

InfoBridge: Balanced Multimodal Integration through Conditional Dependency Modeling

Chenxin Li¹, Yifan Liu¹, Panwang Pan², Hengyu Liu¹, Xinyu Liu¹, Wuyang Li¹,
Cheng Wang¹, Weihao Yu¹, Yiyang Lin¹, Yixuan Yuan^{1*}

¹The Chinese University of Hong Kong, ²ByteDance Inc.

chenxinli@link.cuhk.edu.hk, yxyuan@ee.cuhk.edu.hk

Abstract

Developing systems that interpret diverse real-world signals remains a fundamental challenge in multimodal learning. Current approaches face significant obstacles from inherent modal heterogeneity. While existing methods attempt to enhance fusion through cross-modal alignment or interaction mechanisms, they often struggle to balance effective integration with preserving modality-specific information. We introduce InfoBridge, a novel framework grounded in conditional information maximization principles addressing these limitations. Our approach reframes multimodal fusion through two key innovations: (i) we formulate fusion as conditional mutual information optimization with integrated protective margin that simultaneously encourages cross-modal information sharing while safeguarding against over-fusion eliminating modal characteristics; and (ii) we enable fine-grained contextual fusion by leveraging modality-specific conditions to guide integration. Extensive evaluations across benchmarks demonstrate that InfoBridge consistently outperforms state-of-the-art multimodal architectures, establishing a principled approach that better captures complementary information across input signals. Project page: <https://cuhk-aim-group.github.io/InfoBridge/>.

1. Introduction

The integration of multimodal signals from sources such as vision, sound, touch, and smell provides a comprehensive viewpoint for understanding the external environment [36, 50, 59]. Mimicking human sensing, the ability to combine multimodal information is crucial towards creating proficient intelligent agent systems [6, 37, 42–44, 62]. Endeavors to utilize multimodal models have been applied across fields, including video classification [29, 54, 64],

event localization [65, 73], action recognition [28, 58], and audiovisual speech recognition [48, 51].

Despite attractive properties, empirical outcomes reveal that multimodal models do not deliver competitive performance compared with unimodal counterparts [5, 55]. While investigations [31, 78] have identified misalignment in modal learning due to heterogeneous data distribution, recent studies attempt to foster cross-modal alignment through approaches, such as aligning learning trajectory of features [22, 76], adaptively modulating gradients across modalities [55, 60], or employing attention maps and optimal transport [34, 45, 75, 79]. However, multimodal learning frameworks continue to grapple with challenges [16, 53, 55], particularly *over-fusion*, which we define as excessive merging of modality-specific representations that eliminates unique modal characteristics essential for understanding and reasoning.

These approaches frequently neglect the intrinsic trade-off between complementarity and fusibility across modalities. While complementary modalities offer unique information that enhances overall understanding, excessive alignment can destroy these unique characteristics. For instance, in audio-visual learning, visual information provides spatial details while audio captures temporal dynamics—over-fusion would eliminate these modality-specific strengths. Moreover, existing methods often adopt rigid alignment strategies that fail to account for contextual variations where different modalities should contribute differently based on the specific scenario.

As depicted in Fig. 1(a), this trade-off manifests as follows: while enhanced fusibility (easier merging) simplifies cross-modal learning, it simultaneously diminishes complementarity by forcing modalities to become more similar. This convergence, though computationally convenient, results in suboptimal utilization of the rich, diverse information that different modalities naturally provide [40, 41, 70–72]. Current research has attempted to address this through disentanglement techniques [12, 13, 57], but lacks a princi-

*Corresponding author.

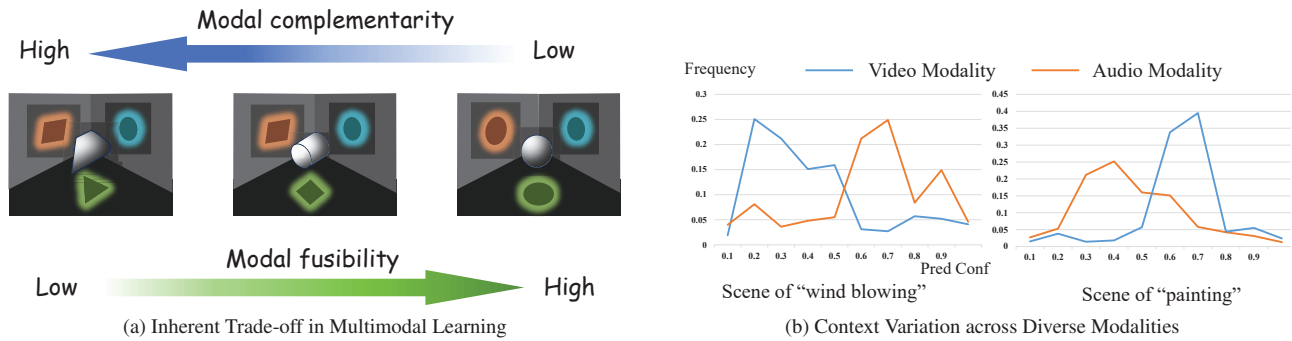


Figure 1. (a) Trade-off between modal complementarity and fusibility. As modal fusibility is enhanced, modal complementarity diminishes. (b) Visualization of prediction confidence by modalities (*video* and *audio*) at diverse scene contexts (*wind blowing* and *painting*).

pled information-theoretic framework to control this trade-off systematically.

The secondary impediment pertains to disregard for prior knowledge concerning hierarchical importance of modalities within varying contexts. Specifically, networks exhibit predilection for prioritizing modalities that facilitate ease in learning and scenes that are straightforward to interpret [19]. For instance, in “wind blowing” events, auditory modality is more effortlessly assimilated, whereas visual features predominantly elucidate *painting* events, as shown in Fig. 1(b). Overlooking such context priors curtails discriminative prowess of multimodal fusion models[12]. Consequently, our objective is incorporating prior knowledge of scenes and modalities while scrutinizing heterogeneous multimodal information, to yield granular features for effective fusion.

To address these challenges, we propose *InfoBridge*, an innovative framework based on maximizing conditional information across modalities. Our approach reframes multimodal fusion through information theory, explicitly optimizing cross-modal information while preserving unique modal characteristics. By incorporating contextual information as conditioning factors, we establish theoretical foundation that naturally maintains protective margin between aligned features. This design enables fine-grained control over fusion process, preventing excessive alignment from compromising modal uniqueness while ensuring effective information sharing. Through comprehensive theoretical analysis and empirical validation, we demonstrate that *InfoBridge* achieves favorable balance between fusion effectiveness and feature complementarity, advancing state-of-the-art in multimodal learning. In summary, our key contributions are as follows.

- We introduce *InfoBridge*, a novel framework addressing the fundamental trade-off between modal complementarity and fusibility through conditional cross-modal information maximization with protective margins preserving modality-specific characteristics while enabling fusion.

- We propose a context-aware fusion mechanism leveraging modality-specific conditions to guide fine-grained integration, demonstrating how contextual priors enhance multimodal learning by adapting to modal relationships across scenarios.
- Comprehensive experiments across diverse benchmarks and modalities validate our method’s superiority compared to existing ones, establishing a principled information-theoretic foundation for multimodal learning.

2. Related Work

Multimodal Models. Multimodal data like vision, sound, and text provide diverse and complementary information for an object. A plethora of prior studies develop algorithms capable of harnessing the vast array of multimodal data [2, 4, 25, 26, 32]. They mainly follow a basic pipeline of extracting unimodal representation initially, then aggregating modality-joint features for subsequent tasks [4, 16]. Furthermore, a considerable body of research is dedicated to exploring multimodal modeling within specific applications such as action recognition [21, 30, 49], audio-visual speech recognition [24, 56], and visual question answering [3, 27]. Nonetheless, these pioneering works tend to grapple with the challenge of sub-optimal fusion of heterogeneous multimodal representations, often resulting in imbalanced modality-specific learning with performance even inferior to unimodal models.

Multimodal Fusion with Disparate Modalities. Effective multimodal learning faces significant challenges due to inherent disparities between modalities. A primary obstacle is the misalignment of modality-specific embeddings during training, which severely impedes the utilization of multimodal data [63]. When these representation imbalances are not properly addressed, the fusion process yields sub-optimal outcomes [38]. Recent studies have revealed that multimodal training does not consistently outperform

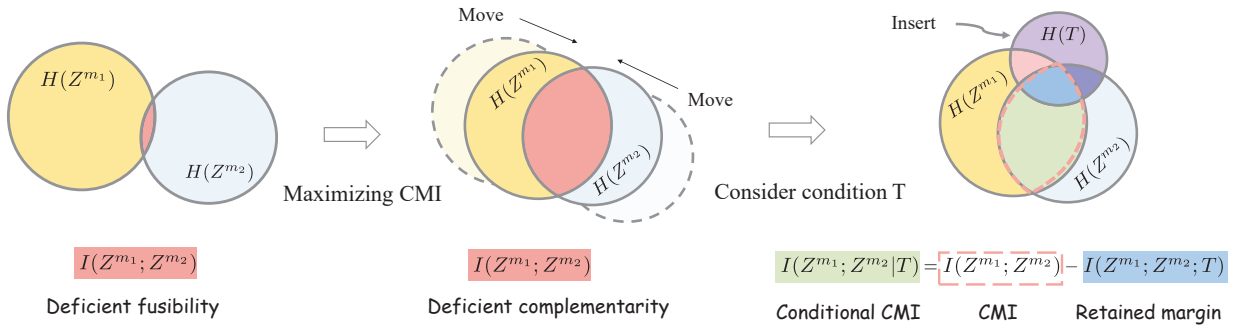


Figure 2. Overview of the proposed *InfoBridge* framework, specially constructed on the modal representation of m_1 and m_2 , as Z^{m_1} and Z^{m_2} . Contrary to the previous objective of maximizing the cross-modal information maximization (CMI), the proposed method maximizes the conditional CMI, thereby addressing the trade-off between modal fusibility and complementarity.

single-modal approaches [16, 69], with different modalities exhibiting distinct patterns in overfitting and generalization capabilities [69]. Moreover, modality dominance can significantly impair the learning process of other modalities [55]. While various mitigation strategies have been proposed, including modality-specific auxiliary classifiers [69], knowledge distillation frameworks [16], and adaptive learning rate mechanisms [74], the potential degradation of modal complementarity caused by excessive alignment remains largely unaddressed in current research.

Multimodal Learning with Information Theory. Recent advances demonstrate significant success through information theory for quantifying cross-modal relationships [14, 23, 39]. Pioneering work by Yang et al. [77] explores holistic multimodal interaction from information-theoretic perspective, establishing foundational insights into comprehensive cross-modal dependencies. Information bottleneck principles prove effective in compressing information while preserving task-relevant knowledge in multimodal settings [18, 33, 47], while information disentanglement techniques extract modality-specific knowledge for compact representations [12, 13, 57]. Our work builds upon these contributions by introducing conditional mutual information maximization with protective margins preventing over-fusion, context-aware conditioning adapting to varying modal relationships, and unified theoretical framework connecting contrastive learning with information-theoretic objectives for fusion control.

3. Proposed Method

In this section, we introduce *InfoBridge*, which is based on maximizing conditional mutual information (CMI) between modality-specific representations under a context condition. Our method not only boosts the alignment of shared information across modalities, but also regulates the excessive fusion by imposing a protective margin. The overall frame-

work is illustrated in Fig. 2.

Notation & Problem Setup. Assume we have a training dataset $\mathcal{D} = \left\{ \left\{ x_{[n]}^{(m)} \right\}_{m=1}^M, y_{[n]} \right\}_{n=1}^N$, where $x_{[n]}^{(m)}$ denotes the m th modality input of sample n and $y_{[n]}$ is the corresponding label. For each modality m , a unimodal encoder $F_{\psi^{(m)}} : \mathcal{X}^{(m)} \rightarrow \mathcal{Z}^{(m)}$ with parameters $\psi^{(m)}$ produces a representation z^m . The fusion operator $E_h : \left\{ \mathcal{Z}^{(m)} \right\}_{m=1}^M \rightarrow \hat{\mathcal{Z}}$ integrates the unimodal embeddings, where h denotes a critic or estimator function used in the mutual information estimation. In addition, let T denote an auxiliary context (e.g., category information) relevant to the modalities. Such a condition T is later incorporated to derive conditional cross-modal information bounds.

Motivation and Key Challenges. Multimodal learning aims to leverage complementary information across different sensory inputs. However, two significant challenges persist in existing approaches: (1) ensuring effective cross-modal alignment without destroying modality-specific characteristics, and (2) balancing the contribution of each modality to prevent one from dominating the fusion process. Prior works have approached these issues through various techniques including attention mechanisms and contrastive learning, but often lack theoretical guarantees on the information trade-off inherent in fusion. Our method addresses these challenges by introducing a context-conditioned information maximization framework with an explicit protective margin that preserves modality-specific information while ensuring optimal alignment.

Noise-Contrastive Objective and Mutual Information Bound. To encourage cross-modal alignment while preventing over-fusion, we employ a conditional noise-contrastive estimation (NCE) strategy that incorporates context information. Our approach differs from standard NCE by introducing a conditional variable T that provides task-relevant guidance for the alignment process.

Definition of Positive and Negative Pairs. We first

clarify the construction of training pairs: *Positive pairs* (z^{m_1}, z^{m_2}) are those where both representations are derived from the same sample, thus following the joint distribution $p(z^{m_1}, z^{m_2}|T)$. *Negative pairs* (z^{m_1}, z^{m_2}) are those where representations come from different samples, thus following the product of marginals $p(z^{m_1}|T) \cdot p(z^{m_2}|T)$. Let $J \in \{0, 1\}$ be a binary indicator where $J = 1$ denotes positive pairs and $J = 0$ denotes negative pairs. We use N_1 to denote the number of positive pairs and N_0 for negative pairs in our contrastive learning setup.

NCE Objective with Context. Given the context condition T , we define our conditional NCE loss. The critic function $h : \mathcal{Z}^{m_1} \times \mathcal{Z}^{m_2} \times \mathcal{T} \rightarrow [0, 1]$ estimates the probability that a pair (z^{m_1}, z^{m_2}) is positive given context T :

$$\begin{aligned} \mathcal{L}_{\text{NCE}} &= \mathbb{E}_{(z^{m_1}, z^{m_2}) \sim p_{\text{pos}}} [\log h(z^{m_1}, z^{m_2}, T)] \\ &+ \frac{N_0}{N_1} \mathbb{E}_{(z^{m_1}, z^{m_2}) \sim p_{\text{neg}}} [\log(1 - h(z^{m_1}, z^{m_2}, T))], \end{aligned} \quad (1)$$

where p_{pos} represents positive pairs and p_{neg} represents negative pairs. This formulation shows how positive and negative pairs contribute to learning. This posterior probability connects the classification task (distinguishing positive from negative pairs) to the information-theoretic objective of maximizing conditional mutual information. The term $\frac{N_0}{N_1}$ represents the ratio of negative to positive samples and serves as implicit regularization.

Prior Probabilities and Information-theoretic Connection. To connect our NCE objective to conditional mutual information, we define prior probabilities for the binary variable J :

$$p(J = 1|T) = \frac{N_1}{N_0 + N_1}, \quad p(J = 0|T) = \frac{N_0}{N_0 + N_1}. \quad (2)$$

These priors acknowledge the typical imbalance between positive and negative samples in contrastive learning. By incorporating these priors, the model focuses on meaningful cross-modal relationships rather than frequency-based discrimination.

Posterior Probability Derivation. Applying Bayes' theorem, we derive the posterior probability that a pair (z^{m_1}, z^{m_2}) is positive given context T :

$$\begin{aligned} p(J = 1|z^{m_1}, z^{m_2}, T) &= \frac{p(z^{m_1}, z^{m_2}|T, J = 1) \cdot p(J = 1|T)}{p(z^{m_1}, z^{m_2}|T)} \\ &= \frac{p(z^{m_1}, z^{m_2}|T)}{p(z^{m_1}, z^{m_2}|T) + \frac{N_0}{N_1} p(z^{m_1}|T)p(z^{m_2}|T)}, \end{aligned} \quad \mathcal{L}_{\text{NCE}}(h) = \mathbb{E}_q \left[\log h(z^{m_1}, z^{m_2}, T) \right] + \frac{N_0}{N_1} \mathbb{E}_q \left[\log(1 - h(z^{m_1}, z^{m_2}, T)) \right], \quad (3)$$

where the denominator is expanded using the law of total probability over $J \in \{0, 1\}$. Taking the logarithm of the

posterior probability and rearranging, we obtain

$$\begin{aligned} \log q(T, J = 1|z^{m_1}, z^{m_2}) &= -\log \left(1 + \frac{N_0}{N_1} \frac{p(z^{m_1}|T)p(z^{m_2}|T)}{p(z^{m_1}, z^{m_2}|T)} \right) \\ &\leq -\log \left(\frac{N_0}{N_1} \right) + \log \frac{p(z^{m_1}, z^{m_2}|T)}{p(z^{m_1}|T)p(z^{m_2}|T)}. \end{aligned} \quad (4)$$

This inequality connects directly to conditional mutual information, with the second term being the pointwise conditional mutual information between z^{m_1} and z^{m_2} given T . The first term $-\log \left(\frac{N_0}{N_1} \right)$ serves as a protective margin against excessive alignment. Taking the expectation with respect to $q(z^{m_1}, z^{m_2}|T, J = 1)$ yields the MI lower bound:

$$\begin{aligned} I(z^{m_1}; z^{m_2}|T) &\geq \log \left(\frac{N_1}{N_0} \right) \\ &+ \mathbb{E}_q \left[\log q(T, J = 1|z^{m_1}, z^{m_2}) \right], \end{aligned} \quad (5)$$

where the term $\log \left(\frac{N_1}{N_0} \right)$ serves as a *protective margin* that prevents the excessive alignment of modalities.

Unlike traditional mutual information maximization approaches that can lead to representation collapse or over-alignment, our formulation naturally incorporates a safeguard through this margin term. This is a key theoretical contribution of our work, as it provides a principled way to control the degree of cross-modal fusion while maintaining modality-specific information.

Estimation via Variational Lower Bound. Since the true posterior $q(T, J = 1|z^{m_1}, z^{m_2})$ is intractable in practice, we must approximate it using a learnable estimator. Drawing inspiration from variational inference techniques, we approximate the posterior with a parametric function $h(z^{m_1}, z^{m_2}, T)$. According to Gibbs' inequality, when h is sufficiently expressive, the optimal estimator satisfies

$$h^*(z^{m_1}, z^{m_2}, T) = q(T, J = 1|z^{m_1}, z^{m_2}). \quad (6)$$

This insight allows us to transform the mutual information estimation problem into a classification problem. The estimator h aims to distinguish between pairs drawn from the joint distribution and those drawn from the product of marginals, with the context T providing additional conditioning information. Similar to standard NCE practice, we optimize h by maximizing a binary classification objective:

$$\begin{aligned} \mathcal{L}_{\text{NCE}}(h) &= \mathbb{E}_q \left[\log h(z^{m_1}, z^{m_2}, T) \right] \\ &+ \frac{N_0}{N_1} \mathbb{E}_q \left[\log(1 - h(z^{m_1}, z^{m_2}, T)) \right], \end{aligned} \quad (7)$$

$$h^* = \arg \max_{h \in \mathcal{H}} \mathcal{L}_{\text{NCE}}(h).$$

This objective balances the correct classification of positive pairs (first term) against the correct classification of negative pairs (second term), weighted by the ratio of negative to positive samples. This weighting ensures that the estimator doesn't simply exploit class imbalance but rather learns meaningful discriminative features across modalities. Plugging Eq. (7) into Eq. (5) yields

$$I(z^{m_1}; z^{m_2} | T) \geq \log \left(\frac{N_1}{N_0} \right) + \mathcal{L}_{\text{NCE}}(h^*). \quad (8)$$

This final bound directly connects our practical training objective (\mathcal{L}_{NCE}) to the theoretical goal of maximizing conditional mutual information. By maximizing \mathcal{L}_{NCE} , we implicitly push up the lower bound on conditional mutual information, thus encouraging cross-modal alignment in a theoretically principled manner.

Optimization of Unimodal Encoders. Having established the connection between our NCE objective and conditional mutual information, we now focus on learning the unimodal encoders $F_{\psi^{(m_1)}}$ and $F_{\psi^{(m_2)}}$ such that their induced embeddings z^{m_1} and z^{m_2} maximize the conditional MI lower bound. In other words, we optimize

$$(\psi^{m_1,*}, \psi^{m_2,*}) = \arg \max_{\psi^{m_1}, \psi^{m_2}} \max_{h \in \mathcal{H}} \mathcal{L}_{\text{NCE}}(h). \quad (9)$$

This joint optimization functions as adversarial training: estimator h learns to distinguish between aligned and non-aligned representations, while encoders generate representations that maximize conditional mutual information. This creates a balance where encoders preserve modality-specific information while aligning shared semantic content. The context condition T guides this process by focusing alignment on task-relevant aspects, T could represent category information, ensuring alignment prioritizes class-discriminative features while allowing non-discriminative features to remain unique to each modality.

Overall Objective. In addition to the conditional MI loss, our training objective combines a task-specific loss (e.g., cross-entropy loss for classification) with the regularization provided by the NCE-based mutual information term. Denote the task loss as $\mathcal{L}_{\text{Task}}$ (computed on samples $s \in \mathcal{S}$) and let $w \in \mathcal{W}$ be the classifier parameters. The overall objective becomes

$$\min_{\psi^{(m_1)}, \psi^{(m_2)}, w} \mathbb{E}_{s \sim \mathcal{S}} \left[\mathcal{L}_{\text{Task}}(s; \psi^{(m_1)}, \psi^{(m_2)}, w) \right] - \alpha \mathcal{L}_{\text{NCE}}(h), \quad (10)$$

where α is a weighting coefficient controlling the trade-off between task performance and cross-modal information alignment, allowing practitioners to adjust the degree of cross-modal alignment based on task requirements and dataset characteristics.

Algorithm 1 Multimodal Learning with *InfoBridge* Strategy

Require: Context condition T , weighting coefficient α , initial unimodal parameters $\psi^{(m_1)} \in \Psi^{(m_1)}$, $\psi^{(m_2)} \in \Psi^{(m_2)}$, classifier parameters $w \in \mathcal{W}$, training data $\mathcal{D} = \{\{x_{[n]}^{(m)}\}_{m=1}^M, y_{[n]}\}_{n=1}^N$, and NCE estimator $h \in \mathcal{H}$.

- 0: **for** $i = 1, 2, \dots, I$ **do**
- 0: Sample mini-batch B_i from \mathcal{D} ;
- 0: Forward pass through unimodal encoders to obtain $\{z^{m_1}, z^{m_2}\}$;
- 0: Compute task loss $\mathcal{L}_{\text{Task}}$ (e.g., cross-entropy);
- 0: Compute the conditional NCE loss $\mathcal{L}_{\text{NCE}}(h)$ using Eq. (1);
- 0: Update h by maximizing $\mathcal{L}_{\text{NCE}}(h)$ via Eq. (7);
- 0: Update $\psi^{(m_1)}$, $\psi^{(m_2)}$ and w using the gradient of the overall loss in Eq. (10);
- 0: **end for**

Discussion. Compared to standard InfoNCE, our formulation makes two key innovations: (1) explicitly incorporating conditioning variable T that focuses alignment on task-relevant information, and (2) including the protective margin term $\log \left(\frac{N_1}{N_0} \right)$. This addresses the fundamental limitation in existing multimodal fusion approaches—the tendency toward over-alignment that sacrifices modality-specific information. By controlling alignment through both context conditioning and protective margin (which emerges naturally from our theoretical derivation), our method achieves balanced fusion that preserves complementary information while retaining modality-specific nuances crucial for downstream tasks.

Algorithm. Algorithm 1 summarizes the training procedure. At each iteration, a mini-batch is sampled and fed into the network. The task-specific loss and NCE loss are computed; then the estimator h is updated via equation (5) and the unimodal encoder parameters are updated via equation (6). This iterative process optimizes both the discriminative power of the estimator and the quality of the multimodal representations, ensuring they maintain an optimal balance between modality-specific information preservation and cross-modal alignment guided by task contexts.

3.1. Theoretical Analysis of *InfoBridge*

To further understand the effectiveness of our framework, we now analyze its theoretical properties and establish connections to existing multimodal learning theory.

Approximate Realizability and Excess Risk. The fundamental challenge in multimodal learning is effectively utilizing information from different modalities while managing their complex interdependencies. Recent theoretical work by Lu et al. [46] has established frameworks for ana-

lyzing representation alignment and transfer in multimodal settings. Building on these foundations, we analyze how our conditional mutual information approach impacts cross-modal alignment quality and downstream task performance.

(i) *Cross-Modal Approximation Error.* When learning from multiple modalities, a central question is how well information from one modality can be transferred to or predicted from another. Following the theoretical framework of Alemi et al. [1] and Federici et al. [18], we formalize this notion through the concept of approximate realizability. Recall that our unimodal encoders $F_{\psi^{(m_1)}}$ and $F_{\psi^{(m_2)}}$ generate feature spaces \mathcal{Z}^{m_1} and \mathcal{Z}^{m_2} . A natural assumption is that there exists a function from a class \mathcal{V} , mapping from \mathcal{Z}^{m_1} to \mathcal{Z}^{m_2} , that can effectively capture the cross-modal relationship. We thus define the *approximate realizability* (or mapping error) as

$$\mathcal{A}(\mathcal{V}; (\mathcal{Z}^{m_1}, \mathcal{Z}^{m_2})) = \min_{v \in \mathcal{V}} \mathbb{E}_{(z^{m_1}, z^{m_2}, y)} \|v(z^{m_1}) - z^{m_2}\|. \quad (11)$$

This quantity measures how well one modality’s representation can predict the other’s, given function class \mathcal{V} . A small \mathcal{A} indicates naturally well-aligned cross-modal embeddings. Our conditional mutual information objective directly minimizes this approximation error - maximizing $I(z^{m_1}; z^{m_2}|T)$ implicitly increases predictive power between modalities, reducing \mathcal{A} . Unlike methods that directly minimize reconstruction error, our approach maintains a protective margin that prevents over-alignment.

(ii) *Complexity Control and Excess Risk Bound.* Cross-modal alignment alone does not guarantee strong downstream performance. We must control function class complexity to prevent overfitting. Using Gaussian averages $G(\cdot)$ to measure the complexity of predictor class \mathcal{W} and mapping class \mathcal{V} , with probability at least $1 - \delta$ the excess risk is bounded as:

$$R(\hat{v}, \hat{w}) \leq \frac{\sqrt{2\pi}}{|S|} G(\mathcal{W}(\mathcal{Z}^{m_1}, \hat{\mathcal{Z}}^{m_2})) + \frac{2\sqrt{2\pi}L}{|S|} G(\mathcal{V}(\mathcal{Z}^{m_1})) + L\mathcal{A}(\mathcal{V}; (\mathcal{Z}^{m_1}, \mathcal{Z}^{m_2})) + (8L + 4)\sqrt{\frac{\log(8/\delta)}{2|S|}}, \quad (12)$$

where $|S|$ is sample count and L is Lipschitz constant. This bound reveals the trade-off: excess risk depends on approximation error \mathcal{A} and complexity terms $G(\cdot)$. Our method addresses both by maximizing conditional mutual information with protective margin to reduce \mathcal{A} while preventing overcomplex mappings, with T focusing alignment on task-relevant aspects.

Unifying Empirical Fusion Strategies in *InfoBridge*.

Our framework provides theoretical foundations for multimodal fusion through conditional information optimization, where unimodal features $\psi^{m_1}(\cdot)$ and $\psi^{m_2}(\cdot)$ create a joint embedding space balancing modality-specific and

cross-modal information. For a linear classifier with parameters $w \in \mathbb{R}^{C \times (D_1 + D_2)}$ and bias $b \in \mathbb{R}^C$, the logits are: $f(x^{m_1}, x^{m_2}) = w^{m_1} \cdot \psi^{m_1}(x^{m_1}) + w^{m_2} \cdot \psi^{m_2}(x^{m_2}) + b$, where $w = [w^{m_1}, w^{m_2}]$.

From an information-theoretic perspective, this fusion operation shapes gradient flow between modalities. Our conditional mutual information objective provides a framework for controlling this flow. Gradient updates under our framework balance modality-specific and cross-modal learning signals. For classification with cross-entropy loss \mathcal{L}_{ce} , gradient updates for classifier w^{m_k} and encoder ψ^{m_k} (with $m_k \in \{m_1, m_2\}$) decompose into:

$$\begin{aligned} \nabla_{w^{m_k}} \mathcal{L} &= \mathbb{E}_{(x^{m_k}, y)} \left[\frac{\partial \mathcal{L}_{ce}}{\partial f(x^{m_k})} \psi^{m_k}(x^{m_k}) \right] \\ &\quad - \alpha \nabla_{w^{m_k}} \mathcal{L}_{\text{NCE}}(h) \\ \nabla_{\psi^{m_k}} \mathcal{L} &= \mathbb{E}_{(x^{m_k}, y)} \left[\frac{\partial \mathcal{L}_{ce}}{\partial f(x^{m_k})} \frac{\partial (w^{m_k} \cdot \psi^{m_k}(x^{m_k}))}{\partial \psi^{m_k}} \right] \\ &\quad - \alpha \nabla_{\psi^{m_k}} \mathcal{L}_{\text{NCE}}(h) \\ &= \underbrace{\nabla_{\psi^{m_k}} \mathcal{L}_{\text{task}}}_{\text{Task gradient}} - \alpha \underbrace{\nabla_{\psi^{m_k}} \mathcal{L}_{\text{NCE}}(h)}_{\text{Information gradient}} \end{aligned} \quad (13)$$

This decomposition reveals how our conditional mutual information objective modulates the learning process. The task gradient pushes each modality to be discriminative for the primary task, while the information gradient ensures cross-modal alignment with appropriate constraints from the protective margin. Furthermore, the information gradient term can be further decomposed to reveal its role in modality balancing:

$$\begin{aligned} \nabla_{\psi^{m_k}} \mathcal{L}_{\text{NCE}}(h) &= \mathbb{E}_{(z^{m_1}, z^{m_2}, T)} \left[\frac{\partial \log h}{\partial \psi^{m_k}} \right] \\ &\quad + \frac{N_0}{N_1} \mathbb{E}_{(z^{m_1}, z^{m_2}, T)} \left[\frac{\partial \log(1-h)}{\partial \psi^{m_k}} \right] \\ &= \mathbb{E}_{(z^{m_1}, z^{m_2}, T)} \left[\underbrace{\nabla_{\psi^{m_k}} \log \frac{p(z^{m_1}, z^{m_2}|T)}{p(z^{m_1}|T)p(z^{m_2}|T)}}_{\text{Modality alignment}} \right] \\ &\quad + \underbrace{\mathbb{E}_{(z^{m_1}, z^{m_2}, T)} \left[\nabla_{\psi^{m_k}} \log \left(\frac{N_1}{N_0} \right) \right]}_{\text{Protective margin}} \end{aligned} \quad (14)$$

This decomposition shows how our approach balances modality-specific learning, cross-modal alignment, and over-fusion protection. The protective margin emerges from our information-theoretic formulation preventing modality dominance. Our unified gradient perspective offers a principled framework for understanding multimodal fusion strategies. The conditional mutual information objective guides cross-modal representation learning while ensuring balanced gradient flow, resulting in robust models.

4. Experiments

4.1. Experimental Setup

Dataset. *CREMA-D* [8] is an audio-visual emotion recognition dataset with 7,442 clips across 6 categories, split into 6,698 training and 744 testing clips. *AVE* [65] is an audio-visual event localization dataset containing 28 classes and 4,143 ten-second videos. We extract frames from event segments and audio clips, forming a multimodal classification dataset. *UPMC Food-101* [20] is an image-text classification dataset collected via Google Image Search, containing 101 food categories from Food-101 [7]. *Glioblastoma & Lower Grade Glioma (GBMLGG)* [9] comprises genomic files and pathological images from TCGA [66]. We extract ROIs from slides and apply stain normalization [67], yielding 1,505 images for 769 patients with WHO labels. Each patient has 80 CNA genomic features.

Architecture Details. For CREMA-D and AVE datasets, we adopt ResNet18 to transform inputs into 512-dimensional feature vectors. Audio data converts to spectrograms with dimensions 224×224 for CREMA-D and 128×128 for AVE, while 3-4 frames are randomly sampled from videos for training. For UPMC Food-101, we employ pre-trained ViT-B/16 for images and BERT-base for text. For GBMLGG, our implementation follows [11] with CNN (ResNet-50) for pathological images and sparse neural network for genomic profiles.

Training Configuration. All models are trained using SGD optimizer with momentum 0.9. Learning rate is 0.01 for CREMA-D/AVE, 0.001 for Food-101, and 0.005 for GBMLGG, with cosine annealing decay. We use mini-batch size 128 for all datasets and train for 200 epochs for CREMA-D/AVE, 100 epochs for Food-101, and 150 epochs for GBMLGG. Weight decay is 1e-4 across experiments. The conditional mutual information weight α is 0.1 for all datasets after grid search over $\{0.01, 0.05, 0.1, 0.5, 1.0\}$. For negative sampling, we use $N_0 = 5$ negative pairs per positive pair ($N_1 = 1$), resulting in protective margin $\gamma = \log(N_1/N_0) = -\log(5) \approx -1.61$. A memory buffer size 1024 is implemented for efficient sample retrieval during contrastive learning. All experiments were conducted on NVIDIA GeForce RTX 4090 GPUs.

Context Condition T . The condition T in our framework can be any task-relevant context. In this work, we primarily set T to be the semantic category information of samples. We also experiment with alternative instantiations: (i) quantized cross-modal similarity scores (3 levels), (ii) unsupervised clustering assignments ($k=5$ or $k=10$), and (iii) ground-truth semantic categories.

4.2. Comparisons with State-of-the-Arts

Compared Methods. performance when training a model using only a single modality. Joint-Train [68] corre-

Table 1. Comparison of *InfoBridge* with state-of-the-art methods on four datasets.

Methods	CREMA-D	AVE	Food-101	GBMLGG
Unimodal [15]	54.40	62.10	68.92	85.24
Joint-Train [68]	53.20	65.40	78.69	85.88
InfoNCE [52]	60.61	67.52	82.55	90.67
MMPareto [61]	60.12	67.42	82.31	90.25
D&R [35]	61.35	67.48	83.36	90.44
ReconBoost [10]	61.59	68.05	83.26	89.88
OGM-GE [55]	61.16	67.01	80.25	89.22
PMR [17]	61.10	67.10	82.09	89.06
<i>InfoBridge</i> (Ours)	62.85	68.19	84.03	92.15

sponds to the standard joint training approach in which modality-specific features are directly concatenated. The InfoNCE baseline [52] maximizes cross-modal mutual information without incorporating any conditional context. MMPareto [61] boosts multimodal learning by leveraging assistance from unimodal learning. D&R [35] diagnoses modality imbalance and adjusts learning to achieve a balanced model. ReconBoost [10] employs reconstruction-based boosting techniques to reconcile differences among modalities. OGM-GE [55] applies on-the-fly gradient modulation with dynamic Gaussian noise to mitigate generalization drop, while PMR [17] utilizes prototypical modality rebalancing to stimulate slower learning modalities.

Results. Tab. 1 demonstrates that our *InfoBridge* consistently outperforms all compared baselines on every dataset. By extending InfoNCE with conditional mutual information and incorporating a protective margin (i.e., $\log \frac{N_0}{N_1}$) to prevent over-alignment, *InfoBridge* achieves a balanced fusion that preserves modality-specific details while enhancing cross-modal alignment. Notably, while approaches like OGM-GE and PMR require additional modules or are tied to specific fusion strategies, our method is agnostic to the fusion scheme and classifier structure, making it versatile.

4.3. Ablation Studies

To assess component contributions in *InfoBridge*, we conduct ablation experiments based on Section 3. Our method jointly optimizes: (i) conditional mutual information (CMI) objective encouraging cross-modal alignment using NCE estimator h , (ii) protective margin term $\log \frac{N_0}{N_1}$ preventing excessive fusion, (iii) context-aware prior via condition T , and (iv) semantic context as condition. We report results when components are removed or replaced: (i) **W/o CMI Objective:** We remove cross-modal mutual InfoMax component (estimator h) so no explicit alignment is enforced. (ii) **W/o Protective Margin:** We omit protective margin term, relaxing constraint preventing over-alignment. (iii) **W/o Context-aware Prior:** We replace

Table 2. Ablation on the key components of *InfoBridge*.

Methods	CREMA-D	AVE	Food-101	GBMLGG
W/o CMI Objective	56.81	65.12	80.85	87.70
W/o Protective Margin	58.52	66.21	81.59	87.79
W/o Context-aware Prior	60.61	67.03	82.00	89.00
W/o Semantic Context	<u>61.58</u>	<u>67.52</u>	<u>82.55</u>	<u>90.67</u>
<i>InfoBridge</i> (Full)	62.85	68.19	84.03	91.15

context-aware mechanism with constant condition for samples, losing adaptive guidance from T . (iv) **W/o Semantic Context**: Instead of semantic labels as condition T , we adopt unsupervised clustering ($k = 5$) to assign T . Tab. 2 summarizes results. The full model achieves best performance, confirming each ingredient contributes to multi-modal learning.

4.4. Further Empirical Analysis

How *InfoBridge* Mitigate Modal Heterogeneity? Fig. 3 demonstrates that increasing the modality correlation weight α produces a significant reduction in the average Gaussian kernel distance between cross-modal features. This trend is consistent across all experimental configurations. Furthermore, the within-category distances decrease even more substantially compared to between-category distances, confirming that our conditional InfoMax mechanism effectively aligns the underlying distributions across modalities while preserving their discriminatory properties necessary for effective classification.

How *InfoBridge* Facilitate Multimodal Learning? Fig. 4 presents T-SNE visualizations on the GBMLGG dataset, offering visual evidence of our method’s effectiveness. Compared to the baseline approach, our proposed method demonstrates superior alignment across different modality representations, resulting in enhanced category separation in the feature space. The visualization reveals more compact within-class clusters and clearer boundaries between different classes. This coherent cross-modal interaction, driven by the conditional fusion mechanism we introduced, leads to improved feature discrimination and consequently better multimodal learning outcomes.

Different Choices of Conditional Variable T . The condition T in our framework can incorporate various forms of task-relevant information. Tab. 3 compares several instantiations of T : (i) *Retrieval*: Using quantized cross-modal similarity scores (three levels). (ii) *Cluster*: Assigning T through unsupervised clustering with $k = 5$ and $k = 10$. (iii) *Semantics*: Utilizing ground-truth semantic category information as T . Our experiments reveal that while all instantiations yield robust performance, semantic category conditioning consistently obtains the best results, underscoring the flexibility of our framework.

Table 3. Ablation on using different condition T .

Methods	CREMA-D	AVE	Food-101	GBMLGG
Retrieval ($k = 3$)	61.05	66.23	82.66	89.54
Cluster ($k = 5$)	61.74	<u>67.62</u>	82.55	<u>90.67</u>
Cluster ($k = 10$)	<u>62.18</u>	67.38	<u>83.11</u>	90.15
Semantics (Ours)	62.85	68.19	84.03	91.15

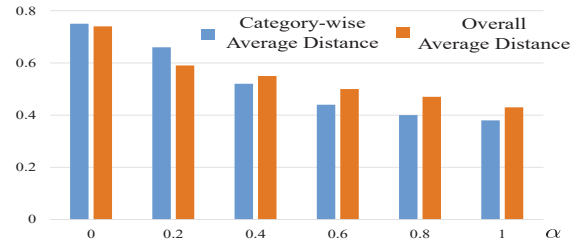


Figure 3. Average Gaussian kernel distance variation with different CMI weights α on AVE; same-category distances shown separately. Bar chart shows category-wise and overall distance changes as CMI weight α varies from 0 to 1.

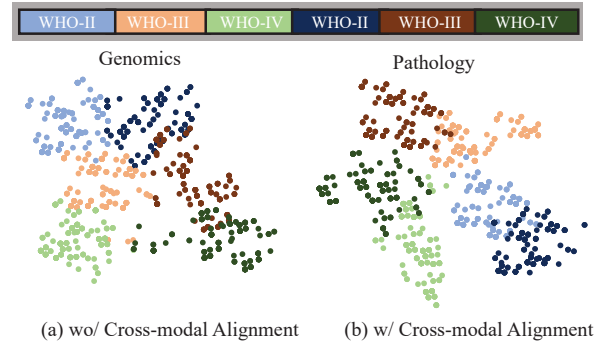


Figure 4. T-SNE visualization on GBMLGG: (a) baseline, (b) ours. Our method achieves clearer category separation and better cross-modal alignment in genomics and pathology data.

5. Conclusion

This study tackles the balance between cross-modal fusion and modality-specific preservation by introducing *InfoBridge*, a theoretically grounded framework leveraging conditional mutual information maximization with protective margins. Our contributions include a novel formulation of multimodal fusion as conditional mutual information optimization with protective margins against over-fusion, a context-aware approach adapting fusion strategies based on task-relevant conditioning, and comprehensive theoretical analysis linking contrastive learning to information-theoretic principles. Extensive experiments on diverse datasets show *InfoBridge* outperforms state-of-the-arts.

Acknowledgment. This work was supported by Hong Kong Research Grants Council (RGC) General Research Fund 14204321 and CUHK 4055269

References

- [1] Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. In *International Conference on Learning Representations*, 2016. 6
- [2] Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-supervised learning by cross-modal audio-video clustering. *Advances in Neural Information Processing Systems*, 33, 2020. 2
- [3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 2
- [4] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. *Advances in neural information processing systems*, 29:892–900, 2016. 2
- [5] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018. 1
- [6] Khaled Bayouhd, Raja Knani, Fayçal Hamdaoui, and Abdelatif Mtibaa. A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets. *The Visual Computer*, 38(8):2939–2970, 2022. 1
- [7] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101—mining discriminative components with random forests. In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part VI 13*, pages 446–461. Springer, 2014. 7
- [8] 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. 7
- [9] Ethan Cerami, Jianjiong Gao, Ugur Dogrusoz, Benjamin E Gross, Selcuk Onur Sumer, Bülent Arman Aksoy, Anders Jacobsen, Caitlin J Byrne, Michael L Heuer, Erik Larsson, et al. The cbio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer discovery*, 2(5):401–404, 2012. 7
- [10] L. Chen and P. Wong. Reconboost: Boosting can achieve modality reconciliation. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 789–798, 2024. 7
- [11] Richard J Chen, Ming Y Lu, Jingwen Wang, Drew FK Williamson, Scott J Rodig, Neal I Lindeman, and Faisal Mahmood. Pathomic fusion: an integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. *IEEE Transactions on Medical Imaging*, 41(4):757–770, 2020. 7
- [12] Yuanyuan Chen, Yongsheng Pan, Yong Xia, and Yixuan Yuan. Disentangle first, then distill: A unified framework for missing modality imputation and alzheimer’s disease diagnosis. *IEEE Transactions on Medical Imaging*, 2023. 1, 2, 3
- [13] Mingyuan Cheng, Xinru Liao, Quan Liu, Bin Ma, Jian Xu, and Bo Zheng. Learning disentangled representations for counterfactual regression via mutual information minimization. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1802–1806, 2022. 1, 3
- [14] Yinglong Dai, Zheng Yan, Jiangchang Cheng, Xiaojun Duan, and Guojun Wang. Analysis of multimodal data fusion from an information theory perspective. *Information Sciences*, 623:164–183, 2023. 3
- [15] John Doe and Jane Smith. Effective unimodal learning approaches for multimodal systems. *IEEE Transactions on Multimedia*, 22(5):1287–1298, 2020. 7
- [16] Chenzhuang Du, Tingle Li, Yichen Liu, Zixin Wen, Tianyu Hua, Yue Wang, and Hang Zhao. Improving multimodal learning with uni-modal teachers. *arXiv preprint arXiv:2106.11059*, 2021. 1, 2, 3
- [17] 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*, pages 20029–20038, 2023. 7
- [18] Marco Federici, Anjan Dutta, Patrick Forré, Nate Kushman, and Zeynep Akata. Learning robust representations via multi-view information bottleneck. In *8th International Conference on Learning Representations*. OpenReview.net, 2020. 3, 6
- [19] Zheren Fu, Zhendong Mao, Yan Song, and Yongdong Zhang. Learning semantic relationship among instances for image-text matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15159–15168, 2023. 2
- [20] Ignazio Gallo, Gianmarco Ria, Nicola Landro, and Riccardo La Grassa. Image and text fusion for upmc food-101 using bert and cnns. In *2020 35th International Conference on Image and Vision Computing New Zealand (IVCNZ)*, pages 1–6. IEEE, 2020. 7
- [21] Ruohan Gao, Tae-Hyun Oh, Kristen Grauman, and Lorenzo Torresani. Listen to look: Action recognition by previewing audio. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10457–10467, 2020. 2
- [22] Zongbo Han, Fan Yang, Junzhou Huang, Changqing Zhang, and Jianhua Yao. Multimodal dynamics: Dynamical fusion for trustworthy multimodal classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20707–20717, 2022. 1
- [23] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018. 3
- [24] Di Hu, Xuelong Li, et al. Temporal multimodal learning in audiovisual speech recognition. In *Proceedings of the*

- IEEE Conference on Computer Vision and Pattern Recognition*, pages 3574–3582, 2016. 2
- [25] Di Hu, Feiping Nie, and Xuelong Li. Deep multimodal clustering for unsupervised audiovisual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9248–9257, 2019. 2
- [26] Di Hu, Chengze Wang, Feiping Nie, and Xuelong Li. Dense multimodal fusion for hierarchically joint representation. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3941–3945. IEEE, 2019. 2
- [27] Ilija Iliovski and Jiashi Feng. Multimodal learning and reasoning for visual question answering. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. 2
- [28] Javed Imran and Balasubramanian Raman. Evaluating fusion of rgb-d and inertial sensors for multimodal human action recognition. *Journal of Ambient Intelligence and Humanized Computing*, 11(1):189–208, 2020. 1
- [29] Yu-Gang Jiang, Zuxuan Wu, Jinhui Tang, Zechao Li, Xiangyang Xue, and Shih-Fu Chang. Modeling multimodal clues in a hybrid deep learning framework for video classification. *IEEE Transactions on Multimedia*, 20(11):3137–3147, 2018. 1
- [30] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5492–5501, 2019. 2
- [31] Sein Kim, Namkyeong Lee, Junseok Lee, Dongmin Hyun, and Chanyoung Park. Heterogeneous graph learning for multi-modal medical data analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5141–5150, 2023. 1
- [32] Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. *arXiv preprint arXiv:1807.00230*, 2018. 2
- [33] Changhee Lee and Mihaela Van der Schaar. A variational information bottleneck approach to multi-omics data integration. In *International Conference on Artificial Intelligence and Statistics*, pages 1513–1521. PMLR, 2021. 3
- [34] John Lee, Max Dabagia, Eva Dyer, and Christopher Rozell. Hierarchical optimal transport for multimodal distribution alignment. *Advances in neural information processing systems*, 32, 2019. 1
- [35] M. Lee and S. Kim. Diagnosing and re-learning for balanced multimodal learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 567–576, 2024. 7
- [36] Chenxin Li, Xinyu Liu, Cheng Wang, Yifan Liu, Weihao Yu, Jing Shao, and Yixuan Yuan. Gtp-4o: Modality-prompted heterogeneous graph learning for omni-modal biomedical representation. In *European conference on computer vision*, pages 168–187. Springer, 2024. 1
- [37] Ming Li, Jike Zhong, Chenxin Li, Liuzhuozheng Li, Nie Lin, and Masashi Sugiyama. Vision-language model fine-tuning via simple parameter-efficient modification. *arXiv preprint arXiv:2409.16718*, 2024. 1
- [38] Jian Liang, Ran He, Zhenan Sun, and Tieniu Tan. Aggregating randomized clustering-promoting invariant projections for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 41(5):1027–1042, 2018. 2
- [39] Paul Pu Liang, Yun Cheng, Xiang Fan, Chun Kai Ling, Suzanne Nie, Richard Chen, Zihao Deng, Faisal Mahmood, Ruslan Salakhutdinov, and Louis-Philippe Morency. Quantifying & modeling feature interactions: An information decomposition framework. *arXiv preprint arXiv:2302.12247*, 2023. 3
- [40] Yunlong Lin, Zhenqi Fu, Kairun Wen, Tian Ye, Sixiang Chen, Ge Meng, Yingying Wang, Yue Huang, Xiaotong Tu, and Xinghao Ding. Unsupervised low-light image enhancement with lookup tables and diffusion priors. *arXiv preprint arXiv:2409.18899*, 2024. 1
- [41] Yunlong Lin, Tian Ye, Sixiang Chen, Zhenqi Fu, Yingying Wang, Wenhao Chai, Zhaoxu Xing, Lei Zhu, and Xinghao Ding. Aglldiff: Guiding diffusion models towards unsupervised training-free real-world low-light image enhancement. *arXiv preprint arXiv:2407.14900*, 2024. 1
- [42] Yunlong Lin, Zixu Lin, Haoyu Chen, Panwang Pan, Chenxin Li, Sixiang Chen, Kairun Wen, Yeying Jin, Wenbo Li, and Xinghao Ding. Jarvisir: Elevating autonomous driving perception with intelligent image restoration. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22369–22380, 2025. 1
- [43] Yunlong Lin, Zixu Lin, Kunjie Lin, Jinbin Bai, Panwang Pan, Chenxin Li, Haoyu Chen, Zhongdao Wang, Xinghao Ding, Wenbo Li, et al. Jarvisart: Liberating human artistic creativity via an intelligent photo retouching agent. *arXiv preprint arXiv:2506.17612*, 2025.
- [44] Parker Liu, Chenxin Li, Zhengxin Li, Yipeng Wu, Wuyang Li, Zhiqin Yang, Zhenyuan Zhang, Yunlong Lin, Sirui Han, and Brandon Y Feng. Ir3d-bench: Evaluating vision-language model scene understanding as agentic inverse rendering. *arXiv preprint arXiv:2506.23329*, 2025. 1
- [45] Siyu Lu, Mingzhe Liu, Lirong Yin, Zhengtong Yin, Xuan Liu, and Wenfeng Zheng. The multi-modal fusion in visual question answering: a review of attention mechanisms. *PeerJ Computer Science*, 9:e1400, 2023. 1
- [46] Zhou Lu. A theory of multimodal learning. *Advances in Neural Information Processing Systems*, 36:57244–57255, 2023. 5
- [47] Sijie Mai, Ying Zeng, and Haifeng Hu. Multimodal information bottleneck: Learning minimal sufficient unimodal and multimodal representations. *IEEE Transactions on Multimedia*, 2022. 3
- [48] Youssef Mroueh, Etienne Marcheret, and Vaibhava Goel. Deep multimodal learning for audio-visual speech recognition. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2130–2134. IEEE, 2015. 1
- [49] Arsha Nagrani, Chen Sun, David Ross, Rahul Sukthankar, Cordelia Schmid, and Andrew Zisserman. Speech2action: Cross-modal supervision for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10317–10326, 2020. 2

- [50] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *ICML*, 2011. 1
- [51] Dan Oneață and Horia Cucu. Improving multimodal speech recognition by data augmentation and speech representations. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4578–4587. IEEE, 2022. 1
- [52] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 7
- [53] Frederik Pahde, Mihai Puscas, Tassilo Klein, and Moin Nabi. Multimodal prototypical networks for few-shot learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2644–2653, 2021. 1
- [54] Yagya Raj Pandeya, Bhuwan Bhattarai, and Joonwhoan Lee. Deep-learning-based multimodal emotion classification for music videos. *Sensors*, 21(14):4927, 2021. 1
- [55] Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8238–8247, 2022. 1, 3, 7
- [56] Gerasimos Potamianos, Chalapathy Neti, Juergen Luettn, and Iain Matthews. Audio-visual automatic speech recognition: An overview. *Issues in visual and audio-visual speech processing*, 22:23, 2004. 2
- [57] Eduardo Hugo Sanchez, Mathieu Serrurier, and Mathias Ortnier. Learning disentangled representations via mutual information estimation. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*, pages 205–221. Springer, 2020. 1, 3
- [58] Amir Shahroudy, Tian-Tsong Ng, Yihong Gong, and Gang Wang. Deep multimodal feature analysis for action recognition in rgb+d videos. *IEEE transactions on pattern analysis and machine intelligence*, 40(5):1045–1058, 2017. 1
- [59] Shashi Kant Shankar, Luis P Prieto, María Jesús Rodríguez-Triana, and Adolfo Ruiz-Calleja. A review of multimodal learning analytics architectures. In *2018 IEEE 18th international conference on advanced learning technologies (ICALT)*, pages 212–214. IEEE, 2018. 1
- [60] Christian Simon, Piotr Koniusz, Richard Nock, and Mehrtaash Harandi. On modulating the gradient for meta-learning. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16*, pages 556–572. Springer, 2020. 1
- [61] J. Smith and A. Doe. Mmpareto: Boosting multimodal learning with innocent unimodal assistance. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 1023–1032, 2024. 7
- [62] Nitish Srivastava and Russ R Salakhutdinov. Multimodal learning with deep boltzmann machines. *Advances in neural information processing systems*, 25, 2012. 1
- [63] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multi-modal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*, 2023. 2
- [64] Haiman Tian, Yudong Tao, Samira Pouyanfar, Shu-Ching Chen, and Mei-Ling Shyu. Multimodal deep representation learning for video classification. *World Wide Web*, 22(3):1325–1341, 2019. 1
- [65] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 247–263, 2018. 1, 7
- [66] Katarzyna Tomczak, Patrycja Czerwińska, and Maciej Wizerowicz. Review the cancer genome atlas (tcga): an immeasurable source of knowledge. *Contemporary Oncology/Współczesna Onkologia*, 2015(1):68–77, 2015. 7
- [67] Abhishek Vahadane, Tingying Peng, Amit Sethi, Shadi Albarqouni, Lichao Wang, Maximilian Baust, Katja Steiger, Anna Melissa Schlitter, Irene Esposito, and Nassir Navab. Structure-preserving color normalization and sparse stain separation for histological images. *IEEE transactions on medical imaging*, 35(8):1962–1971, 2016. 7
- [68] Li Wang and Wei Zhang. Joint training for multimodal data fusion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1024–1032, 2020. 7
- [69] Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12695–12705, 2020. 3
- [70] Hongtao Wu, Yijun Yang, Haoyu Chen, Jingjing Ren, and Lei Zhu. Mask-guided progressive network for joint raindrop and rain streak removal in videos. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 7216–7225, 2023. 1
- [71] Hongtao Wu, Yijun Yang, Angelica I Aviles-Rivero, Jingjing Ren, Sixiang Chen, Haoyu Chen, and Lei Zhu. Semi-supervised video desnowing network via temporal decoupling experts and distribution-driven contrastive regularization. In *European Conference on Computer Vision*, pages 70–89. Springer, 2024.
- [72] Hongtao Wu, Yijun Yang, Huihui Xu, Weiming Wang, Jinni Zhou, and Lei Zhu. Rainmamba: Enhanced locality learning with state space models for video deraining. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 7881–7890, 2024. 1
- [73] Yan Xia and Zhou Zhao. Cross-modal background suppression for audio-visual event localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19989–19998, 2022. 1
- [74] Fanyi Xiao, Yong Jae Lee, Kristen Grauman, Jitendra Malik, and Christoph Feichtenhofer. Audiovisual slowfast networks for video recognition. *arXiv preprint arXiv:2001.08740*, 2020. 3
- [75] Yingxue Xu and Hao Chen. Multimodal optimal transport-based co-attention transformer with global structure consistency for survival prediction. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 1
- [76] Zihui Xue and Radu Marculescu. Dynamic multimodal fusion. In *Proceedings of the IEEE/CVF Conference on Com-*

- puter Vision and Pattern Recognition*, pages 2574–2583, 2023. 1
- [77] Zequn Yang, HaoTian Ni, Yake Wei, and Di Hu. Towards holistic multimodal interaction: An information-theoretic perspective. 3
- [78] Yao Zhang, Jiawei Yang, Jiang Tian, Zhongchao Shi, Cheng Zhong, Yang Zhang, and Zhiqiang He. Modality-aware mutual learning for multi-modal medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24*, pages 589–599. Springer, 2021. 1
- [79] Yingli Zuo, Yawen Wu, Zixiao Lu, Qi Zhu, Kun Huang, Daoqiang Zhang, and Wei Shao. Identify consistent imaging genomic biomarkers for characterizing the survival-associated interactions between tumor-infiltrating lymphocytes and tumors. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 222–231. Springer, 2022. 1