($\mathcal{L}_{\text{density}}$) defined in Section 2.2, and the auxiliary representation loss ($\mathcal{L}_{\text{rep}}$) defined in Section 2.3:

$$\mathcal{L} = \mathcal{L}_{\text{MLP}} + \lambda_{\text{density}}\mathcal{L}_{\text{density}} + \lambda_{\text{rep}}\mathcal{L}_{\text{rep}}. \tag{7}$$

Optimization. In practice, the total loss $\mathcal{L}$ is averaged over the training set and minimized using stochastic gradient descent (SGD). All parameters—including source encoders, normalizing flows, structured covariance parameters, and the final MLP classifier—are optimized jointly end-to-end.

3 CONNECTIONS TO OTHER MULTIMODAL LEARNING FRAMEWORKS

To understand how our model captures complex multimodal interactions, we analyze it through the lens of Partial Information Decomposition (PID) (Williams & Beer, 2010; Liang et al., 2024b), a framework that decomposes joint predictive information into unique, redundant, and synergistic components. Our S2MLVM connects deeply to this formalism. PID achieves this decomposition by defining a constraint space Δ_P of adversarial joint distributions Q that preserve the true source-target marginals:

$$\Delta_P = \{Q \mid Q(u^{(1)}, \tilde{y}) = P(u^{(1)}, \tilde{y}),\ Q(u^{(2)}, \tilde{y}) = P(u^{(2)}, \tilde{y})\}. \tag{8}$$

While the latent factors in our model (z^{12}, z^1, z^2, z^c) possess clear intuitive meanings, they do not map one-to-one onto the information-theoretic atoms of PID. Instead, as we detail in Appendix B, the formal PID definition extracts synergy by identifying a worst-case adversary q^*_{MI} that minimizes the joint mutual information:

$$q^*_{MI} = \arg\min_{Q \in \Delta_P} I_Q(u^{(1)}, u^{(2)}; \tilde{y}). \tag{9}$$

We show that this adversary is remarkably strong: it correlates the private noise across modalities to align with the combined task signal, maximally reducing the joint predictive information.

Furthermore, we demonstrate that the assumption of conditional independence ($u^{(1)} \perp\!\!\!\perp u^{(2)} \mid \tilde{y}$)—a widespread premise in multimodal machine learning (Blum & Mitchell, 1998; Chaudhuri et al., 2009)—provides a tractable analytic proxy q^*_{CE} for the intractable PID optimization. This adversary is defined by maximizing the conditional entropy:

$$q^*_{CE} = \arg\max_{Q \in \Delta_P} H_Q(u^{(1)}, u^{(2)} \mid \tilde{y}). \tag{10}$$

Because our generative structure models the latent features as jointly Gaussian, their information quantities are fully determined by the joint covariance matrix. Consequently, the constraint $Q \in \Delta_P$, which fixes the source-target marginals, restricts the adversary to exclusively manipulating the cross-source covariance block. This means the adversarial covariance matrix $\Sigma^{(Q)}$ is identically the original covariance matrix Σ plus an off-diagonal cross-source perturbation block C. As detailed in Appendix B, the optimal conditional entropy adversary acts as a decoupling filter by injecting the exact negative conditional cross-covariance $C_{CE} = -\Sigma_{u^{(1)}u^{(2)}|\tilde{y}}$ as this perturbation. Under our S2MLVM parameterization, this injected covariance takes a closed form:

$$C_{CE} = -\Big(\underbrace{\Lambda_1^c (\Lambda_2^c)^\top}_{\text{shared nuisance}} + \underbrace{\Lambda_1^{12} (\Lambda_2^{12})^\top}_{\text{shared signal}} - \underbrace{\Sigma_{u^{(1)}\tilde{y}} \Sigma_{\tilde{y}\tilde{y}}^{-1} \Sigma_{\tilde{y}u^{(2)}}}_{\text{induced V-structure}} \Big). \tag{11}$$

This term-by-term algebraic decomposition reveals that *all* latent factors provide opportunities for synergistic predictive power. To eliminate synergy and achieve conditional independence, the adversary C_{CE} must systematically inject noise to cancel three distinct mechanisms: the cooperative suppression of the shared nuisance z^c, the noise-averaging of the redundant shared signal z^{12}, and the V-structure "explaining-away" correlation induced by the independent private signals z^1 and z^2.

Because q^*_{CE} provides a closed-form but sub-optimal minimization of the joint mutual information ($I_{q^*_{MI}} \leq I_{q^*_{CE}}$), substituting it into the PID equations yields analytic bounds on the true optimal quantities. Specifically, our tractable proxy safely underestimates both Synergy ($S_{CE} \leq S_{MI}$) and Redundancy ($R_{CE} \leq R_{MI}$), while overestimating the Unique Information ($U_j^{CE} \geq U_j^{MI}$). Finally, our structured model explicitly accommodates the spectrum of multimodal assumptions. Many classical multi-view frameworks rely on the *multi-view redundancy assumption* (Federici et al., 2020; Tsai et al., 2021; Tosh et al., 2021), positing that all task-relevant information is shared (z^1, z^2 are empty). By explicitly isolating the private predictive factors (z^1, z^2) from the shared redundancy (z^{12}), S2MLVM provides a generalization that efficiently learns in both highly redundant and highly unique multimodal settings without discarding synergistic interactions.

4 RELATED WORK

Unsupervised multimodal representation learning. Multimodal representation learning traditionally relies on correspondence and partial observations to retain information across sources. Contrastive learning and mutual information-based methods (Oord et al., 2018; Belghazi et al., 2018;

Chen et al., 2020; Tian et al., 2020; Radford et al., 2021) learn transferable representations by aligning paired views. The Multi-View Information Bottleneck (Federici et al., 2020) formalizes this by providing an information-theoretic account for retaining shared predictive content under strict multi-view redundancy assumptions. To move beyond purely shared information, recent works explicitly isolate modality-specific or complementary content. Deep generative models (Wang et al., 2016; Lee & Pavlovic, 2021a; Palumbo et al., 2023; Zhang et al., 2026) separate shared and private variations through structured likelihood modeling. Parallel efforts in self-supervised learning, such as FactorCL (Liang et al., 2023b) and DisentangledSSL (Wang et al., 2025), achieve this factorization using complex augmentations and contrastive learning. Other approaches actively encourage complementary interactions: CoMM (Dufumier et al., 2025) models joint spaces to capture beyond-redundancy interactions, while InfMasking (Wen et al., 2026) contrasts partially masked features against complete fusions to elicit synergy. While these self-supervised methods demonstrate the necessity of learning from more than just cross-modal agreement, they largely rely on complex ad-hoc regularizations or masking strategies. We generalize this principle of likelihood-based separation to the supervised setting, providing a rigorous statistical formalization for complex source–target dependencies.

Balancing supervised multimodal optimization. In joint multimodal training, models often disproportionately rely on a single dominant modality (Peng et al., 2022; Huang et al., 2022), a phenomenon formally linked to inter-modality correlations (Zhang et al., 2024b). To mitigate this bias, a prominent line of work directly intervenes in the optimization process. Techniques like OGM (Peng et al., 2022), AGM (Li et al., 2023), and MLB (Kontras et al., 2024) dynamically scale gradient updates based on ongoing modality contributions, while scheduling methods like MMPareto (Wei & Hu, 2024), MLA (Zhang et al., 2024a), and ReconBoost (Hua et al., 2024) explicitly reconcile conflicting unimodal and multimodal objectives. More recently, methods use game-theoretic valuation (Wei et al., 2024) or mutual-information decomposition (Kontras et al., 2025) to formalize modality cooperation. We share the motivation of using task-related dependence to guide learning; however, S2MLVM achieves this balance inherently via its generative objective, bypassing the need for explicit gradient modulation.

Target-relative information decomposition. Partial information decomposition (PID) isolates the unique, redundant, and synergistic information provided by multiple sources about a target (Williams & Beer, 2010; Bertschinger et al., 2014). Recent works have extended these concepts to multimodal learning, developing scalable estimators for quantifying interactions (Liang et al., 2023a) and deriving learning guarantee (Liang et al., 2024a). While these works focus on estimating information-theoretic quantities from datasets or fixed representations, evaluating PID on continuous, high-dimensional neural representations presents significant computational challenges. Existing continuous estimators rely on variational optimization (Pakman et al., 2021), neural estimation (Kleinman et al., 2021), convex approximations for Gaussians (Venkatesh & Schamberg, 2022; Venkatesh et al., 2023), or invertible transformations to latent Gaussian spaces (Zhao et al., 2025). We build on this latent-Gaussian principle rather than proposing a new PID definition. By explicitly disentangling shared and modality-specific predictive factors, our framework naturally yields the covariance structures needed to analyze synergistic information.

5 EXPERIMENTS

We adopt the experimental setups, including datasets and base architectures, of recent related works. For all experiments, we tune hyperparameters and perform ablation studies exclusively on the validation set. We select the best-performing model configuration on the validation set and evaluate its task MLP predictions to report the final metrics on the held-out test set. We report accuracy for classification tasks and mean squared error (MSE) for regression tasks. Our repeated-run results use 5 seeds and are summarized by the mean and sample standard deviation. We focus on comparing with the current state-of-the-art unsupervised (e.g., COMM (Dufumier et al., 2025), InfMasking (Wen et al., 2026)) and supervised (MCR (Kontras et al., 2025)) methods using their configurations (data, architecture, and implementation); results taken from these prior works are annotated with *.

Table 1: Recovery of S2MLVM subspaces. Here $dim = d_1 = d_2$ and $k = k_{12} = k_1 = k_2 = k_c$.

Setting		Canonical correlation (CC) ↑
$k = 4$, $\sigma = 0.4$	$dim = 12$	0.99479 ± 0.00315
	$dim = 24$	0.98816 ± 0.00396
	$dim = 48$	0.94867 ± 0.03073
	$dim = 96$	0.93104 ± 0.03477
$dim = 24$, $k = 4$	$\sigma = 0.2$	0.98624 ± 0.00523
	$\sigma = 0.4$	0.98816 ± 0.00396
	$\sigma = 0.6$	0.99109 ± 0.00206
	$\sigma = 0.8$	0.99137 ± 0.00155
$dim = 48$, $\sigma = 0.4$	$k = 4$	0.94867 ± 0.03073
	$k = 8$	0.96488 ± 0.01118
	$k = 12$	0.97333 ± 0.00559
	$k = 16$	0.96502 ± 0.00804

5.1 Subspace Recovery on Synthetic Data

We first verify that maximum likelihood modeling via SGD recovers the true task-relevant latent structure. We generate two observed sources and the response from independent Gaussian latent blocks (z^{12}, z^1, z^2, and z^c) following the S2MLVM loading pattern in equation 2. We use isotropic observation noise with covariance $\Psi = \sigma^2 I$. we then apply fixed random nonlinear flow transformations to both observed sources (without the nonlinear mappings, the recovery would be perfect across settings). For the experiments, we first vary the ambient source dimension over $d_1 = d_2 \in \{12, 24, 48, 96\}$ while fixing $k = k_{12} = k_1 = k_2 = k_c = 4$ and $\sigma = 0.4$. Second, we vary the observation-noise standard deviation over $\sigma \in \{0.2, 0.4, 0.6, 0.8\}$ while fixing $d_1 = d_2 = 24$ and $k = 4$. Then, we vary the dimensionality of each latent block over $k \in \{4, 8, 12, 16\}$ while fixing $d_1 = d_2 = 48$ and $\sigma = 0.4$. Models are trained on $10{,}000$ samples for $12{,}000$ steps using SGD (momentum 0.9, batch size 512, learning rate 3×10^{-3}). The checkpoint achieving the lowest negative log-likelihood on a $2{,}000$-sample validation set is evaluated on $4{,}096$ held-out test samples. Results are averaged over five random seeds.

Because individual loading coordinates are only identifiable up to rotations within each latent block, we evaluate recovery at the subspace level using a rotation-invariant metric:

$$\mathrm{CC} = \frac{1}{k_{12} + k_1 + k_2} \sum_{i=1}^{k_{12}+k_1+k_2} \sigma_i\Big(Q^\top \widehat{Q}\Big). \tag{12}$$

Here, Q and $\widehat{Q}$ denote orthonormal bases for the true and estimated task-relevant subspaces, respectively, while σ_i indicates the i-th singular value. Mean canonical correlation (CC) measures subspace alignment through their principal angles (larger is better), and is bounded in $[0, 1]$. Table 1 demonstrates that S2MLVM consistently recovers the task-relevant subspace under varying conditions. While performance degrades smoothly as the estimation problem becomes more challenging—such as with higher ambient or latent dimensions—the recovered predictive geometry remains robust, maintaining high canonical correlation across all settings.

5.2 Experiments on TriFeature

We use TriFeatures (Dufumier et al., 2025) as a controlled diagnostic to examine whether the latent variables learned by our model exhibit the intended functional specialization. By providing explicit control over target labels, TriFeatures enables us to systematically evaluate the extraction of shared, source-specific, and cross-source information. Each input image comprises three categorical attributes—shape, texture, and color—which are used to construct four targeted diagnostic tasks. Redundancy (R) targets the shape attribute shared across both images. The two unique tasks, U_1 and U_2, correspond to the independent texture attributes of the first and second images, respectively. These three attribute-level tasks are evaluated as 10-way classification problems. Finally, Synergy (S) is a binary classification task evaluating a cross-source relation determined jointly by the texture

Table 2: Classification accuracy (%) on TriFeature, for different task labels R/U_1/U_2/S.

Method	R-ACC ↑	U_1-ACC ↑	U_2-ACC ↑	U-ACC ↑	S-ACC ↑
FactorCL	99.8^*			62.5^*	46.5^*
CoMM	$99.9 \pm 0.1^*$	$84.4 \pm 2.4^*$	$91.2 \pm 1.0^*$	$86.8 \pm 3.0^*$	$71.4 \pm 3.5^*$
InfMasking	$99.9 \pm 0.1^*$	$90.7 \pm 2.1^*$	$91.4 \pm 3.0^*$	$90.6 \pm 2.3^*$	$77.0 \pm 4.2^*$
MCR	99.9 ± 0.1	92.1 ± 1.4	93.2 ± 0.9	92.7 ± 1.2	90.2 ± 3.9
S2MLVM-MASK	93.4 ± 1.1	91.3 ± 2.2	84.9 ± 3.5	88.1 ± 2.7	85.4 ± 4.3
S2MLVM-CONTRAST	$\mathbf{99.9 \pm 0.1}$	$\mathbf{95.1 \pm 0.7}$	$\mathbf{95.3 \pm 0.8}$	$\mathbf{95.2 \pm 0.5}$	$\mathbf{92.3 \pm 0.9}$
z^{12}	93.6 ± 3.3	34.2 ± 7.1	32.7 ± 11.8	33.5 ± 9.5	56.2 ± 2.6
z^1	70.4 ± 4.3	83.5 ± 9.2	14.1 ± 7.5	48.8 ± 8.4	48.6 ± 1.6
z^2	64.3 ± 9.5	10.6 ± 5.2	78.4 ± 14.0	44.5 ± 9.6	49.8 ± 1.7

Table 3: Test results on two modality benchmarks, by unsupervised, supervised, and our methods.

Method	V&T EE ↓	MIMIC ↑	CREMA-D ↑	UCF101 ↑	MOSEI ↑
FactorCL	$10.8 \pm 0.6^*$	$67.3 \pm 0.0^*$	61.1 ± 1.3	50.3 ± 2.2	78.5 ± 0.1
CoMM	$8.0 \pm 2.1^*$	$66.4 \pm 0.4^*$	59.9 ± 2.5	51.0 ± 2.1	79.7 ± 0.3
InfMasking	$4.2 \pm 0.4^*$	$68.1 \pm 0.4^*$	63.9 ± 2.7	51.3 ± 1.9	80.4 ± 0.4
Unimodals	I : 1.9 ± 0.0 F : 87.7 ± 0.5	S : 56.3 ± 0.5 T : 68.0 ± 0.4	V : 55.6 ± 3.4 A : 56.5 ± 3.7	V : 42.7 ± 1.2 A : 28.8 ± 1.7	V : 65.2 ± 0.2 T : 78.4 ± 1.3
Joint Training	1.9 ± 0.1	66.8 ± 1.4	$62.6 \pm 5.8^*$	$47.7 \pm 1.5^*$	$80.5 \pm 0.2^*$
MCR	2.0 ± 0.18	68.6 ± 0.8	$76.1 \pm 1.6^*$	$55.2 \pm 1.8^*$	$\mathbf{80.8 \pm 0.4}^*$
S2MLVM					
-MASK	$\mathbf{1.6 \pm 0.1}$	$\mathbf{69.1 \pm 0.4}$	74.7 ± 1.7	57.0 ± 1.8	80.5 ± 0.8
-CONTRAST	2.6 ± 0.2	68.7 ± 0.5	$\mathbf{76.2 \pm 2.0}$	$\mathbf{57.2 \pm 2.0}$	80.3 ± 0.6

of the first image and the color of the second. Crucially, neither attribute alone is sufficient to predict the target; the synergistic relation can only be resolved by integrating information contributed by both sources. Following InfMasking (Wen et al., 2026), we train on $10{,}000$ synthetic image pairs and evaluate on $4{,}096$ held-out pairs. For the encoders of both S2MLVM-MASK and S2MLVM-CONTRAST, each view uses an AlexNet frontend and an independent width-512 Transformer with its own CLS token. We train both architectures for 100 epochs. The representation dimension is 512 for $u^{(1)}$, $u^{(2)}$, and $\tilde{y}$. The latent dimensions are $k_{12} = k_1 = k_2 = k_c = 8$.

We provide the classification accuracy on test set of several methods in Table 2. In general, supervised methods (MCR and S2MLVM) outperform unsupervised methods (e.g., COMM and InfMasking), due to joint training of predictors and predictive representations, and contrastive learning turns out to be a better representation learning objective for this dataset. To see that our latent factors well extracts the desired predictive information, after model training (with task loss on (z^{12}, z^1, z^2)), we train a small MLP on each extracted latent factor by S2MLVM-CONTRAST and obtain 3 additional accuracies. As expected, For the tasks R, U_1 and U_2, performing classification only on z^{12}, z^1, z^2 correspondingly maintains most accuracy. For the synergy task S however, each individual factor is not sufficient to maintain desired predictive information.

5.3 REAL-WORLD MULTIMODAL BENCHMARKS

Datasets and tasks. We evaluate two groups of real-world benchmarks. The first follows the InfMasking setup (Wen et al., 2026) on MultiBench (Liang et al., 2021), covering Vision&Touch end-effector (V&T EE) prediction, MIMIC diagnostic-group prediction, MOSI sentiment classification, UR-FUNNY humor recognition, and MUSTARD sarcasm recognition. The second follows the MCR configurations for CREMA-D, UCF101, and MOSEI (Kontras et al., 2025). Complete dataset definitions and evaluation conventions are provided in Appendix A.1; dataset-specific model architetures are summarized in Appendix A.2.

Table 4: Test results on two modality benchmarks MOSI (MO), UR-FUNNY (UR), MUSTARD (MU), by unsupervised, supervised, and our methods.

Method	MO ↑	UR ↑	MU ↑
FactorCL	$51.2 \pm 1.6^*$	$60.5 \pm 0.8^*$	$55.8 \pm 0.9^*$
CoMM	$63.7 \pm 2.5^*$	$63.3 \pm 0.5^*$	$64.4 \pm 1.1^*$
InfMasking	$69.0 \pm 1.2^*$	$64.3 \pm 0.9^*$	$66.8 \pm 2.5^*$
Unimodals	V : 55.1 ± 1.3	52.4 ± 0.5	57.7 ± 2.2
	T : 72.0 ± 1.4	61.7 ± 0.7	63.4 ± 2.0
Joint Training	72.9 ± 1.6	61.7 ± 1.0	60.6 ± 2.0
MCR	75.0 ± 1.2	54.4 ± 2.1	60.5 ± 3.5
S2MLVM			
-MASK	$\mathbf{75.8 \pm 0.8}$	63.7 ± 2.0	$\mathbf{68.6 \pm 5.5}$
-CONTRAST	74.3 ± 1.9	$\mathbf{65.0 \pm 1.8}$	66.0 ± 2.9

Table 5: Architecture ablation on MOSI (MO), UR-FUNNY (UR), MUSTARD (MU) validation sets.

S2MLVM	MO ↑	UR ↑	MU ↑
-MASK	73.6	63.8	63.8
w/o Flow	68.7	59.6	48.6
w/o z^c	64.6	62.0	57.3
-CONTRAST	76.2	63.6	59.4
w/o Flow	70.2	61.4	55.8
w/o z^c	72.9	58.9	52.9

Table 6: Test accuracies (%) on three-modality benchmarks.

Method	UR-FUNNY ↑	V&T Contact ↑	MOSEI ↑
CoMM	$64.8 \pm 1.1^*$	$94.1 \pm 0.2^*$	79.0 ± 0.3
InfMasking	$65.6 \pm 1.2^*$	$94.1 \pm 0.1^*$	78.9 ± 0.8
MCR	66.6 ± 1.5	94.1 ± 0.3	$81.1 \pm 0.4^*$
S2MLVM-MASK	66.6 ± 1.8	94.0 ± 0.3	80.7 ± 0.7
S2MLVM-CONTRAST	66.0 ± 0.9	93.6 ± 0.9	81.4 ± 0.5

Experiments with Two Input Modalities. Tables 3 and 4 summarize the performance on the two-modality benchmarks. We observe that the optimal auxiliary objective varies depending on the dataset: S2MLVM-MASK excels on V&T EE, MIMIC, MOSI, and MUSTARD, whereas S2MLVM-CONTRAST achieves the best performance on UR-FUNNY, CREMA-D, and UCF101. Across all tasks, S2MLVM consistently outperforms previous unsupervised and supervised methods, indicating better utilization of all decomposed predictive information. Optimal hyperparameters and sensitivity analysis are discussed in Appendix (cf. Table 7 and Table 8).

Ablation studies. We conduct two sets of ablation studies to examine the contributions of the proposed model components and training objectives. As shown in Table 5, removing the invertible transformations (w/o Flow) or the separate nuisance component (w/o z^c) significantly degrades performance across all tasks. In the appendix, we ablate the loss terms in Table 9. The results confirms that both auxiliary representation objective and likelihood are indispensable for effectively capturing predictive information, validating our structural design.

Experiments with Three Input Modalities. We extend our LVM with one more modality, and consider UR-FUNNY with visual, textual, and acoustic features; V&T contact prediction with image, force, and proprioceptive observations; and MOSEI with visual, textual, and acoustic features. As shown in Table 6, S2MLVM consistently outperforms baselines, demonstrating that it can scale to multiple modalities while preserving strong predictive performance.

6 CONCLUSION

We introduced S2MLVM, a structured supervised latent variable model that tackles modality competition in multimodal fusion. By disentangling representations into shared predictive, modality-specific predictive, and task-irrelevant nuisance factors, our framework isolates and better utilizes diverse multimodal evidence. Integrating this structured likelihood model with representation learning—via dimension-preserving flows—bridges the gap between classic statistical models and high-dimensional deep networks. Empirically, S2MLVM mitigates unimodal bias, outperform-

ing strong baselines across diverse real-world benchmarks, and scales seamlessly to more than two modalities. Ultimately, our approach offers a unified latent-variable lens for characterizing complex continuous multimodal interactions. Given the inherent complexity of synergy, cleanly extracting all PID components into completely separate latent variables remains an open question. Future work will explore extending our framework to handle partially missing modalities (Wu & Goodman, 2018; Lee & Pavlovic, 2021b; Lee & van der Schaar, 2021).

AI USE STATEMENT

Parts of this manuscript were drafted with assistance from AI. All AI-assisted content was subsequently reviewed, verified, and edited by the authors.

REPRODUCIBILITY STATEMENT

We provide the information needed throughout the main paper and appendix. The main paper describes the proposed model, learning objectives, and common experimental protocol. The appendix provides additional details on dataset preprocessing and evaluation protocols, source-specific architectures, hyperparameter configurations, and experimental settings.

REFERENCES

Francis R. Bach and Michael I. Jordan. A probabilistic interpretation of canonical correlation analysis. Technical Report 688, Dept. of Statistics, University of California, Berkeley, 2005.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. data2vec: A general framework for self-supervised learning in speech, vision and language. In *ICML*, 2022.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *ICML*, 2018.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

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

Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pp. 92–100, 1998.

Remi Cadene, Corentin Dancette, Hedi Ben-younes, Matthieu Cord, and Devi Parikh. RUBi: Reducing unimodal biases for visual question answering. In *NeurIPS*, 2019.

Kamalika Chaudhuri, Sham M. Kakade, Karen Livescu, and Karthik Sridharan. Multi-view clustering via canonical correlation analysis. In *ICML*, 2009.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. second edition, 2006.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using Real NVP. In *ICLR*, 2017.

Benoit Dufumier, Javiera Castillo Navarro, Devis Tuia, and Jean-Philippe Thiran. What to align in multimodal contrastive learning? In *ICLR*, 2025. URL https://openreview.net/forum?id=Pe3AxLq6Wf.

Yunfeng Fan, Wenchao Xu, Haozhao Wang, Junxiao Wang, and Song Guo. Pmr: Prototypical modal rebalance for multimodal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023.

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

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, 2017.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Doll'ar, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022.

Cong Hua, Qianqian Xu, Shilong Bao, Zhiyong Yang, and Qingming Huang. ReconBoost: Boosting can achieve modality reconcilement. In *ICML*, 2024.

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

Michael Kleinman, Alessandro Achille, Stefano Soatto, and Jonathan C Kao. Redundant information neural estimation. *Entropy*, 23(7):922, 2021.

Konstantinos Kontras, Christos Chatzichristos, Matthew Blaschko, and Maarten De Vos. Improving multimodal learning with multi-loss gradient modulation. In *BMVC*, 2024.

Konstantinos Kontras, Thomas Strypsteen, Christos Chatzichristos, Paul Pu Liang, Matthew B. Blaschko, and Maarten De Vos. Balancing multimodal training through game-theoretic regularization. In *NeurIPS*, 2025. URL https://openreview.net/forum?id=auiURbhoYx.

Changhee Lee and Mihaela van der Schaar. A variational information bottleneck approach to multi-omics data integration. In *AISTATS*, 2021.

Mihee Lee and Vladimir Pavlovic. Private-shared disentangled multimodal vae for learning of latent representations. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1692–1700. IEEE, 2021a.

Mihee Lee and Vladimir Pavlovic. Private-shared disentangled multimodal VAE for learning of latent representations. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2021b.

Hong Li, Xingyu Li, Pengbo Hu, Yinuo Lei, Chunxiao Li, and Yi Zhou. Boosting multi-modal model performance with adaptive gradient modulation. In *ICCV*, 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, 2021. URL https://arxiv.org/abs/2107.07502.

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

Paul Pu Liang, Zihao Deng, Martin Q. Ma, James Zou, Louis-Philippe Morency, and Russ Salakhutdinov. Factorized contrastive learning: Going beyond multi-view redundancy. In *NeurIPS*, 2023b.

Paul Pu Liang, Chun Kai Ling, Yun Cheng, Alex Obolenskiy, Yudong Liu, Rohan Pandey, Alex Wilf, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal learning without labeled multimodal data: Guarantees and applications. In *ICLR*, 2024a.

Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Foundations & trends in multimodal machine learning: Principles, challenges, and open questions. *ACM Comput. Surv.*, 56(10), 2024b.

Francesco Locatello, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. Weakly-supervised disentanglement without compromises. In *ICML*, 2020.

Eric F. Lock, Katherine A. Hoadley, J. S. Marron, and Andrew B. Nobel. Joint and individual variation explained (jive) for integrated analysis of multiple data types. *Annals of Applied Statistics*, 7(1):523–542, 2013.

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

Ari Pakman, Amin Nejatbakhsh, Dar Gilboa, Abdullah Makkeh, Luca Mazzucato, Michael Wibral, and Elad Schneidman. Estimating the unique information of continuous variables. In *NeurIPS*, 2021.

Emanuele Palumbo, Imant Daunhawer, and Julia E. Vogt. MMVAE+: Enhancing the generative quality of multimodal VAEs without compromises. In *ICLR*, 2023.

Elise F. Palzer, Christine H. Wendt, Russell P. Bowler, Craig P. Hersh, Sandra E. Safo, and Eric F. Lock. sJIVE: Supervised joint and individual variation explained. *Computational Statistics & Data Analysis*, 175:107547, 2022.

George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021.

Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *CVPR*, 2022.

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 *ICML*, 2021.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *ECCV*, 2020.

Christopher Tosh, Akshay Krishnamurthy, and Daniel Hsu. Contrastive learning, multi-view redundancy, and linear models. In *Algorithmic Learning Theory*, 2021.

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

Praveen Venkatesh and Gabriel Schamberg. Partial information decomposition via deficiency for multivariate gaussians. In *IEEE International Symposium on Information Theory (ISIT)*, 2022. doi: 10.1109/ISIT50566.2022.9834649.

Praveen Venkatesh, Corbett Bennett, Sam Gale, Tamina K. Ramirez, Greggory Heller, Severine Durand, Shawn Olsen, and Stefan Mihalas. Gaussian partial information decomposition: Bias correction and application to high-dimensional data. In *NeurIPS*, 2023.

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 *ICLR*, 2025.

Weiran Wang, Xinchen Yan, Honglak Lee, and Karen Livescu. Deep variational canonical correlation analysis. *arXiv:1610.03454*, 2016.

Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *CVPR*, 2020.

Yake Wei and Di Hu. MMPareto: Boosting multimodal learning with innocent unimodal assistance. In *ICML*, 2024.

Yake Wei, Ruoxuan Feng, Zihe Wang, and Di Hu. Enhancing multimodal cooperation via sample-level modality valuation. In *CVPR*, 2024.

Liangjian Wen, Qun Dai, Jianzhuang Liu, Jiangtao Zheng, Yong Dai, Dongkai Wang, Zhao Kang, Jun Wang, Zenglin Xu, and Jiang Duan. Infmasking: Unleashing synergistic information by contrastive multimodal interactions. *Advances in Neural Information Processing Systems*, 38: 15529–15555, 2026.

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

Mike Wu and Noah Goodman. Multimodal generative models for scalable weakly-supervised learning. In *NeurIPS*, 2018.

Xiaohui Zhang, Jaehong Yoon, Mohit Bansal, and Huaxiu Yao. Multimodal representation learning by alternating unimodal adaptation. In *CVPR*, 2024a.

Yedi Zhang, Peter E. Latham, and Andrew Saxe. Understanding unimodal bias in multimodal deep linear networks. In *ICML*, 2024b.

Yijie Zhang, Yiyang Shen, and Weiran Wang. Disentanglement of variations with multimodal generative modeling. In *ICLR*, 2026.

Wenyuan Zhao, Adithya Balachandran, Chao Tian, and Paul Pu Liang. Partial information decomposition via normalizing flows in latent gaussian distributions. In *NeurIPS*, 2025. URL https://openreview.net/forum?id=X13jOIhnog.

A APPENDIX

A.1 DATASETS

We evaluate the proposed framework across affective computing, multimodal language understanding, robotics, and audio–visual recognition benchmarks.

MultiBench. **MOSI** contains $2{,}199$ opinion segments paired with visual, acoustic, and textual observations. In the InfMasking-style evaluation, we use the two-modality setting and evaluate the representation with binary sentiment classification.

MOSEI extends this setting to approximately $23{,}000$ monologue clips and is evaluated under both the two-modality supervised protocol inherited from MCR and the visual–acoustic–textual three-modality setting.

UR-FUNNY contains $16{,}514$ examples from $1{,}866$ TED-talk videos and evaluates humor recognition. We use both its two-modality configuration and its visual–acoustic–textual configuration for the three-modality experiments.

MUSTARD contains 690 balanced dialogue utterances and evaluates sarcasm recognition (Liang et al., 2021; Wen et al., 2026; Kontras et al., 2025).

Robotics benchmark. **Vision&Touch** contains 150 robotic manipulation trajectories with $1{,}000$ time steps per trajectory. We consider two tasks. End-effector prediction is a regression task evaluated from the two-modality representation, whereas contact prediction is evaluated as a three-modality classification task using image, force, and proprioceptive observations. We report MSE$\times 10^{-4}$ for end-effector prediction and classification accuracy for contact prediction.

Audio–visual benchmarks. **CREMA-D** evaluates six-way emotion recognition from synchronized face video and speech produced by 91 actors.

UCF101 is used for audio–visual action recognition; following the MCR protocol, we retain the subset of action categories for which both modalities are available. These benchmarks follow the supervised evaluation protocol used by MCR (Kontras et al., 2025).

MIMIC combines a 24-hour sequence of clinical measurements with static patient descriptors and defines a binary diagnostic-group prediction task. It contains $53{,}423$ admissions from $38{,}597$ patients. We retain MIMIC in the comparison where required to reproduce the benchmark coverage and averages reported by prior multimodal interaction methods.

A.2 ARCHITECTURE

As illustrated in Figure 1,the overall architecture consists of dataset-specific neural encoders, normalizing flows, and the S2MLVM latent variable model. Each source pathway consists of a dataset-specific backbone and input adapter, followed by an independent Transformer with its own learnable CLS token. The clean, unprojected CLS representations are passed separately to their respective flows, without cross-modal attention or target inputs in the source encoders. We evaluate two model variants, S2MLVM-MASK and S2MLVM-CONTRAST, which share this identical architecture and differ only in their auxiliary representation loss. To make downstream predictions, the concatenated posterior means of the predictive latent blocks (z^{12}, z^1, and z^2) are extracted from the LVM and passed through a task MLP comprising two linear layers with an intervening ReLU activation, where the supervised task loss is applied.

However, specific details are handled differently for each dataset. For MOSI, UR-FUNNY, MUSTARD, and MOSEI, we use separate temporal Transformers with modality-specific input projections to encode the precomputed visual and textual feature sequences. For MIMIC, an MLP encodes the static patient descriptors, while a GRU encodes the multivariate clinical time series. For Vision&Touch end-effector regression, we use a ResNet-18 image backbone and a temporal force encoder as separate source pathways. The three-modality contact-prediction configuration additionally includes a dedicated proprioception encoder. For CREMA-D and UCF101, we adopt the ResNet-

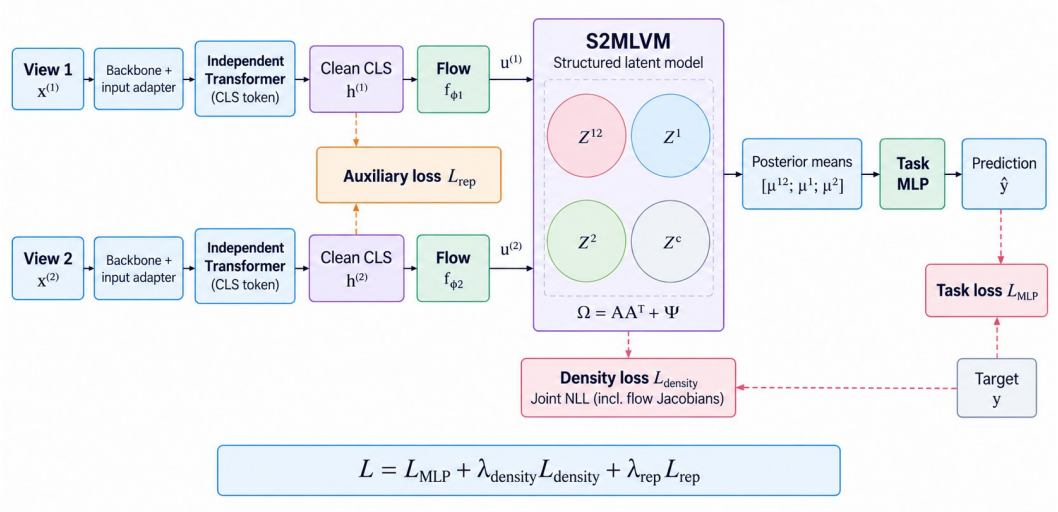

Figure 1: Overview architecture.

Table 7: Selected hyperparameter for each dataset.

Dataset	S2MLVM-MASK		S2MLVM-CONTRAST		dim k	loss
	Mask ratio	Batch size	Temperature	Batch size		$(\lambda_{\text{density}}, \lambda_{\text{rep}})$
MOSI	0.55	96	0.10	32	4	(1,1)
MIMIC	0.60	96	0.20	48	12	(1,1)
UR-FUNNY	0.30	64	0.10	128	4	(1,1)
MUSTARD	0.45	64	0.10	128	4	(1,1)
CREMA-D	0.55	32	0.15	8	16	(0.2,0.8)
UCF101	0.55	16	0.05	16	12	(0.2,1)
MOSEI	0.85	64	0.05	32	12	(0.5,0.8)

18 backbone configuration of MCR (Kontras et al., 2025), with independent visual and acoustic encoders operating on video frames and audio spectrograms, respectively.

Besides, the unimodal and joint-training baselines are reproduced following MCR (Kontras et al., 2025). Unimodal training optimizes each modality independently with the supervised task loss, whereas joint training optimizes all modalities together using a single supervised task objective.

A.3 HYPERPARAMETER TUNING

The final comparison uses validation-selected hyperparameters, summarized in Table 7. For S2MLVM-MASK, these are the mask ratio and batch size; for S2MLVM-CONTRAST, they are the InfoNCE temperature and batch size. Here, mask ratio $m \in [0, 1]$ denotes the fraction of input feature tokens randomly masked for the masked-reconstruction objective, temperature τ is the temperature parameter that scales the similarity logits before the softmax in the InfoNCE objective, and batch size B denotes the number of training examples in each minibatch. The validation sensitivity of these hyperparameters is reported in Table 8.

A.4 LOSS ABLATION

We use a strict one-factor-at-a-time validation procedure. We conduct an ablation study on the loss coefficients to examine the contribution of each loss term, with the results reported in Table 9.

Table 8: Hyperparameter sensitivity on validation sets.

Study	Candidate	MOSI ↑	MUSTARD ↑
Mask ratio (S2MLVM-MASK)	$m = 0.40$	77.6	65.2
	$m = 0.45$	79.1	**70.3**
	$m = 0.50$	77.8	66.7
	$m = 0.55$	**80.0**	70.0
	$m = 0.60$	75.6	62.7
	$m = 0.65$	77.9	65.2
	$m = 0.70$	79.7	64.4
	$m = 0.75$	74.4	63.1
Temperature (S2MLVM-CONTRAST)	$\tau = 0.05$	77.7	64.9
	$\tau = 0.10$	**79.9**	**69.2**
	$\tau = 0.15$	79.0	68.7
	$\tau = 0.20$	79.2	67.1
Batch size (S2MLVM-MASK)	$B = 32$	75.5	68.0
	$B = 48$	78.9	65.8
	$B = 64$	79.1	**70.2**
	$B = 96$	**79.6**	64.7
	$B = 128$	77.0	64.6
Batch size (S2MLVM-CONTRAST)	$B = 32$	**82.7**	64.9
	$B = 48$	80.8	68.8
	$B = 64$	79.0	67.7
	$B = 96$	79.0	65.9
	$B = 128$	78.6	**72.2**

Table 9: Loss-coefficient ablations on validation datasets.

Variant	MOSI ↑	MUSTARD ↑
S2MLVM-MASK ($\lambda_{\text{density}} = 1, \lambda_{\text{rep}} = 1$)	79.0	68.6
w/o L_{density}	64.6	63.2
w/o L_{rep}	73.3	66.5
S2MLVM-CONTRAST ($\lambda_{\text{density}} = 1, \lambda_{\text{rep}} = 1$)	79.9	66.4
w/o L_{density}	75.4	64.2
w/o L_{rep}	77.5	62.8

B CONNECTION TO PARTIAL INFORMATION DECOMPOSITION (PID)

In this section, we analyze the theoretical origins of predictive synergy within our proposed latent variable model. While the previous sections established the practical benefits of explicitly disentangling predictive and nuisance factors for supervised fusion, we now formalize how this structure captures synergistic interactions between modalities. We utilize the Partial Information Decomposition (PID) framework (Williams & Beer, 2010; Bertschinger et al., 2014) to isolate synergistic information, demonstrating how specific latent factors—such as shared nuisance variables and task-relevant signals—interact to provide predictive power that is only accessible when multiple modalities are observed jointly.

Specifically, PID decomposes the joint and marginal mutual informations into four non-negative components: the unique information provided by each individual view (Unique_1 and Unique_2), the redundant information shared by both views (Redundancy), and the synergistic information that arises only from their combination (Synergy). These components satisfy the classic PID system of

equations (Williams & Beer, 2010):

$$\begin{aligned} I_P(U^1, U^2; Y) &= \text{Unique}_1 + \text{Unique}_2 + \text{Redundancy} + \text{Synergy}, \\ I_P(U^1; Y) &= \text{Unique}_1 + \text{Redundancy}, \\ I_P(U^2; Y) &= \text{Unique}_2 + \text{Redundancy}. \end{aligned}$$

Because this system is underdetermined, modern PID formulations define a constraint space Δ_P of *adversarial joint distributions* Q that preserve the pairwise source-target marginals of the true distribution P:

$$\Delta_P = \left\{ Q(U^1, U^2, Y) \;\middle|\; Q(U^1, Y) = P(U^1, Y),\ Q(U^2, Y) = P(U^2, Y) \right\}.$$

By minimizing the joint information over $Q \in \Delta_P$, the framework isolates the synergistic interactions from the baseline predictive information. Following Bertschinger et al. (2014) and Liang et al. (2024a), this adversarial optimization yields formal definitions for the unique and synergistic information:

$$\begin{aligned} \text{Unique}_j &= \min_{Q \in \Delta_P} I_Q(U^j; Y \mid U^{-j}), \\ \text{Synergy} &= I_P(U^1, U^2; Y) - \min_{Q \in \Delta_P} I_Q(U^1, U^2; Y), \\ \text{Redundancy} &= \max_{Q \in \Delta_P} I_Q(U^1; U^2; Y) = I_P(U^1; Y) + I_P(U^2; Y) - \min_{Q \in \Delta_P} I_Q(U^1, U^2; Y). \end{aligned}$$

B.1 The Adversarial Latent Framework

To provide a self-contained analysis of synergy, we briefly recall the generative structure of our latent variable model. We assume the multi-view observations U^1, U^2 and the target Y are generated from a set of mutually orthogonal Gaussian latent factors:[1]

$$\begin{aligned} U^1 &= \Lambda_1^{12} z^{12} + \Lambda_1^c z^c + \Lambda_1^1 z^1 + \epsilon^1 \\ U^2 &= \Lambda_2^{12} z^{12} + \Lambda_2^c z^c + \Lambda_2^2 z^2 + \epsilon^2 \\ Y &= B^{12} z^{12} + B^1 z^1 + B^2 z^2 + \epsilon^Y \end{aligned}$$

Here, z^{12} is the task-relevant shared factor, z^c is the task-irrelevant shared nuisance factor, z^1, z^2 are view-specific task-relevant private factors, and $\epsilon^j \sim \mathcal{N}(\mathbf{0}, \Sigma_{\epsilon^j})$ is the full-rank independent private noise. We assume the nuisance loadings Λ_j^c are orthogonal to the combined task-relevant loadings $\tilde{\Lambda}_j = [\Lambda_j^{12}, \Lambda_j^j]$ (i.e., $(\tilde{\Lambda}_j)^\top \Lambda_j^c = \mathbf{0}$), and the task-relevant loading matrices $\tilde{\Lambda}_j$ have full column rank.

For our Gaussian model, any adversarial distribution $Q \in \Delta_P$ must preserve the true marginal predictive covariances ($\Sigma_{U^1U^1}, \Sigma_{U^2U^2}, \Sigma_{U^1Y}, \Sigma_{U^2Y}, \Sigma_{YY}$). Because these marginals are fixed, the only parameter the adversary can manipulate is the cross-source covariance $\Sigma_{U^1U^2}$. This allows us to frame any valid adversary $q \in \Delta_P$ as a pure noise-injection strategy: the adversary retains the true generative process p (leaving the latent factors and marginal variances untouched), but injects a cross-covariance matrix C directly between the independent private noises ϵ^1, ϵ^2. This creates the adversarial joint covariance:

$$\Sigma_{UU}^{(Q)} = \Sigma_{UU}^{(p)} + \begin{bmatrix} \mathbf{0} & C \\ C^\top & \mathbf{0} \end{bmatrix}.$$

Because the injected matrix C alters only the off-diagonal cross-source covariance, any choice of C that keeps the resulting $\Sigma_{UU}^{(Q)}$ positive semi-definite guarantees $Q \in \Delta_P$. The question then becomes: what constitutes the 'worst-case" cross-covariance C? This depends on the chosen adversarial objective.

[1] For notational simplicity in this section, we use uppercase U^1, U^2, Y to denote the neural features $u^{(1)}, u^{(2)}$ and target embedding $\tilde{y}$ respectively, and we drop the parentheses on the block indices of the noise ϵ.

B.2 THE CONDITIONAL INDEPENDENCE ADVERSARY (q^*_{CE})

To formally capture the notion of decoupling the views, we first define the least-informative distribution using the Maximum Entropy (MaxEnt) principle applied to the sources. We seek the distribution q^*_{CE} that maximizes the conditional entropy of the sources $h_Q(U \mid Y)$, which is equivalent to maximizing the determinant $|\Sigma^{(Q)}_{U|Y}|$:

$$q^*_{CE} = \arg\max_Q |\Sigma^{(Q)}_{U|Y}| \quad \text{subject to} \quad \Sigma^{(Q)} \succeq 0.$$

By maximizing the residual uncertainty of the sources, the MaxEnt objective forces the observations to be as unstructured and independent as possible once the target Y is known. This adversary implies that, conditioned on Y, we have independent U^1 and U^2.

Lemma 1 (Least-Informative Coupling via Maximum Conditional Entropy)**.** *Suppose the true data distribution p is generated by the model. The true model contains synergy because the shared nuisance factor z^c allows the views to cooperatively suppress shared noise, yielding a cleaner prediction of Y. Furthermore, because the independent task-relevant factors (z^{12}, z^1, z^2) all independently cause Y, they form a classical V-structure. While the private factors z^1 and z^2 are unconditionally independent, conditioning on their shared effect Y renders all task factors conditionally dependent via the "explaining away" effect. At the observation level, this structure induces a negative conditional correlation between the views, mathematically captured by $-\Sigma_{U^1Y}\Sigma^{-1}_{YY}\Sigma_{YU^2}$.*

To achieve conditional independence, the optimal adversary q^*_{CE} must cancel both of these mechanisms. Under the noise-injection framework, this is achieved by injecting the negative of the true conditional cross-covariance into the private noise:

$$C_{CE} = -\Sigma^{(p)}_{U^1U^2|Y}.$$

Proof. The objective is to maximize the determinant of the joint conditional covariance matrix $|\Sigma^{(Q)}_{U|Y}|$. By the properties of multivariate Gaussians, $\Sigma^{(Q)}_{U|Y}$ is given by the Schur complement of the target variance Σ_{YY}. Since Q is formed by injecting C into the off-diagonal of the observations, the conditional covariance under Q is the true conditional covariance plus the injected noise block:

$$\Sigma^{(Q)}_{U|Y} = \Sigma^{(Q)}_{UU} - \Sigma_{UY}\Sigma^{-1}_{YY}\Sigma_{YU} = \Sigma^{(p)}_{U|Y} + \begin{bmatrix} \mathbf{0} & C \\ C^\top & \mathbf{0} \end{bmatrix} = \begin{bmatrix} \Sigma^{(p)}_{U^1|Y} & \Sigma^{(p)}_{U^1U^2|Y} + C \\ \Sigma^{(p)}_{U^2U^1|Y} + C^\top & \Sigma^{(p)}_{U^2|Y} \end{bmatrix}.$$

By Hadamard's inequality (Cover & Thomas, 2006, Theorem 17.9.4), the determinant of a block matrix with fixed diagonal blocks is maximized when the off-diagonal blocks are zero. Therefore, to maximize $|\Sigma^{(Q)}_{U|Y}|$, the adversary must set $C_{CE} = -\Sigma^{(p)}_{U^1U^2|Y}$. For jointly Gaussian variables, a zero conditional cross-covariance implies conditional independence, achieving $U^1 \perp\!\!\!\perp U^2 \mid Y$. Furthermore, because this choice makes $\Sigma^{(Q)}_{U|Y}$ a block diagonal matrix whose diagonal blocks are the valid marginals from p, the conditional covariance $\Sigma^{(Q)}_{U|Y}$ is positive semi-definite (PSD). By the properties of the Schur complement, a joint block covariance matrix is PSD if and only if its lower-right block (Σ_{YY}) and its Schur complement ($\Sigma^{(Q)}_{U|Y}$) are both PSD. Since the target variance $\Sigma_{YY} \succ 0$ is fixed by the constraint space, and our choice of C_{CE} ensures $\Sigma^{(Q)}_{U|Y} \succeq 0$, the resulting full joint covariance matrix $\Sigma^{(Q)}$ is guaranteed to be a valid PSD matrix. This confirms that q^*_{CE} is a feasible point in the constraint space Δ_P. □

Remark 1 (The Three Mechanisms of Synergy: Suppression, Averaging, and Explaining-Away)**.** *This algebraic solution reveals how q^*_{CE} removes the sources of multi-view synergy. By substituting the generative components, the true conditional cross-covariance decomposes into three terms, each corresponding to a synergistic mechanism:*

$$\Sigma^{(p)}_{U^1U^2|Y} = \underbrace{\Lambda^c_1(\Lambda^c_2)^\top}_{\text{1. Shared Nuisance}} + \underbrace{\Lambda^{12}_1(\Lambda^{12}_2)^\top}_{\text{2. True Shared Signal}} - \underbrace{\Sigma_{U^1Y}\Sigma^{-1}_{YY}\Sigma_{YU^2}}_{\text{3. Induced V-Structure Effect}}.$$

By injecting $C_{CE} = -\Sigma^{(p)}_{U^1U^2|Y}$ into the noise, the adversary acts as a filter that cancels all three mechanisms:

1. ***Suppression (Driven by z^c):*** *When the views share a nuisance factor z^c, the optimal joint regression linearly cross-references the views to subtract and cancel out the shared interference. Consider a toy system where $U^1 = Y + z^c$ and $U^2 = z^c$. Marginally, U^2 is uninformative about Y. However, a joint regression model computing the optimal linear estimator ($W = \Sigma_{UU}^{-1}\Sigma_{UY}$) assigns a negative weight to U^2. This allows the model to subtract the shared noise, computing $U^1 - U^2 = Y$ and recovering the target. q^*_{CE} eliminates this mechanism by injecting $-\Lambda_1^c(\Lambda_2^c)^\top$ to decouple z^c across the views, turning it into independent private noise so it can no longer be subtracted.*

2. ***Averaging (Driven by z^{12}):*** *Because both views observe the true shared signal z^{12} corrupted by independent private noise, they provide redundant measurements. For example, if $U^1 = Y + \epsilon^1$ and $U^2 = Y + \epsilon^2$ (with $\epsilon^1 \perp\!\!\!\perp \epsilon^2$), the optimal joint estimator applies positive weights to both views to compute their average. This averaging boosts the signal-to-noise ratio by suppressing the uncorrelated private noises, yielding a cleaner prediction of Y than either view could individually. q^*_{CE} breaks this capability by injecting $-\Lambda_1^{12}(\Lambda_2^{12})^\top$, which cancels the true shared signal correlation across the views, destroying the redundancy that enabled the averaging.*

3. ***V-Structure Exploitation (Driven by z^{12}, z^1, z^2):*** *Because the independent task-relevant factors all independently cause the target Y, they form a classical joint V-structure. Observing Y renders all task factors conditionally dependent via the "explaining away" effect, which a joint model exploits. q^*_{CE} breaks this cooperative dependence because its injected noise ($+\Sigma_{U^1Y}\Sigma_{YY}^{-1}\Sigma_{YU^2}$) cancels the induced negative explaining-away correlation.*

Together, these mechanisms dictate that any latent structure—whether shared task-relevant, shared nuisance, or conditionally induced by a V-structure—provides an opportunity for synergistic predictive power. The baseline distribution q^*_{CE} eliminates them all.

B.3 The Minimum Mutual Information Adversary (q^*_{MI})

The original PID framework (Bertschinger et al., 2014) defines its adversary q^*_{MI} through an alternative objective: to find the distribution that minimizes the joint mutual information $I_Q(Y;U)$. Because $H(Y)$ is fixed across Δ_P, this is equivalent to maximizing the residual generalized variance $|\Sigma^{(Q)}_{Y|U}|$:

$$q^*_{MI} = \arg\max_{Q \in \Delta_P} |\Sigma^{(Q)}_{Y|U}|.$$

The connection between the two adversarial objectives is governed by the determinant identity:

$$|\Sigma^{(Q)}_{Y|U}| = |\Sigma_{YY}| \frac{|\Sigma^{(Q)}_{U|Y}|}{|\Sigma^{(Q)}_{UU}|}.$$

This identity follows directly from factoring the determinant of the joint covariance matrix $\Sigma^{(Q)}$ in two equivalent ways via its Schur complements: $|\Sigma^{(Q)}| = |\Sigma_{YY}||\Sigma^{(Q)}_{U|Y}| = |\Sigma^{(Q)}_{UU}||\Sigma^{(Q)}_{Y|U}|$. To minimize the mutual information, q^*_{MI} manipulates the cross-covariance matrix C to reduce the denominator $|\Sigma^{(Q)}_{UU}|$. Mathematically, injecting correlation reduces the generalized variance of a joint distribution. Because the marginal blocks $\Sigma_{U^1U^1}$ and $\Sigma_{U^2U^2}$ are fixed by Δ_P, the joint determinant factors via the Schur complement as $|\Sigma^{(Q)}_{UU}| = |\Sigma_{U^1U^1}||\Sigma_{U^2U^2} - \Sigma^{(Q)}_{U^2U^1}\Sigma^{-1}_{U^1U^1}\Sigma^{(Q)}_{U^1U^2}|$. Injecting a cross-covariance subtracts a positive semi-definite matrix from the second term, thereby reducing the overall determinant. q^*_{MI} exploits this property by injecting positive correlation between the private noises across the views. This compresses the joint noise distribution to align with the signal, maximizing the residual target variance.

Like q^*_{CE}, q^*_{MI} must first decouple the response-irrelevant nuisance factor z^c. If z^c remains shared across the views, the optimal estimator will exploit this correlation to subtract the views and cancel the nuisance interference (the suppression mechanism). To prevent this cooperative cancellation and minimize the mutual information $I_Q(Y;U)$, q^*_{MI} adopts the same strategy as q^*_{CE} within the nuisance subspace: it injects negative correlation to decouple z^c, turning it into independent private noise.

However, the two adversaries diverge fundamentally in how they treat the task-relevant subspace. The optimization objective of q^*_{CE} yields a clean, term-by-term matrix decomposition that explicitly decouples the task variables to achieve conditional independence ($I_{q^*_{CE}}(U^1; U^2 \mid Y) = 0$). In contrast, the objective of q^*_{MI} mathematically couples these variables together through the inverse covariance matrix. Because this coupling prevents a simple discrete factorization for the task subspace, the adversary must instead holistically inject positive correlation to align the joint noise ellipse with the combined signal direction. While this alignment effectively hinders the regression model, it inevitably introduces a conditional dependence between the views. Thus, the optimal mutual information adversary sacrifices conditional independence ($I_{q^*_{MI}}(U^1; U^2 \mid Y) > 0$) to achieve its objective.

To see this mechanically, consider a simplified scalar system where the private task-relevant factors z^1, z^2 are absent. The target is defined purely by the shared factor $Y = z^{12}$, with $z^{12} \sim \mathcal{N}(0, 1)$. The observations are $U = \Lambda z^{12} + n$, where the stacked vector $\Lambda = \begin{bmatrix}\lambda_1 \\ \lambda_2\end{bmatrix}$ defines the specific 1-dimensional "signal direction". Here, $n \sim \mathcal{N}(\mathbf{0}, \Sigma_N^{(p)})$ represents the sum of the full-rank private noise and the completely decoupled nuisance factors. Under the noise-injection framework, the adversary injects its cross-covariance c exclusively into this noise block, resulting in the total adversarial noise covariance $\Sigma_N^{(Q)} = \begin{bmatrix}v_1 & c \\ c & v_2\end{bmatrix}$.

The joint mutual information is minimized when $|\Sigma^{(Q)}_{Y|U}|$ is maximized. Under our explicit model, $\Sigma_{YY} = 1$, $\Sigma_{YU} = \Lambda^\top$, and $\Sigma^{(Q)}_{UU} = \Lambda\Lambda^\top + \Sigma_N^{(Q)}$. Substituting these into the residual variance yields:

$$|\Sigma^{(Q)}_{Y|U}| = 1 - \Lambda^\top \left(\Lambda\Lambda^\top + \Sigma_N^{(Q)}\right)^{-1} \Lambda.$$

To simplify the inverted term, we apply the Woodbury matrix identity, $(\mathbf{A} + \mathbf{UCV})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{U}(\mathbf{C}^{-1} + \mathbf{VA}^{-1}\mathbf{U})^{-1}\mathbf{VA}^{-1}$, setting $\mathbf{A} = \Sigma_N^{(Q)}$, $\mathbf{U} = \Lambda$, $\mathbf{V} = \Lambda^\top$, and $\mathbf{C} = 1$. Letting $K(c) = \Lambda^\top(\Sigma_N^{(Q)})^{-1}\Lambda$ represent the Signal-to-Noise Ratio (SNR) penalty, the inverse becomes:

$$\left(\Sigma_N^{(Q)} + \Lambda\Lambda^\top\right)^{-1} = (\Sigma_N^{(Q)})^{-1} - (\Sigma_N^{(Q)})^{-1}\Lambda\left(1 + K(c)\right)^{-1}\Lambda^\top(\Sigma_N^{(Q)})^{-1}.$$

Substituting this expanded inverse back into the residual variance equation yields a step-by-step algebraic reduction:

$$\begin{aligned}
|\Sigma^{(Q)}_{Y|U}| &= 1 - \Lambda^\top\left[(\Sigma_N^{(Q)})^{-1} - (\Sigma_N^{(Q)})^{-1}\Lambda\left(1 + K(c)\right)^{-1}\Lambda^\top(\Sigma_N^{(Q)})^{-1}\right]\Lambda \\
&= 1 - \underbrace{\Lambda^\top(\Sigma_N^{(Q)})^{-1}\Lambda}_{K(c)} + \underbrace{\Lambda^\top(\Sigma_N^{(Q)})^{-1}\Lambda}_{K(c)}\left(1 + K(c)\right)^{-1}\underbrace{\Lambda^\top(\Sigma_N^{(Q)})^{-1}\Lambda}_{K(c)} \\
&= 1 - K(c) + \frac{K(c)^2}{1 + K(c)} \\
&= \frac{1}{1 + K(c)}.
\end{aligned}$$

Thus, maximizing the residual variance is exactly equivalent to minimizing the SNR penalty term $K(c)$:

$$K(c) = \begin{bmatrix}\lambda_1 & \lambda_2\end{bmatrix}\begin{bmatrix}v_1 & c \\ c & v_2\end{bmatrix}^{-1}\begin{bmatrix}\lambda_1 \\ \lambda_2\end{bmatrix} = \frac{\lambda_1^2 v_2 + \lambda_2^2 v_1 - 2c\lambda_1\lambda_2}{v_1 v_2 - c^2}.$$

By taking the derivative with respect to c and setting it to zero, we find the adversary optimally minimizes the SNR penalty at the roots $c = \lambda_1 v_2/\lambda_2$ and $c = \lambda_2 v_1/\lambda_1$. The true optimal adversarial correlation c^* is the unique root that satisfies the positive semi-definite covariance constraint $c^2 < v_1 v_2$. Because $v_1, v_2 > 0$, both roots strictly match the sign of the signal correlation $\lambda_1\lambda_2$, ensuring c^* always matches this sign as well. This sign-matching mathematically guarantees that the adversarial noise is geometrically aligned with the task signal. For instance, if the views have positively aligned signal loadings ($\lambda_1\lambda_2 > 0$), an optimal estimator would average the views to boost the signal while canceling independent noise. By injecting a positive noise correlation ($c^* > 0$),

the adversary ensures that the noise components constructively reinforce each other upon addition. Conversely, if the signal extraction requires subtraction, the adversary injects negative correlation to ensure the subtracted noises similarly reinforce. Consequently, any linear combination that attempts to boost the signal will simultaneously amplify the geometrically aligned adversarial noise, structurally preventing the model from achieving synergistic noise cancellation.

Remark 2 (The Mechanics of Adversarial Correlation)**.** *The noise-injection framework captures the distinction between the adversaries: while both counteract cooperative nuisance suppression (i.e., injecting negative correlation to render the nuisance factor z^c independent), their treatment of the task-relevant predictive subspace differs.*

To achieve conditional independence, q^*_{CE} acts as a decoupling filter: it injects the negative conditional cross-covariance $C_{CE} = -\Sigma^{(p)}_{U^1U^2|Y}$. In our scalar example, this corresponds to setting $c = 0$, making the noise conditionally independent so the regression model could average the views to extract a cleaner signal.

In contrast, q^*_{MI} goes further to minimize mutual information. Rather than just canceling the conditional correlation, it injects a positive correlation C_{MI}. By stretching and rotating the Gaussian noise ellipse $\Sigma^{(Q)}_N$ so that it mimics the signal direction Λ, the adversary ensures that any linear estimator weight vector w attempting to extract the signal (maximizing the projection $w^\top\Lambda$) simultaneously captures adversarial noise variance ($w^\top\Sigma^{(Q)}_N w$). Mathematically, this minimizes the maximum achievable signal-to-noise ratio ($K(c) = \max_w \frac{(w^\top\Lambda)^2}{w^\top\Sigma^{(Q)}_N w}$). Because the noise aligns with the signal, the estimator cannot project away the noise without also annihilating the signal itself, creating a structural dependency.

B.4 Shared Properties and Analytic Bounds

Despite their differing mathematical objectives, both q^*_{CE} and q^*_{MI} are constrained within Δ_P. Because of this shared constraint space, they share structural properties that define the PID decomposition.

Remark 3 (The Marginal Preservation)**.** *Because any valid adversary $Q \in \Delta_P$ matches the true predictor-target marginals ($I_Q(U^j;Y) = I_P(U^j;Y)$), the views maintain their individual marginal predictive relationships with Y even when the shared task factor z^{12} is adversarially corrupted. By the chain rule, $I_Q(U^1,U^2;Y) = I_P(U^{-j};Y) + I_Q(U^j;Y \mid U^{-j})$. Since the marginal term is fixed, minimizing the joint information is identical to minimizing the conditional mutual information. Thus, because q^*_{CE} is a sub-optimal minimizer of joint information ($I_{q^*_{CE}} \geq I_{q^*_{MI}}$), its resulting conditional mutual information is larger, acting as an upper bound on the true unique information.*

Remark 4 (Analytic Bounds via the Tractable Proxy)**.** *Because q^*_{CE} provides a closed-form but sub-optimal minimization of the joint information ($I_{q^*_{MI}} \leq I_{q^*_{CE}}$), substituting it into the PID equations yields analytic bounds on the true optimal quantities. Specifically, it underestimates Synergy ($S_{CE} \leq S_{MI}$) and Redundancy ($R_{CE} \leq R_{MI}$), while overestimating Unique Information ($U^{CE}_j \geq U^{MI}_j$).*

Finally, this formulation reveals that conditional mutual information is distinct from unique information. By the chain rule, $I_P(U^1;Y \mid U^2) = \text{Unique}_1 + \text{Synergy}$. By removing the synergy, the adversaries reduce the conditional mutual information to its unique information component.

B.5 Bridge to Multi-View Learning

This framework provides a theoretical bridge to classic multi-view learning. Classic multi-view algorithms typically invoke two distinct premises that are often bundled together: the *redundancy assumption* (that the views provide identical predictive information, meaning zero unique information), which is relied upon by modern contrastive and bottleneck methods (Federici et al., 2020; Tosh et al., 2021), and the *conditional independence assumption* (that $U^1 \perp\!\!\!\perp U^2 \mid Y$), which is foundational to classic algorithms like Co-training (Blum & Mitchell, 1998) and CCA dimension reduction (Chaudhuri et al., 2009). Our baseline distribution q^*_{CE} is the mathematical embodiment of the latter. Conditional independence does *not* imply redundancy. Because q^*_{CE} preserves the true marginals, it allows for amounts of unique predictive information; it merely forbids the views from

interacting synergistically. By measuring our tractable proxy S_{CE}, we approximate the true synergy and directly quantify how much the true data violates the conditional independence assumption. When $S_{CE} \gg 0$, the views cooperate to cancel noise, indicating that algorithms assuming conditional independence will be suboptimal compared to joint models that can exploit synergistic cancellation.