Information-Theoretic Decomposition for Multimodal Interaction Learning

Supplementary Material

In the Supplementary Material, we first present the proofs and detailed analysis for the information-based interaction bound and the decomposition architecture in Appendix A. Specifically, we provide the theoretical proof for Theorem 1 to clarify the significance of interaction subsection A.1. And we provide the variational explanation for the interaction decomposition architecture subsection A.3. Next, we describe the experimental setup and model architecture in greater detail subsection B.1. Finally, we conduct expanded evaluations across diverse domains with varying modalities and backbones to further validate the applicability of our DMIL method subsection B.3.

A. Proof and Analysis

A.1. Proof for Theorem 1

Theorem 1. *Let C be the interaction composition random variable with realization c, Z be the learned representation from a multimodal input X, and Y be the target variable. The mutual information $I(Z;Y)$ is lower-bounded as follows:*

$$I(Z;Y) \geq \mathbb{E}_c[I(Z;Y|c)] - H(C|Z) + I(Y;C), \quad (1)$$

where $H(C|Z)$ is the conditional entropy of C given Z.

Proof. The proof aims to establish a lower bound for $I(Z;Y)$ by analyzing the difference between the total mutual information and the expected conditional mutual information, a quantity we denote as Δ. This term, $\Delta = I(Z;Y) - \mathbb{E}_c[I(Z;Y|c)]$, quantifies the information synergy; that is, how much information about Y is revealed in Z due to knowledge of the interaction context C.

Our derivation rests on two key assumptions regarding the data generation process:

1. **Deterministic Encoding**: The representation Z is generated by a deterministic encoder ϕ from the multimodal input X, i.e., $Z = \phi(X)$. This implies that given X, there is no uncertainty about Z, so the conditional entropy $H(Z|X) = 0$. This establishes the Markov chain $Y \to X \to Z$.
2. **Deterministic Interaction**: The interaction variable C is a deterministic function of the input X and the target Y, i.e., $C = f(X,Y)$. This means the conditional entropy $H(C|X,Y) = 0$. The variable C is designed to capture specific interaction patterns between modalities in X that are relevant for predicting Y.

We begin by expressing Δ using a standard information-theoretic identity. We have the following identity:

$$\begin{aligned}\Delta &= \mathbb{E}_{z,y} \log \frac{p(z,y)}{p(z)p(y)} - \mathbb{E}_{z,y,c} \log \frac{p(z,y|c)}{p(z|c)p(y|c)} \\ &= \mathbb{E}_{z,y,c} \log \frac{p(y|c)p(z|c)p(z,y)}{p(y)p(z)p(z,y|c)} \\ &= I(Y;C) + I(Z;C) - I(Z,Y;C).\end{aligned} \quad (2)$$

This identity follows from the chain rule of mutual information, as $I(Z,Y;C) = I(Z;C) + I(Y;C|Z)$ and $I(Z;Y) - I(Z;Y|C) = I(Y;C) - I(Y;C|Z)$.

Next, we apply the Data Processing Inequality (DPI). From our first assumption ($Z = \phi(X)$), the pair (Z,Y) is a function of the pair (X,Y). This implies the Markov chain $(X,Y) \to (Z,Y)$. The DPI states that post-processing cannot increase information, so for any variable C, we have:

$$I(Z,Y;C) \leq I(X,Y;C). \quad (3)$$

Substituting this into our expression for Δ yields the inequality:

$$\Delta \geq I(Y;C) + I(Z;C) - I(X,Y;C). \quad (4)$$

Now, we use our second assumption, that C is a deterministic function of X and Y. This means the conditional entropy $H(C|X,Y) = 0$. By the definition of mutual information:

$$I(X,Y;C) = H(C) - H(C|X,Y) = H(C) - 0 = H(C). \quad (5)$$

Replacing $I(X,Y;C)$ with $H(C)$ in our inequality for Δ, we get:

$$\Delta \geq I(Y;C) + I(Z;C) - H(C). \quad (6)$$

To simplify further, we use the definition of mutual information for $I(Z;C)$:

$$I(Z;C) = H(C) - H(C|Z). \quad (7)$$

Rearranging this gives $I(Z;C) - H(C) = -H(C|Z)$. Substituting this into the inequality for Δ:

$$\Delta \geq I(Y;C) - H(C|Z). \quad (8)$$

Finally, by substituting back the definition of Δ, we arrive at the desired lower bound:

$$I(Z;Y) - \mathbb{E}_c[I(Z;Y|c)] \geq I(Y;C) - H(C|Z), \quad (9)$$

which can be rearranged to conclude the proof:

$$I(Z;Y) \geq \mathbb{E}_c[I(Z;Y|c)] - H(C|Z) + I(Y;C). \quad (10)$$

□

A.2. Analysis for Theorem 1

Theorem 1 establishes a lower bound on the learned multimodal information $I(Z;Y)$. As derived in its proof, this bound is given by:

$$I(Z;Y) \geq \mathbb{E}_c[I(Z;Y|c)] - H(C|Z) + I(Y;C).$$

To maximize this lower bound and thus enhance the overall multimodal information captured by the model, we must consider the roles of its constituent terms. The term $I(Y;C)$ is determined by the intrinsic data distribution and represents the total information about the target Y contained in the interaction composition C. The other two terms, $\mathbb{E}_c[I(Z;Y|c)]$ and $H(C|Z)$, are directly influenced by the learned representation Z.

Firstly, $\mathbb{E}_c[I(Z;Y|c)]$ represents the expected conditional mutual information between the representation Z and the target Y, given the interaction composition c. This term quantifies how well the model captures task-relevant information for specific interaction patterns. Maximizing this term implies that the model effectively learns information across diverse interaction dynamics, which is crucial for robust multimodal learning.

Secondly, $H(C|Z)$ is the conditional entropy of the interaction variable C given the learned representation Z. This term measures the uncertainty remaining about the interaction composition C after observing Z. A lower $H(C|Z)$ indicates that Z is a better predictor of C, meaning the representation effectively captures the underlying interaction patterns. According to the Data Processing Inequality, $H(C|Z) \geq H(C|X)$, which means $H(C|X)$ serves as a fundamental lower bound for $H(C|Z)$, representing the inherent uncertainty of interaction given the raw data. To maximize the overall multimodal information $I(Z;Y)$, it is crucial for the representation Z to minimize $H(C|Z)$, thereby learning to predict the intrinsic interaction compositions C as accurately as possible.

In summary, the proof of Theorem 1 underscores that the ability of multimodal representations to effectively capture information across diverse interaction types (maximizing $\mathbb{E}_c[I(Z;Y|c)]$) and to accurately predict these intrinsic interaction patterns (minimizing $H(C|Z)$) are critical factors for achieving comprehensive and high-quality multimodal information learning.

A.3. Explanation of decomposition

A.3.1. Intra-Modal Decomposition

Here, we provide a detailed explanation of how our proposed decomposition framework operates. The core objective is to decompose a given representation Z into two latent components, N and M. The decomposition aims to satisfy two primary conditions: (1) the components N and M should be statistically independent, minimizing their mutual information $I(N;M)$, and (2) they should collectively preserve the information content of the original representation Z. To achieve this, we employ a variational approach.

We begin by deriving the Evidence Lower Bound (ELBO) for the marginal log-likelihood of Z, $\log p(z)$. We assume that z is generated from latent variables n and m drawn from independent priors, i.e., $p(n,m) = p(n)p(m)$. We introduce a factorized variational posterior $q(n,m|z) = q(n|z)q(m|z)$ to approximate the true posterior $p(n,m|z)$. By applying Jensen's inequality, we obtain the ELBO:

$$\begin{aligned}
\log p(z) &= \log \int p(z|n,m)p(n)p(m)\,dv\,dt \\
&\geq \mathbb{E}_{q(n|z),q(m|z)}\left[\log\left(\frac{p(z|n,m)p(n)p(m)}{q(n|z)q(m|z)}\right)\right] \\
&= \mathbb{E}_{q(n|z),q(m|z)}[\log p(z|n,m)] \\
&\quad + \mathbb{E}_{q(n|z)}\left[\log\frac{p(n)}{q(n|z)}\right] + \mathbb{E}_{q(m|z)}\left[\log\frac{p(m)}{q(m|z)}\right] \\
&= \mathbb{E}_{q(n|z),q(m|z)}[\log p(z|n,m)] \\
&\quad - KL(q(n|z)||p(n)) - KL(q(m|z)||p(m)).
\end{aligned} \tag{11}$$

Maximizing this lower bound involves optimizing three terms. The first term, $\mathbb{E}_{q(n|z),q(m|z)}[\log p(z|n,m)]$, is the reconstruction term, which encourages the latent components N and M to accurately reconstruct the original representation Z. The other two terms are regularization terms that minimize the Kullback-Leibler (KL) divergence between the variational posteriors ($q(n|z)$, $q(m|z)$) and their respective priors ($p(n)$, $p(m)$). Typically, the priors are chosen to be standard Gaussian distributions, $\mathcal{N}(0,I)$, to encourage well-structured latent spaces.

To understand why this decomposition architecture achieves disentanglement, we can analyze its connection to information theory. The goal is to decompose a representation Z into two statistically independent components, N and M. This can be framed as an optimization problem aimed at minimizing their mutual information, $I(N;M)$, while ensuring that N and M collectively retain the information from Z.

From an information-theoretic standpoint, minimizing $I(N;M)$ is equivalent to maximizing the negative interaction information, $I(Z;M,N) - I(Z;M) - I(Z;N)$. This objective encourages the joint representation (N,M) to be predictive of Z (maximizing $I(Z;M,N)$) while penalizing the information that N and M individually share with Z (minimizing $I(Z;M)$ and $I(Z;N)$). This process effectively isolates the unique contributions of N and M and minimizes their redundant overlap.

This equivalence relies on the assumption that N and M are conditionally independent given Z, i.e., $p(n,m|z) = p(n|z)p(m|z)$. This is a natural assumption for a model designed to decompose Z into distinct factors and is explicitly enforced in our variational framework by the factorized

posterior $q(n, m|z) = q(n|z)q(m|z)$. The derivation is as follows:

$$\begin{aligned} & I(Z; M, N) - I(Z; M) - I(Z; N) \\ &= \mathbb{E}_{p(z,v,m)} \left[\log \frac{p(z, m, n)p(z)p(m)p(n)}{p(m, n)p(z, m)p(z, v)} \right] \\ &= \mathbb{E}_{p(z,v,m)} \left[\log \frac{p(m, n|z)p(m)p(n)}{p(m, n)p(m|z)p(n|z)} \right] \\ &= \mathbb{E}_{p(z,v,m)} \left[\log \frac{p(m|z)p(n|z)p(m)p(n)}{p(m, n)p(m|z)p(n|z)} \right] \\ &= \mathbb{E}_{p(m,n)} \left[\log \frac{p(m)p(n)}{p(m, n)} \right] = -I(M; N). \end{aligned} \tag{12}$$

Thus, maximizing the interaction information objective is equivalent to minimizing the mutual information $I(M; N)$. We now show that maximizing the ELBO from Equation 11 aligns with this information-theoretic objective. The ELBO can be expressed as:

$$\begin{aligned} \mathcal{L}_{\text{ELBO}} = & \underbrace{\mathbb{E}_{q(n|z),q(m|z)}[\log p(z|n, m)]}_{\text{Reconstruction}} \\ & - \underbrace{KL(q(n|z)||p(n))}_{\text{Regularization for } N} \\ & - \underbrace{KL(q(m|z)||p(m))}_{\text{Regularization for } M}. \end{aligned} \tag{13}$$

Each term in the ELBO corresponds to a term in our information-theoretic objective:

1. **Reconstruction Term**: Maximizing the reconstruction term, $\mathbb{E}[\log p(z|n, m)]$, is equivalent to minimizing the conditional entropy $H(Z|V, M)$. Since $I(Z; V, M) = H(Z) - H(Z|V, M)$, this term effectively maximizes the joint mutual information $I(Z; V, M)$, ensuring that the latent components collectively preserve information about Z.
2. **Regularization Terms**: The KL divergence terms serve as variational upper bounds on the mutual information between the representation and the latent components. Specifically, we have:

$$\begin{aligned} I(Z; M) &= \mathbb{E}_{p(z,m)} \left[\log \frac{p(m|z)}{p(m)} \right] \\ &\leq \mathbb{E}_{p(z)}[KL(q(m|z)||p(m))], \\ I(Z; N) &= \mathbb{E}_{p(z,v)} \left[\log \frac{p(n|z)}{p(n)} \right] \\ &\leq \mathbb{E}_{p(z)}[KL(q(n|z)||p(n))]. \end{aligned} \tag{14}$$

   Maximizing the ELBO involves minimizing these KL terms, which in turn minimizes the upper bounds on $I(Z; M)$ and $I(Z; N)$.

Therefore, maximizing the ELBO in Equation 11 implicitly encourages maximizing $I(Z; V, M)$ while minimizing $I(Z; M)$ and $I(Z; N)$. This aligns directly with the objective of maximizing $I(Z; M, N) - I(Z; M) - I(Z; N)$, which, as shown in Equation 12, is equivalent to minimizing $I(M; N)$. This connection demonstrates that our variational decomposition framework is principled and effectively promotes the learning of disentangled latent representations.

A.3.2. Consistency Decomposition

After obtaining the intra-modality feature $M^{(m)}$, we propose a consistency decomposition to separate it into a modality-specific vector $U^{(m)}$ and a consistency/shared vector R. The objective is to learn representations that capture shared information in R while isolating unique, modality-specific information in $U^{(m)}$. This is achieved by maximizing the following objective function:

$$\max \; 2I(M^{(1)}; M^{(2)}; R) - I(U^{(1)}; R) - I(U^{(2)}; R). \tag{15}$$

This objective encourages R to capture information common to both $M^{(1)}$ and $M^{(2)}$ (interaction information $I(M^{(1)}; M^{(2)}; R)$) while being independent of the unique components $U^{(1)}$ and $U^{(2)}$.

To better understand the optimization process, we can decompose this objective. By focusing on the components related to modality $M^{(1)}$, the objective can be rewritten into three interpretable terms. The derivation is as follows, assuming a symmetric treatment for $M^{(2)}$:

$$\begin{aligned} & I(M^{(1)}; M^{(2)}; R) - I(U^{(1)}; R) \\ &= I(M^{(1)}; R) + I(M^{(2)}; R) \\ &\quad - I(M^{(1)}, M^{(2)}; R) - I(U^{(1)}; R) \\ &= \underbrace{I(M^{(1)}; U^{(1)}, R)}_{\text{Reconstruction}} - \underbrace{I(M^{(1)}; U^{(1)})}_{\text{Compactness}} - \underbrace{I(M^{(2)}; R|M^{(1)})}_{\text{Redundancy}}. \end{aligned} \tag{16}$$

The last equation is similar to Equation 12. This decomposition reveals three key optimization goals:

1. **Maximizing Reconstruction**: The term $I(M^{(1)}; U^{(1)}, R)$ corresponds to a reconstruction objective. Maximizing it is equivalent to minimizing the conditional entropy $H(M^{(1)}|U^{(1)}, R)$, ensuring that the original feature $M^{(1)}$ can be accurately reconstructed from its specific component $U^{(1)}$ and the shared component R.
2. **Maximizing Compactness**: The term $-I(M^{(1)}; U^{(1)})$ encourages the specific representation $U^{(1)}$ to be a compact, minimal representation of the information in $M^{(1)}$, following the information bottleneck principle. This is analogous to the KL divergence regularization term in Equation 14.
3. **Minimizing Redundancy**: The term $-I(M^{(2)}; R|M^{(1)})$ aims to minimize the conditional mutual information between $M^{(2)}$ and R given $M^{(1)}$. This encourages R to only contain information that is shared between $M^{(1)}$

and $M^{(2)}$, effectively isolating the redundant (shared) information from the unique aspects of each modality.

The third term, the conditional mutual information $I(M^{(2)}; R|M^{(1)})$, is often intractable to compute directly because it requires evaluating $q(r|m^{(1)}) = \int q(r|m^{(1)}, m^{(2)})p(m^{(2)}|m^{(1)})dm^{(2)}$.

$$I(M^{(2)}; R|M^{(1)}) = \mathbb{E}_{p(m^{(1)},m^{(2)})q(r|m^{(1)},m^{(2)})}\left[\log \frac{q(r|m^{(1)}, m^{(2)})}{q(r|m^{(1)})}\right]. \tag{17}$$

To address this, we introduce a variational distribution $v_\phi(r|m^{(1)})$ to approximate the true posterior $q(r|m^{(1)})$. This allows us to derive a tractable variational upper bound on this term:

$$\begin{aligned}
&\mathbb{E}_{p(m^{(1)},m^{(2)})q(r|m^{(1)},m^{(2)})}\left[\log \frac{q(r|m^{(1)}, m^{(2)})v_\phi(r|m^{(1)})}{v_\phi(r|m^{(1)})q(r|m^{(1)})}\right] \\
&= \mathbb{E}_{p(m^{(1)},m^{(2)})}\left[\mathrm{KL}(q(r|m^{(1)}, m^{(2)})||v_\phi(r|m^{(1)}))\right] \\
&- \mathbb{E}_{p(m^{(1)})}\left[\mathrm{KL}(q(r|m^{(1)})||v_\phi(r|m^{(1)}))\right] \\
&\leq \mathbb{E}_{p(m^{(1)},m^{(2)})}\left[\mathrm{KL}(q(r|m^{(1)}, m^{(2)})||v_\phi(r|m^{(1)}))\right].
\end{aligned} \tag{18}$$

The inequality holds because the KL divergence is non-negative. By minimizing this upper bound, we effectively minimize $I(M^{(2)}; R|M^{(1)})$. This strategy allows for the practical decomposition of features into their unique and redundant components. In summary, the complete objective function $\mathcal{L}_\mathrm{c}$ is given by:

$$\begin{aligned}
\mathcal{L}_\mathrm{c} = &\underbrace{\mathbb{E}_{q(r,u^{(1)}|m^{(1)})p(m^{(2)}|m^{(1)})}[\log p(m^{(1)}|r, u^{(1)})]}_{\text{Reconstruction}} \\
&+ \underbrace{\mathbb{E}_{q(r,u^{(2)}|m^{(2)})p(m^{(1)}|m^{(2)})}[\log p(m^{(2)}|r, u^{(2)})]}_{\text{Reconstruction}} \\
&- \underbrace{KL((q(r|m^{(1)})||v(r)) + KL(q(r|m^{(2)})||v(r)))}_{\text{Regularization for } R} \\
&- \underbrace{KL(q(u^{(1)}|m^{(1)})||p(u^{(1)}))}_{\text{Regularization for } U^{(1)}} \\
&- \underbrace{KL(q(u^{(2)}|m^{(2)})||p(u^{(2)}))}_{\text{Regularization for } U^{(2)}}.
\end{aligned} \tag{19}$$

The final loss function is a composition of the Evidence Lower Bound (ELBO), denoted as $f(z, y) = \mathcal{L}_\mathrm{ELBO}$, and the consistency loss, $g(z, y) = \mathcal{L}_\mathrm{c}$.

B. Experiment

B.1. Experimental setting

Real-world Dataset **BRCA** [10]: A dataset for breast invasive carcinoma PAM50 subtype classification, including 875 samples across 5 classes. We utilize mRNA expression and DNA methylation modalities.

ROSMAP [6]: A dataset targeted at Alzheimer's Disease diagnosis, containing 351 samples spanning 2 classes. Consistent with BRCA, we utilize mRNA expression and DNA methylation modalities.

CREMA-D [4]: An emotion recognition dataset consisting of 7,442 video clips. It covers six emotional states (anger, happiness, sadness, neutrality, disgust, and fear) using Audio and Visual modalities.

Kinetic-Sounds (KS) [1]: A multimodal action recognition dataset containing 19,000 ten-second clips. It includes 31 human action classes selected from the Kinetics dataset, utilizing Audio and Visual modalities.

CMU-MOSEI [11]: A large-scale sentiment analysis dataset including 23,453 video segments with Audio, Visual, and Text modalities. In this work, we focus specifically on the Visual and Text modalities for a challenging three-way sentiment classification task (positive, negative, neutral).

UCF101 [9]: An action recognition benchmark containing 13,320 videos across 101 human action classes. We employ RGB and Optical Flow modalities.

UR-FUNNY [7]: A large-scale dataset for humor detection, including over 16,000 samples from TED talks. It integrates Text, Visual, and Acoustic modalities to capture diverse speakers and topics.

VGGSound [5]: A large-scale audio-visual dataset consisting of short clips from over 200,000 videos, designed to capture sound events in diverse acoustic environments.

Training Details All experiments used a batch size of 64. CNN-based models were optimized using SGD (momentum 0.9, weight decay 1e-4) with dataset-specific learning rates: 1e-2 (KS/UCF) and 4e-3 (CREMA-D). For the Transformer-based MOSEI experiments, we used Adam with a learning rate of 1e-4. Training for our proposed method proceeded in three stages: (1) training the intra-modal decomposition for 10 epochs; (2) freezing this module while training subsequent consistency decomposition for 5–10 epochs; and (3) unfreezing the decomposition module to jointly fine-tune the full model with a learning rate between 1e-4 and 1e-5.

Data Preprocessing For datasets containing videos, we extract frames at 1 fps. In the KS dataset, we uniformly sample 3 frames from each 10-frame clip as visual inputs. For CREMA-D, 1 frame is extracted from each video. In UCF101, we select 2 RGB frames and 5 optical flow frames per video. For VGGSound, 3 frames are used as visual inputs. Additionally, we conduct experiments using more frames for both the KS and CREMA-D datasets; see Section B.3.2 for details.

Table 1. Performance comparison on diverse tasks with various modality combinations. ACC: accuracy, mAP: mean average precision. Best results in each column are **bolded**.

Method	UR-FUNNY (V+T)		ROSMAP (mRNA+METH)		BRCA (mRNA+METH)	
	ACC	mAP	ACC	mAP	ACC	mAP
Joint	63.71	67.99	82.08	85.57	88.97	86.41
Ensemble	62.57	65.92	83.02	86.32	89.35	86.62
DMIL	**65.50**	**69.15**	**85.85**	**88.84**	**89.73**	**87.33**

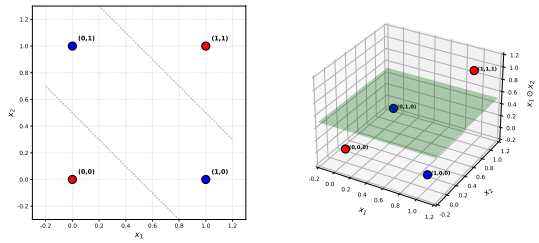

Figure 1. Illustration of the XOR data, which is a simple example of synergistic interaction. Left: The features (x_1, x_2) are not linearly separable, making it hard for a linear model to distinguish classes. Right: By adding the interaction term $(x_1 \cdot x_2)$, the data is separated in three-dimensional space, enabling simple linear separation.

B.2. Model architecture

Architecture Description. In this section, we detail the architecture of our Decomposition-based Multimodal Interaction Learning (DMIL) model, as illustrated in Figure 3. The DMIL framework incorporates two distinct decomposition modules. First, raw inputs from m modalities, denoted as $X^{(m)}$, are mapped into feature representations $Z^{(m)}$ via modality-specific encoders $\phi^{(m)}$. These features are subsequently decomposed to capture inter-modal interactions. Each decomposition module is built upon a Variational Autoencoder (VAE) framework, wherein the encoders—composed of Multi-Layer Perceptrons (MLPs)—predict the mean and variance of the latent distributions. Correspondingly, the decoders employ multilayer networks designed to minimize information loss during reconstruction. To extract the redundant component, we enforce feature alignment across modalities by minimizing the Kullback-Leibler (KL) divergence between their corresponding distributions.

Synergistic Interaction. The synergistic component is designed to capture information that is not present in any single modality but emerges solely through their integration. Therefore, designing an effective fusion mechanism is crucial. While simple concatenation is a common approach, it is often insufficient for modeling complex interactions. For instance, in a typical synergistic scenario such as the XOR problem, data points are not linearly separable in the original feature space. This non-linearity hinders the model's ability to learn synergistic interactions effectively. As shown in Figure 1 a, linear separation is unachievable with simple inputs. To address this, we introduce an element-wise product term. The formulation is defined as: $S = f_{\text{linear}}(n^{(1)}, n^{(2)}, n^{(1)} \odot n^{(2)})$, where $\odot$ denotes element-wise multiplication. The introduction of this multiplicative term simplifies the interaction modeling: it renders the XOR data linearly separable, as demonstrated in Figure 1 b. Furthermore, this product term explicitly captures high-order interactions, which are essential for defining synergy.

Table 2. Comparison of performance with richer temporal dynamics on CREMA-D (5 frames) and KS (8 frames) datasets.

Method	CREMA-D (5 Frames)		KS (8 Frames)	
	ACC	mAP	ACC	mAP
Joint	79.44	87.69	85.39	91.59
Ensemble	81.32	90.46	86.34	92.76
DMIL	**83.16**	**91.83**	**87.44**	**93.15**

B.3. Expanded evaluation on diverse domains

B.3.1. Generalization across modalities and tasks

To comprehensively validate the robustness of our method, we introduced three additional datasets: UR-FUNNY, ROSMAP, and BRCA. These were selected to cover distinct tasks and modality combinations beyond standard audio-visual benchmarks. UR-FUNNY focuses on humor detection using Visual (V) and Text (T) modalities, which are known to contain strong synergistic information [8]. ROSMAP and BRCA are biological datasets involving mRNA and DNA methylation (METH) modalities, presenting a different challenge in feature interaction.

The results are presented in Table 1. On the UR-FUNNY dataset, the Joint model outperforms the Ensemble baseline. This supports the idea that humor detection relies heavily on the synergy between vision and text, which Ensemble methods—designed to isolate modalities—often fail to cap-

Table 3. Performance comparison across different backbone architectures (CNN and Vision Transformer) on KS and VGG datasets.

Method	KS (CNN)		VGG (CNN)		KS (ViT)		VGG (ViT)	
	ACC	mAP	ACC	mAP	ACC	mAP	ACC	mAP
Joint	85.07	90.93	56.79	59.21	67.81	73.11	45.09	44.84
Ensemble	85.86	91.90	57.46	59.85	71.80	79.03	48.16	50.03
DMIL	**86.72**	**92.15**	**58.05**	**60.66**	**74.20**	**81.50**	**52.28**	**53.91**

ture. Conversely, on the biological datasets (ROSMAP and BRCA), the Ensemble model performs slightly better than the Joint model. Despite these variations in baseline performance, our DMIL method consistently achieves the highest accuracy and mAP across all datasets. This demonstrates that DMIL can adaptively learn interactions, whether the task requires capturing strong synergy (as in humor) or handling more independent modalities (as in biological analysis).

B.3.2. Impact of richer temporal dynamics

Previous studies highlight the importance of temporal dynamics in multimodal learning [2]. Therefore, we investigated how increased temporal resolution affects model performance. We conducted experiments on the CREMA-D and KS datasets using inputs with richer temporal information: 5 frames for CREMA-D and 8 frames for KS. The results are presented in Table 2. Compared to the baseline results in the main manuscript, CREMA-D shows significant improvement with added temporal information, whereas the gain on KS is more moderate. Notably, on CREMA-D, the ensemble model benefits more from the increased frame count than the joint model. This suggests that richer temporal dynamics strengthen the individual unimodal features, which the ensemble method captures more effectively. Crucially, our DMIL method continues to outperform these baselines, demonstrating its effectiveness with varying temporal dynamics.

B.3.3. Various backbone

Distinct backbone architectures process data differently, which affects how they capture and learn interactions. Following the initial experiments in Table 1, we conducted further validation on the KS and VGGSound datasets to examine the impact of architecture choice. Specifically, we introduced the widely used Vision Transformer (ViT) as the backbone for both modality encoders to provide a comparison with CNN-based models.

The results, presented in Table 3, indicate that the choice of backbone significantly influences baseline performance. We observed that ViT demonstrates weaker modeling capabilities for audio-visual understanding compared to CNN-based models. Additionally, the joint model marginally underperforms the ensemble model, particularly when using ViT. This is likely because audio-visual tasks are complex to learn, and ensemble methods are often more effective at preserving distinct uni-modal information. In contrast, our DMIL method consistently delivers superior performance across different backbones, demonstrating that it not only outperforms other approaches but also effectively learns interactions regardless of the architecture.

Table 4. Weights of different interaction types learned on the CREMA-D and KS datasets.

Interaction Type	CREMA-D	KS
Redundancy	0.7471	0.5732
Unique-Visual	0.0650	0.2093
Unique-Audio	0.1329	0.2096
Synergy	0.0549	0.0078

B.3.4. Interaction Weight

Our method employs an adaptive weighting mechanism to assign importance to different interaction components. Table 4 presents the average weights learned across samples for the CREMA-D and KS datasets. We observe that Redundancy is the predominant component in both datasets. Notably, on CREMA-D, the weight for Audio Uniqueness significantly surpasses that of Visual Uniqueness. This suggests that the audio modality possesses stronger discriminative power than the visual modality for this task. In contrast, KS shows a balanced distribution between audio and visual unique weights (0.2096 vs. 0.2093). These results demonstrate that our method effectively captures the intrinsic properties and primary information carriers of each dataset.

Interaction learning demonstration. While the previous section confirms DMIL's overall performance, it is crucial to verify that our proposed training strategy effectively learns the intended interaction components. To this end, we construct synthetic datasets from bit-wise features governed by logical relationships with the labels, including mixtures such as $1/2$ AND and $1/2$ XOR, $1/2$ OR and $1/2$ XOR, and $1/3$ AND, $1/3$ OR, and $1/3$ XOR. We compare the in-

Table 5. Validation of interaction decomposition on synthetic Boolean datasets. The table compares the estimated proportions (%) of Redundancy ($\tilde{R}$), Uniqueness ($\tilde{U}^{(1)}$,$\tilde{U}^{(2)}$), and Synergy ($\tilde{S}$) interaction for various logical combinations with the ground truth (Truth).

Method	AND				AND+XOR				OR+XOR				AND+OR+XOR			
	$\tilde{R}$	$\tilde{U}^{(1)}$	$\tilde{U}^{(2)}$	$\tilde{S}$	$\tilde{R}$	$\tilde{U}^{(1)}$	$\tilde{U}^{(2)}$	$\tilde{S}$	$\tilde{R}$	$\tilde{U}^{(1)}$	$\tilde{U}^{(2)}$	$\tilde{S}$	$\tilde{R}$	$\tilde{U}^{(1)}$	$\tilde{U}^{(2)}$	$\tilde{S}$
CVX	38.0	0.0	0.9	61.1	21.4	0.0	0.3	78.2	21.2	0.0	0.6	78.3	33.7	0.4	0.1	65.8
DMIL	36.4	0.0	1.7	61.9	19.0	0.4	0.0	80.6	20.7	0.0	1.9	77.4	26.8	4.1	0.0	69.1
Truth	38.3	0.0	0.0	61.7	19.1	0.0	0.0	80.9	19.1	0.0	0.0	80.9	25.5	0.0	0.0	74.5

teraction proportions estimated by DMIL against the CVX estimator [8], a specialized method for quantifying interaction quantity of Redundancy, Uniqueness, and Synergy. For DMIL, we derive these proportions from its decomposed feature components. The results in Table 5 show that although DMIL is not explicitly designed for quantification, it implicitly learns the interaction composition. Our framework's estimates are highly comparable to those of the CVX estimator, validating the effectiveness of our decomposition approach. Furthermore, in complex datasets mixing two or three interaction types, DMIL often achieves a closer approximation to the ground truth, demonstrating its robustness in capturing multifaceted interactions.

B.4. Synthetic dataset

In Section 4.3, we construct two types of synthetic data to facilitate our study. The first type is shown in Figure 1 of the main paper, where the data is crafted with pre-defined interactions to elucidate specific interaction dynamics. The second type derives from Boolean logic variables. Here, the interactions are inherently embedded within the Boolean logic itself, providing a general consideration for interaction analysis.

The data generation process for predefined interactions is executed in two sequential steps. Initially, the type of interaction for each sample is identified. We categorize potential interactions as *R*edundancy, *U*niqueness, and *S*ynergy for each dataset. As illustrated in the teaser figure in the introduction, each dataset is composed of samples exhibiting one or two types of interactions. The proportion of each interaction type is quantified using a fractional notation, such as $\frac{1}{4}U + \frac{3}{4}R$. This indicates that $\frac{1}{4}$ of the samples display **Unique** interactions, while the remaining samples demonstrate **Redundant** interactions.

In the second step, data corresponding to the predefined interactions is constructed. Different networks are employed to encode specific interactions into some dimensions of input space, which are then concatenated to form a comprehensive sample. If a sample is defined as a certain interaction, other types of interaction are suppressed by introducing Gaussian noise into their respective dimensions. This approach ensures that each sample exclusively embodies one type of interaction, thereby facilitating the construction of datasets with specified interaction properties.

The dataset, derived from Boolean logical data, is generated in a structured manner. Initially, the specific Boolean logic within each sample is determined. We consider two to three types of Boolean logics—*AND*, *OR*, and *XOR*—with each sample containing only one type. Each type of logic occupies an equivalent proportion within the dataset. Subsequently, these logics are encoded into the input space. Given that Boolean data inherently contains measurable interactions [3], we utilize this data to validate our method for interaction decomposition.

References

[1] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE international conference on computer vision*, pages 609–617, 2017. 4

[2] Arne Bernin, Larissa Müller, Sobin Ghose, Christos Grecos, Qi Wang, Ralf Jettke, Kai von Luck, and Florian Vogt. Automatic classification and shift detection of facial expressions in event-aware smart environments. In *Proceedings of the 11th PErvasive Technologies Related to Assistive Environments Conference*, pages 194–201, 2018. 6

[3] Nils Bertschinger, Johannes Rauh, Eckehard Olbrich, Jürgen Jost, and Nihat Ay. Quantifying unique information. *Entropy*, 16(4):2161–2183, 2014. 7

[4] Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5(4):377–390, 2014. 4

[5] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725. IEEE, 2020. 4

[6] Philip L De Jager, Yiyi Ma, Cristin McCabe, Jishu Xu, Badri N Vardarajan, Daniel Felsky, Hans-Ulrich Klein, Charles C White, Mette A Peters, Ben Lodgson, et al. A multi-omic atlas of the human frontal cortex for aging and alzheimer's disease research. *Scientific data*, 5(1):1–13, 2018. 4

[7] Md Kamrul Hasan, Wasifur Rahman, Amir Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, et al. Ur-funny: A multimodal language dataset for understanding humor. *arXiv preprint arXiv:1904.06618*, 2019. 4

[8] Paul Pu Liang, Chun Kai Ling, Yun Cheng, Alexander Obolenskiy, Yudong Liu, Rohan Pandey, Alex Wilf, Louis-Philippe Morency, and Russ Salakhutdinov. Quantifying interactions in semi-supervised multimodal learning: Guarantees

and applications. In *The Twelfth International Conference on Learning Representations*, 2023. 5, 7

[9] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 4

[10] Tongxin Wang, Wei Shao, Zhi Huang, Haixu Tang, Jie Zhang, Zhengming Ding, and Kun Huang. Moronet: multi-omics integration via graph convolutional networks for biomedical data classification. *BioRxiv*, pages 2020–07, 2020. 4

[11] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246, 2018. 4