Multimodal Learning Without Labeled Multimodal Data: Guarantees and Applications

Anonymous Author(s)
Affiliation
Address
email

Abstract

In many machine learning systems that jointly learn from multiple modalities, a core research question is to understand the nature of *multimodal interactions*: the emergence of new task-relevant information during learning from both modalities that was not present in either alone. We study this challenge of interaction quantification in a semi-supervised setting with only labeled unimodal data and naturally co-occurring multimodal data (e.g., unlabeled images and captions, video and corresponding audio) but when labeling them is time-consuming. Using a precise information-theoretic definition of interactions, our key contributions are the derivations of lower and upper bounds to quantify the amount of multimodal interactions in this semi-supervised setting. We propose two lower bounds based on the amount of *shared information* between modalities and the *disagreement* between separately trained unimodal classifiers, and derive an upper bound through connections to approximate algorithms for *min-entropy couplings*. We validate these estimated bounds and show how they accurately track true interactions. Finally, two semi-supervised multimodal applications are explored based on these theoretical results: (1) analyzing the relationship between multimodal performance and estimated interactions, and (2) self-supervised learning that embraces *disagreement* between modalities beyond agreement as is typically done.

1 Introduction

A core research question in multimodal learning is to understand the nature of *multimodal interactions* across modalities in the context of a task: the emergence of new task-relevant information during learning from both modalities that was not present in either modality alone [6, 56]. In settings where labeled multimodal data is abundant, the study of multimodal interactions has inspired advances in theoretical analysis [1, 37, 57, 71, 82] and representation learning [43, 64, 79, 92] in language and vision [3], multimedia [9], healthcare [45], and robotics [49]. In this paper, we study the problem of interaction quantification in a setting where there is only *unlabeled multimodal data* $\mathcal{D}_M = \{(x_1, x_2)\}$ but some *labeled unimodal data* $\mathcal{D}_i = \{(x_i, y)\}$ collected separately for each modality. This multimodal semi-supervised paradigm is reminiscent of many real-world settings with the emergence of separate unimodal datasets like large-scale visual recognition [20] and text classification [84], as well as the collection of data in multimodal settings (e.g., unlabeled images and captions or video and audio [54, 75, 64, 95]) but when labeling them is time-consuming [40, 41].

Using a precise information-theoretic definition of interactions [10, 87], our key contributions are the derivations of lower and upper bounds to quantify the amount of multimodal interactions in this semi-supervised setting with only $\mathcal{D}_i$ and $\mathcal{D}_M$. We propose two lower bounds for interaction quantification: our first lower bound relates multimodal interactions with the amount of *shared information* between modalities, and our second lower bound introduces the concept of *modality disagreement* which quantifies the differences of classifiers trained separately on each modality. Finally, we propose an upper bound through connections to approximate algorithms for *min-entropy couplings* [14]. To validate our derivations, we experiment on large-scale synthetic and real-world datasets with varying amounts of interactions. In addition, these theoretical results naturally yield new algorithms for two applications involving semi-supervised multimodal data:

1. We first analyze the relationship between interaction estimates and downstream task performance when optimal multimodal classifiers are learned access to multimodal data. This analysis can help develop new guidelines for deciding when to *collect* and *fuse* labeled multimodal data.
2. As the result of our analysis, we further design a new family of self-supervised learning objectives that capture *disagreement* on unlabeled multimodal data, and show that this learns interactions beyond agreement conventionally used in the literature [64, 68, 95]. Our experiments show strong results on four datasets: relating cartoon images and captions [38], predicting expressions of humor and sarcasm from videos [12, 35], and reasoning about multi-party social interactions [93].

More importantly, we believe these results shed light on the intriguing connections between disagreement, multimodal interactions, and performance. We release our code and models at `<anon>`.

2 Preliminaries

2.1 Definitions and setup

Let $\mathcal{X}_i$ and $\mathcal{Y}$ be finite sample spaces for features and labels. Define Δ to be the set of joint distributions over $(\mathcal{X}_1, \mathcal{X}_2, \mathcal{Y})$. We are concerned with features X_1, X_2 (with support $\mathcal{X}_i$) and labels Y (with support $\mathcal{Y}$) drawn from some distribution $p \in \Delta$. We denote the probability mass function by $p(x_1, x_2, y)$, where omitted parameters imply marginalization. In many real-world applications [54, 64, 68, 90, 95], we only have partial datasets from p rather than the full distribution:

- *Labeled unimodal* data $\mathcal{D}_1 = \{(x_1, y) : \mathcal{X}_1 \times \mathcal{Y}\}$, $\mathcal{D}_2 = \{(x_2, y) : \mathcal{X}_2 \times \mathcal{Y}\}$.
- *Unlabeled multimodal* data $\mathcal{D}_M = \{(x_1, x_2) : \mathcal{X}_1 \times \mathcal{X}_2\}$.

$\mathcal{D}_1, \mathcal{D}_2$ and $\mathcal{D}_M$ follow the *pairwise marginals* $p(x_1, y)$, $p(x_2, y)$ and $p(x_1, x_2)$. We define $\Delta_{p_{1,2}} = \{q \in \Delta : q(x_i, y) = p(x_i, y) \ \forall y \in \mathcal{Y}, x_i \in \mathcal{X}_i, i \in [2]\}$ as the set of joint distributions which agree with the labeled unimodal data $\mathcal{D}_1$ and $\mathcal{D}_2$, and $\Delta_{p_{1,2,12}} = \{r \in \Delta : r(x_1, x_2) = p(x_1, x_2), r(x_i, y) = p(x_i, y)\}$ as the set of joint distributions which agree with all $\mathcal{D}_1, \mathcal{D}_2$ and $\mathcal{D}_M$.

Despite partial observability, we often still want to understand the degree to which two modalities can interact to contribute new information not present in either modality alone, in order to inform our decisions on multimodal data collection and modeling [43, 52, 57, 92]. We now cover relevant background towards a formal information-theoretic definition of interactions and their approximation.

2.2 Information theory, partial information decomposition, and synergy

Information theory formalizes the amount of information that a variable (X_1) provides about another (X_2), and is quantified by Shannon's mutual information (MI) and conditional MI [67]:

$$I(X_1; X_2) = \int p(x_1, x_2) \log \frac{p(x_1, x_2)}{p(x_1)p(x_2)} d\boldsymbol{x}, \quad I(X_1; X_2|Y) = \int p(x_1, x_2|y) \log \frac{p(x_1, x_2|y)}{p(x_1|y)p(x_2|y)} d\boldsymbol{x}dy.$$

The MI of two random variables X_1 and X_2 measures the amount of information (in bits) obtained about X_1 by observing X_2, and by extension, conditional MI is the expected value of MI given the value of a third (e.g., Y). However, the extension of information theory to three or more variables to describe the synergy between two modalities for a task remains an open challenge. Among many proposed frameworks, Partial information decomposition (PID) [87] posits a decomposition of the total information 2 variables X_1, X_2 provide about a task Y into 4 quantities: $I_p(\{X_1, X_2\}; Y) = R + U_1 + U_2 + S$ where $I_p(\{X_1, X_2\}; Y)$ is the MI between the joint random variable (X_1, X_2) and Y, redundancy R describes task-relevant information shared between X_1 and X_2, uniqueness U_1 and U_2 studies the task-relevant information present in only X_1 or X_2 respectively, and synergy S investigates the emergence of new information only when both X_1 and X_2 are present [10, 33]:

Definition 1. *(Multimodal interactions) Given X_1, X_2, and a target Y, we define their redundant (R), unique (U_1 and U_2), and synergistic (S) interactions as:*

$$R = \max_{q \in \Delta_{p_{1,2}}} I_q(X_1; X_2; Y), \quad U_1 = \min_{q \in \Delta_{p_{1,2}}} I_q(X_1; Y|X_2), \quad U_2 = \min_{q \in \Delta_{p_{1,2}}} I_q(X_2; Y|X_1), \tag{1}$$

$$S = I_p(\{X_1, X_2\}; Y) - \min_{q \in \Delta_{p_{1,2}}} I_q(\{X_1, X_2\}; Y), \tag{2}$$

where the notation $I_p(\cdot)$ and $I_q(\cdot)$ disambiguates mutual information (MI) under p and q respectively.

$I(X_1; X_2; Y) = I(X_1; X_2) - I(X_1; X_2|Y)$ is a multivariate extension of information theory [8, 60]. Most importantly, R, U_1, and U_2 can be computed exactly using convex programming over distributions $q \in \Delta_{p_{1,2}}$ with access only to the marginals $p(x_1, y)$ and $p(x_2, y)$ by solving an

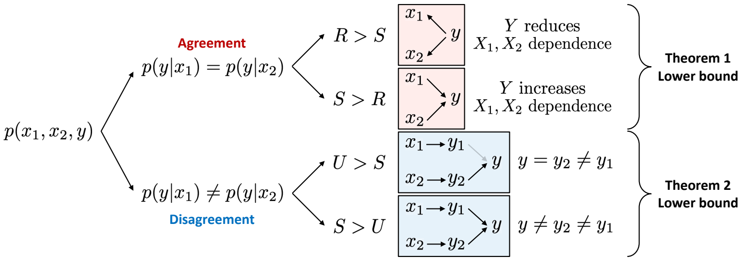

Figure 1: We estimate two types of synergy: (1) *agreement synergy* that arises as a result of Y increasing the agreeing shared information between X_1 and X_2 (reminiscent of common cause structures as opposed to redundancy in common effect), and (2) *disagreement synergy* that emerges due to the disagreement between unimodal predictors resulting in a new prediction $y \neq y_1 \neq y_2$ (rather than uniqueness where $y = y_2 \neq y_1$).

equivalent max-entropy optimization problem $q^* = \arg\max_{q \in \Delta_{p_{1,2}}} H_q(Y|X_1, X_2)$ [10, 57]. This is a convex optimization problem with linear marginal-matching constraints (see Appendix A.2). This gives us an elegant interpretation that we need only labeled unimodal data in each feature from $\mathcal{D}_1$ and $\mathcal{D}_2$ to estimate redundant and unique interactions.

3 Estimating Synergy Without Multimodal Data

Unfortunately, S is impossible to compute via equation (2) when we do not have access to the full joint distribution p, since the first term $I_p(X_1, X_2; Y)$ is unknown. Instead, we will aim to provide lower and upper bounds in the form $\underline{S} \leq S \leq \overline{S}$ which depend *only* on $\mathcal{D}_1$, $\mathcal{D}_2$, and $\mathcal{D}_M$.

3.1 Lower bounds on synergy

Our first insight is that while labeled multimodal data is unavailable, the output of unimodal classifiers may be compared against each other. Let $\delta_\mathcal{Y} = \{r \in \mathbb{R}_+^{|\mathcal{Y}|} \mid ||r||_1 = 1\}$ be the probability simplex over labels $\mathcal{Y}$. Consider the set of unimodal classifiers $\mathcal{F}_i \ni f_i : \mathcal{X}_i \to \delta_\mathcal{Y}$ and multimodal classifiers $\mathcal{F}_M \ni f_M : \mathcal{X}_1 \times \mathcal{X}_2 \to \delta_\mathcal{Y}$. The crux of our method is to establish a connection between *modality disagreement* and a lower bound on synergy.

Definition 2. *(Modality disagreement) Given X_1, X_2, and a target Y, as well as unimodal classifiers f_1 and f_2, we define modality disagreement as $\alpha(f_1, f_2) = \mathbb{E}_{p(x_1,x_2)}[d(f_1, f_2)]$ where $d : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^{\geq 0}$ is a distance function in label space scoring the disagreement of f_1 and f_2's predictions.*

Quantifying modality disagreement gives rise to two types of synergy as illustrated in Figure 1: agreement synergy and disagreement synergy. As their names suggest, *agreement synergy* happens when two modalities agree in predicting the label and synergy arises within this agreeing information. On the other hand, *disagreement synergy* happens when two modalities disagree in predicting the label, and synergy arises due to disagreeing information.

Agreement synergy We first consider the case when two modalities contain shared information that leads to agreement in predicting the outcome. In studying these situations, a driving force for estimating S is the amount of shared information $I(X_1; X_2)$ between modalities, with the intuition that more shared information naturally leads to redundancy which gives less opportunity for new synergistic interactions. Mathematically, we formalize this by relating S to R [87],

$$S = R - I_p(X_1; X_2; Y) = R - I_p(X_1; X_2) + I_p(X_1; X_2|Y). \qquad (3)$$

implying that synergy exists when there is high redundancy and low (or even negative) three-way MI $I_p(X_1; X_2; Y)$ [7, 31]. By comparing the difference in X_1, X_2 dependence with and without the task (i.e., $I_p(X_1; X_2)$ vs $I_p(X_1; X_2|Y)$), 2 cases naturally emerge (see top half of Figure 1):

1. $\mathbf{S} > \mathbf{R}$: When both modalities do not share a lot of information as measured by low $I(X_1; X_2)$, but conditioning on Y *increases* their dependence: $I(X_1; X_2|Y) > I(X_1; X_2)$, then there is synergy between modalities when combining them for task Y. This setting is reminiscent of common cause structures. Examples of these distributions in the real world are multimodal question answering, where the image and question are less dependent (some questions like 'what is the color of the car' or 'how many people are there' can be asked for many images), but the answer (e.g., 'blue car') connects the two modalities, resulting in dependence given the label. As expected, $S = 4.92, R = 0.79$ for the VQA 2.0 dataset [32].

2. **R** > **S**: Both modalities share a lot of information but conditioning on Y *reduces* their dependence: $I(X_1; X_2) > I(X_1; X_2|Y)$, which results in more redundant than synergistic information. This setting is reminiscent of common effect structures. A real-world example is in detecting sentiment from multimodal videos, where text and video are highly dependent since they are emitted by the same speaker, but the sentiment label explains away some of the dependencies between both modalities. Indeed, for multimodal sentiment analysis from text, video, and audio of monologue videos on MOSEI [51, 94], $R = 0.26$ and $S = 0.04$.

However, $I_p(X_1; X_2|Y)$ cannot be computed without access to the full distribution p. In Theorem 1, we obtain a lower bound on $I_p(X_1; X_2|Y)$, resulting in a lower bound $\underline{S}_{\text{agree}}$ for synergy.

Theorem 1. *(Lower-bound on synergy via redundancy) We can relate S to R as follows*

$$\underline{S}_{agree} = R - I_p(X_1; X_2) + \min_{r \in \Delta_{p_{1,2,12}}} I_r(X_1; X_2|Y) \le S \tag{4}$$

We include the full proof in Appendix A.3, but note that $\min_{r \in \Delta_{p_{1,2,12}}} I_r(X_1; X_2|Y)$ is equivalent to a max-entropy optimization problem solvable using convex programming. This implies that $\underline{S}_{\text{agree}}$ can be computed efficiently using only unimodal data $\mathcal{D}_i$ and unlabeled multimodal data $\mathcal{D}_M$.

Disagreement synergy We now consider settings where two modalities disagree in predicting the outcome: suppose $y_1 = \arg\max_y p(y|x_1)$ is the most likely prediction from the first modality, $y_2 = \arg\max_y p(y|x_2)$ for the second modality, and $y = \arg\max_y p(y|x_1, x_2)$ the true multimodal prediction. During disagreement, there are again 2 cases (see bottom half of Figure 1):

1. **U** > **S**: Multimodal prediction $y = \arg\max_y p(y|x_1, x_2)$ is the same as one of the unimodal predictions (e.g., $y = y_2$), in which case unique information in modality 2 leads to the outcome. A real-world dataset that we categorize in this case is MIMIC involving mortality and disease prediction from tabular patient data and time-series medical sensors [45] which primarily shows unique information in the tabular modality. The disagreement on MIMIC is high $\alpha = 0.13$, but since disagreement is due to a lot of unique information, there is less synergy $S = 0.01$.
2. **S** > **U**: Multimodal prediction y is different from both y_1 and y_2, then both modalities interact synergistically to give rise to a final outcome different from both disagreeing unimodal predictions. This type of joint distribution is indicative of real-world examples such as predicting sarcasm from language and speech - the presence of sarcasm is typically detected due to a contradiction between what is expressed in language and speech, as we observe from the experiments on MUSTARD [12] where $S = 0.44$ and $\alpha = 0.12$ are both relatively large.

We formalize these intuitions via Theorem 2, yielding a lower bound $\underline{S}_{\text{disagree}}$ based on disagreement minus the maximum unique information in both modalities:

Theorem 2. *(Lower-bound on synergy via disagreement, informal) We can relate synergy S and uniqueness U to modality disagreement $\alpha(f_1, f_2)$ of optimal unimodal classifiers f_1, f_2 as follows:*

$$\underline{S}_{disagree} = \alpha(f_1, f_2) \cdot c - \max(U_1, U_2) \le S \tag{5}$$

for some constant c depending on the label dimension $|\mathcal{Y}|$ and choice of label distance function d.

Theorem 2 implies that if there is substantial disagreement $\alpha(f_1, f_2)$ between unimodal classifiers, it must be due to the presence of unique or synergistic information. If uniqueness is small, then disagreement must be accounted for by synergy, thereby yielding a lower bound $\underline{S}_{\text{disagree}}$. Note that the notion of optimality in unimodal classifiers is important: poorly-trained unimodal classifiers could show high disagreement but would be uninformative about true interactions. We include the formal version of the theorem based on Bayes' optimality and a full proof in Appendix A.4.

Hence, agreement and disagreement synergy yield separate lower bounds $\underline{S}_{\text{agree}}$ and $\underline{S}_{\text{disagree}}$. Note that these bounds *always* hold, so we could take $\underline{S} = \max\{\underline{S}_{\text{agree}}, \underline{S}_{\text{disagree}}\}$.

3.2 Upper bound on synergy

While the lower bounds tell us the least amount of synergy possible in a distribution, we also want to obtain an upper bound on the possible synergy, which together with the above lower bounds sandwich S. By definition, $S = I_p(\{X_1, X_2\}; Y) - \max_{q \in \Delta_{p_{1,2}}} I_q(\{X_1, X_2\}; Y)$. Thus, upper bounding synergy is the same as *maximizing* the MI $I_p(X_1, X_2; Y)$, which can be rewritten as

$$\max_{r\in\Delta_{p_{1,2,12}}} I_r(\{X_1, X_2\}; Y) = \max_{r\in\Delta_{p_{1,2,12}}} \{H_r(X_1, X_2) + H_r(Y) - H_r(X_1, X_2, Y)\} \tag{6}$$

$$= H_p(X_1, X_2) + H_p(Y) - \min_{r\in\Delta_{p_{1,2,12}}} H_r(X_1, X_2, Y), \tag{7}$$

where the second line follows from the definition of $\Delta_{p_{1,2,12}}$. Since the first two terms are constant, an upper bound on S requires us to look amongst all multimodal distributions $r \in \Delta$ which match the unimodal $\mathcal{D}_i$ and unlabeled multimodal data $\mathcal{D}_M$, and find the one with minimum entropy.

Theorem 3. *Solving $r^* = \arg\min_{r\in\Delta_{p_{1,2,12}}} H_r(X_1, X_2, Y)$ is NP-hard, even for a fixed $|\mathcal{Y}| \geq 4$.*

Theorem 3 suggests we cannot tractably find a joint distribution which tightly upper bounds synergy when the feature spaces are large. Thus, our proposed upper bound $\overline{S}$ is based on a lower bound on $\min_{r\in\Delta_{p_{1,2,12}}} H_r(X_1, X_2, Y)$, which yields

Theorem 4. *(Upper-bound on synergy)*

$$S \leq H_p(X_1, X_2) + H_p(Y) - \min_{r\in\Delta_{p_{12,y}}} H_r(X_1, X_2, Y) - \max_{q\in\Delta_{p_{1,2}}} I_q(\{X_1, X_2\}; Y) = \overline{S} \tag{8}$$

where $\Delta_{p_{12,y}} = \{r \in \Delta : r(x_1, x_2) = p(x_1, x_2), r(y) = p(y)\}$. The second optimization problem is solved with convex optimization. The first is the classic *min-entropy coupling* over (X_1, X_2) and Y, which is still NP-hard but admits good approximations [14, 15, 47, 65, 17, 18]. Proofs of Theorem 3, 4, and approximations for min-entropy couplings are deferred to Appendix A.5 and A.6.

4 Experiments

We design comprehensive experiments to validate these estimated bounds and show new relationships between disagreement, multimodal interactions, and performance, before describing two applications in (1) estimating optimal multimodal performance without multimodal data to prioritize the *collection* and *fusion* data sources, and (2) a new disagreement-based self-supervised learning method.

4.1 Verifying predicted guarantees and analysis of multimodal distributions

Synthetic bitwise datasets: We enumerate joint distributions over $\mathcal{X}_1, \mathcal{X}_2, \mathcal{Y} \in \{0, 1\}$ by sampling 100,000 vectors in the 8-dimensional probability simplex and assigning them to each $p(x_1, x_2, y)$. Using these distributions, we estimate $\hat{p}(y|x_1)$ and $\hat{p}(y|x_2)$ to compute disagreement and the marginals $\hat{p}(x_1, y)$, $\hat{p}(x_2, y)$, and $\hat{p}(x_1, x_2)$ to estimate the lower and upper bounds.

Large real-world multimodal datasets: We also use the large collection of real-world datasets in MultiBench [53]: (1) MOSI: video-based sentiment analysis [91], (2) MOSEI: video-based sentiment and emotion analysis [94], (3) MUSTARD: video-based sarcasm detection [12], (5) MIMIC: mortality and disease prediction from tabular patient data and medical sensors [45], and (6) ENRICO: classification of mobile user interfaces and screenshots [50]. While the previous bitwise datasets with small and discrete support yield exact lower and upper bounds, this new setting with high-dimensional continuous modalities requires the approximation of disagreement and information-theoretic quantities: we train unimodal neural network classifiers $\hat{f}_\theta(y|x_1)$ and $\hat{f}_\theta(y|x_2)$ to estimate disagreement, and we cluster representations of X_i to approximate the continuous modalities by discrete distributions with finite support to compute lower and upper bounds. We summarize the following regarding the utility of each bound (see details in Appendix B):

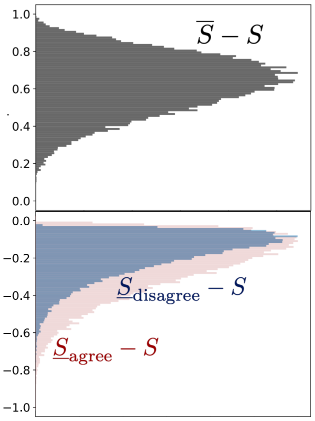

Figure 2: Our two lower bounds $\underline{S}_{\text{agree}}$ and $\underline{S}_{\text{disagree}}$ track actual synergy S from below, and the upper bound $\overline{S}$ tracks S from above. We find that $\underline{S}_{\text{agree}}$, $\underline{S}_{\text{disagree}}$ tend to approximate S better than $\overline{S}$.

1. Overall trends: For the 100,000 bitwise distributions, we compute S, the true value of synergy assuming oracle knowledge of the full multimodal distribution, and compute $\underline{S}_{\text{agree}} - S$, $\underline{S}_{\text{disagree}} - S$, and $S - \overline{S}$ for each point. Plotting these points as a histogram in Figure 2, we find that the two lower bounds track actual synergy from below ($\underline{S}_{\text{agree}} - S$ and $\underline{S}_{\text{disagree}} - S$ approaching 0 from below), and the upper bound tracks

Table 1: We compute lower and upper bounds on S without labeled multimodal data and compare them to the true S assuming knowledge of the full joint distribution p: the bounds track S well on MUSTARD and MIMIC.

	MOSEI	UR-FUNNY	MOSI	MUSTARD	MIMIC	ENRICO
$\overline{S}$	0.97	0.97	0.92	0.79	0.41	2.09
S	0.03	0.18	0.24	0.44	0.02	0.34
$\underline{S}_{\text{agree}}$	0	0	0.01	0.04	0	0.01
$\underline{S}_{\text{disagree}}$	0.01	0.01	0.03	0.11	-0.12	-0.55

x_1	x_2	y	p
0	0	0	0
0	0	1	0.05
0	1	0	0.03
0	1	1	0.28
1	0	0	0.53
1	0	1	0.03
1	1	0	0.01
1	1	1	0.06

(a) Disagreement XOR

x_1	x_2	y	p
0	0	0	0.25
0	1	1	0.25
1	0	1	0.25
1	1	0	0.25

(b) Agreement XOR

x_1	x_2	y	p
0	0	0	0.25
0	1	0	0.25
1	0	1	0.25
1	1	1	0.25

(c) $y = x_1$

x_1	x_2	y	p
0	0	0	0.5
1	1	1	0.5

(d) $y = x_1 = x_2$

Table 2: Four representative examples: (a) disagreement XOR has high disagreement and high synergy, (b) agreement XOR has no disagreement and high synergy, (c) $y = x_1$ has high disagreement and uniqueness but no synergy, and (d) $y = x_1 = x_2$ has all agreement and redundancy but no synergy.

synergy from above ($S - \overline{S}$ approaching 0 from above). The two lower bounds are quite tight, as we see that for many points $\underline{S}_{\text{agree}} - S$ and $\underline{S}_{\text{disagree}} - S$ are approaching close to 0, with an average gap of 0.18. The disagreement bound seems to be tighter empirically than the agreement bound: for half the points, $\underline{S}_{\text{disagree}}$ is within 0.14 and $\underline{S}_{\text{agree}}$ is within 0.2 of S. For the upper bound, there is an average gap of 0.62. However, it performs especially well on high synergy data. When $S > 0.6$, the average gap is 0.24, with more than half of the points within 0.25 of S.

On real-world MultiBench datasets, we show the estimated bounds and actual S (assuming knowledge of full p) in Table 1. The lower and upper bounds track true S: as estimated $\underline{S}_{\text{agree}}$ and $\underline{S}_{\text{disagree}}$ increases from MOSEI to UR-FUNNY to MOSI to MUSTARD, true S also increases. For datasets like MIMIC with disagreement but high uniqueness, $\underline{S}_{\text{disagree}}$ can be negative, but we can rely on $\underline{S}_{\text{agree}}$ to give a tight estimate on low synergy. Unfortunately, our bounds do not track synergy well on ENRICO. We believe this is because ENRICO displays all interactions: $R = 0.73, U_1 = 0.38, U_2 = 0.53, S = 0.34$, which makes it difficult to distinguish between R and S using $\underline{S}_{\text{agree}}$ or U and S using $\underline{S}_{\text{disagree}}$ since no interaction dominates over others, and $\overline{S}$ is also quite loose relative to the lower bounds. Given these general observations, we now carefully analyze the relationships between interactions, agreement, and disagreement.

2. The relationship between redundancy and synergy: In Table 2b we show the classic AGREEMENT XOR distribution where X_1 and X_2 are independent, but $Y = 1$ sets $X_1 \neq X_2$ to increase their dependence. $I(X_1; X_2; Y)$ is negative, and $\underline{S}_{\text{agree}} = 1 \leq 1 = S$ is tight. On the other hand, Table 2d is an extreme example where the probability mass distributed uniformly only when $y = x_1 = x_2$ and 0 elsewhere. As a result, X_1 is always equal to X_2 (perfect dependence), and yet Y perfectly explains away the dependence between X_1 and X_2 so $I(X_1; X_2|Y) = 0$: $\underline{S}_{\text{agree}} = 0 \leq 0 = S$. A real-world example is multimodal sentiment analysis from text, video, and audio on MOSEI, $R = 0.26$ and $S = 0.03$, and as expected the lower bound is small $\underline{S}_{\text{agree}} = 0 \leq 0.03 = S$ (Table 1).

3. The relationship between disagreement and synergy: In Table 2a we show an example called DISAGREEMENT XOR. There is maximum disagreement between marginals $p(y|x_1)$ and $p(y|x_2)$: the likelihood for y is high when y is the opposite bit as x_1, but reversed for x_2. Given both x_1 and x_2: y seems to take a 'disagreement' XOR of the individual marginals, i.e. $p(y|x_1, x_2) = \arg\max_y p(y|x_1)$ XOR $\arg\max_y p(y|x_2)$, which indicates synergy (note that an exact XOR would imply perfect agreement and high synergy). The actual disagreement is 0.15, synergy is 0.16, and uniqueness is 0.02, indicating a very strong lower bound $\underline{S}_{\text{disagree}} = 0.14 \leq 0.16 = S$. A real-world equivalent dataset is MUSTARD, where the presence of sarcasm is often due to a contradiction between what is expressed in language and speech, so disagreement $\alpha = 0.12$ is the highest out of all the video datasets, giving a lower bound $\underline{S}_{\text{disagree}} = 0.11 \leq 0.44 = S$.

Table 3: Estimated bounds $(\underline{P}_{\text{acc}}(f_M^*), \overline{P}_{\text{acc}}(f_M^*))$ on optimal multimodal performance in comparison with the best unimodal performance $P_{\text{acc}}(f_i)$, best simple fusion $P_{\text{acc}}(f_{M\text{simple}})$, and best complex fusion $P_{\text{acc}}(f_{M\text{complex}})$.

	MOSEI	UR-FUNNY	MOSI	MUSTARD	MIMIC	ENRICO
$\overline{P}_{\text{acc}}(f_M^*)$	1.07	1.21	1.29	1.63	1.27	0.88
$P_{\text{acc}}(f_{M\text{complex}})$	0.88	0.77	0.86	0.79	0.92	0.51
$P_{\text{acc}}(f_{M\text{simple}})$	0.85	0.76	0.84	0.74	0.92	0.49
$P_{\text{acc}}(f_i)$	0.82	0.74	0.83	0.74	0.92	0.47
$\underline{P}_{\text{acc}}(f_M^*)$	0.52	0.58	0.62	0.78	0.76	0.48

On the contrary, the lower bound is low when all disagreement is explained by uniqueness (e.g., $y = x_1$, Table 2c), which results in $\underline{S}_{\text{disagree}} = 0 \le 0 = S$ (α and U cancel each other out). A real-world equivalent is MIMIC: from Table 1, disagreement is high $\alpha = 0.13$ due to unique information $U_1 = 0.25$, so the lower bound informs us about the lack of synergy $\underline{S}_{\text{disagree}} = -0.12 \le 0.02 = S$. Finally, the lower bound is loose when there is synergy without disagreement, such as AGREEMENT XOR ($y = x_1$ XOR x_2, Table 2b) where the marginals $p(y|x_i)$ are both uniform, but there is full synergy: $\underline{S}_{\text{disagree}} = 0 \le 1 = S$. Real-world datasets which fall into agreement synergy include UR-FUNNY where there is low disagreement in predicting humor $\alpha = 0.03$, and relatively high synergy $S = 0.18$, which results in a loose lower bound $\underline{S}_{\text{disagree}} = 0.01 \le 0.18 = S$.

4. On upper bounds for synergy: Finally, we find that the upper bound for MUSTARD is quite close to real synergy, $\overline{S} = 0.79 \ge 0.44 = S$. On MIMIC, the upper bound is the lowest $\overline{S} = 0.41$, matching the lowest $S = 0.02$. Some of the other examples in Table 1 show bounds that are quite weak. This could be because (i) there indeed exists high synergy distributions which match $\mathcal{D}_i$ and $\mathcal{D}_M$, but these are rare in the real world, or (ii) our approximation used in Theorem 4 is mathematically loose. We leave these as open directions for future work.

4.2 Application 1: Estimating multimodal performance for multimodal fusion

Now that we have validated the accuracy of these lower and upper bounds, we can apply them towards estimating multimodal performance without labeled multimodal data. This serves as a strong signal for deciding (1) whether to collect paired and labeled data from a second modality, and (2) whether one should use complex fusion techniques on collected multimodal data.

Method: Our approach for answering these two questions is as follows: given $\mathcal{D}_1$, $\mathcal{D}_2$, and $\mathcal{D}_M$, we can estimate synergistic information based on our derived lower and upper bounds $\underline{S}$ and $\overline{S}$. Together with redundant and unique information which can be computed exactly, we will use the total information to estimate the performance of multimodal models trained optimally on the full multimodal distribution. Formally, we estimate optimal performance via a result from Feder and Merhav [25] and Fano's inequality [23], which together yield tight bounds of performance as a function of total information $I_p(\{X_1, X_2\}; Y)$.

Theorem 5. *Let $P_{acc}(f_M^*) = \mathbb{E}_p\left[\mathbf{1}\left[f_M^*(x_1, x_2) = y\right]\right]$ denote the accuracy of the Bayes' optimal multimodal model f_M^* (i.e., $P_{acc}(f_M^*) \ge P_{acc}(f'_M)$ for all $f'_M \in \mathcal{F}_M$). We have that*

$$2^{I_p(\{X_1,X_2\};Y) - H(Y)} \le P_{acc}(f_M^*) \le \frac{I_p(\{X_1, X_2\}; Y) + 1}{\log |\mathcal{Y}|}, \tag{9}$$

where we can plug in $R + U_1, U_2 + \underline{S} \le I_p(\{X_1, X_2\}; Y) \le R + U_1, U_2 + \overline{S}$ to obtain lower $\underline{P}_{\text{acc}}(f_M^*)$ and upper $\overline{P}_{\text{acc}}(f_M^*)$ bounds on optimal multimodal performance (refer to Appendix C for full proof). Finally, we summarize estimated multimodal performance as the average $\hat{P}_M = (\underline{P}_{\text{acc}}(f_M^*) + \overline{P}_{\text{acc}}(f_M^*))/2$. A high $\hat{P}_M$ suggests the presence of important joint information from both modalities (not present in each) which could boost accuracy, so it is worthwhile to collect the full distribution p and explore multimodal fusion [56] to learn joint information over unimodal methods.

Results: For each MultiBench dataset, we implement a suite of unimodal and multimodel models spanning simple and complex fusion. Unimodal models are trained and evaluated separately on each modality. Simple fusion includes ensembling by taking an additive or majority vote between unimodal models [36]. Complex fusion is designed to learn higher-order interactions as exemplified by bilinear pooling [28], multiplicative interactions [43], tensor fusion [92, 39, 52, 58], and cross-modal self-attention [78, 88]. See Appendix C for models and training details. We include unimodal, simple and complex multimodal performance, as well as estimated lower and upper bounds on optimal multimodal performance in Table 3.

RQ1: Should I collect multimodal data? We compare estimated performance $\hat{P}_M$ with the actual difference between unimodal and best multimodal performance in Figure 3 (left). Higher estimated $\hat{P}_M$ correlates with a larger gain from unimodal to multimodal. MUSTARD and ENRICO show the most opportunity for multimodal modeling, but MIMIC shows less improvement.

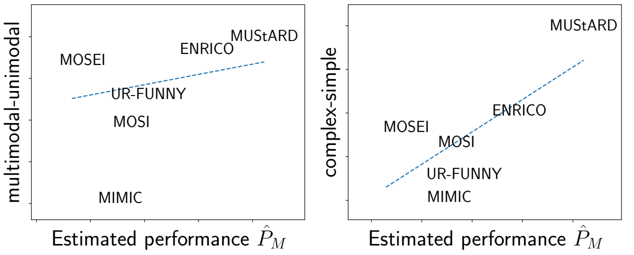

Figure 3: Datasets with higher estimated multimodal performance $\hat{P}_M$ tend to show improvements from unimodal to multimodal (left) and from simple to complex multimodal fusion (right).

RQ2: Should I investigate multimodal fusion? From Table 3, synergistic datasets like MUSTARD and ENRICO show best reported multimodal performance only slightly above the estimated lower bound, indicating more work to be done in multimodal fusion. For datasets with less synergy like MOSEI and MIMIC, the best multimodal performance is much higher than the estimated lower bound, indicating that existing fusion methods may already be quite optimal. We compare $\hat{P}_M$ with the performance gap between complex and simple fusion methods in Figure 3 (right). We again observe trends between higher $\hat{P}_M$ and improvements with complex fusion, with large gains on MUSTARD and ENRICO. We expect new methods to further improve the state-of-the-art on these datasets due to their generally high interaction values and low multimodal performance relative to estimated lower bound $\underline{P}_{\text{acc}}(f^*_M)$.

4.3 Application 2: Self-supervised multimodal learning via disagreement

Finally, we highlight an application of our analysis towards self-supervised pre-training, which is generally performed by encouraging agreement as a pre-training signal on large-scale unlabeled data [64, 68] before supervised fine-tuning [61]. However, our results suggest that there are regimes where disagreement can lead to synergy that may otherwise be ignored when only training for agreement. We therefore design a new family of self-supervised learning objectives that capture *disagreement* on unlabeled multimodal data.

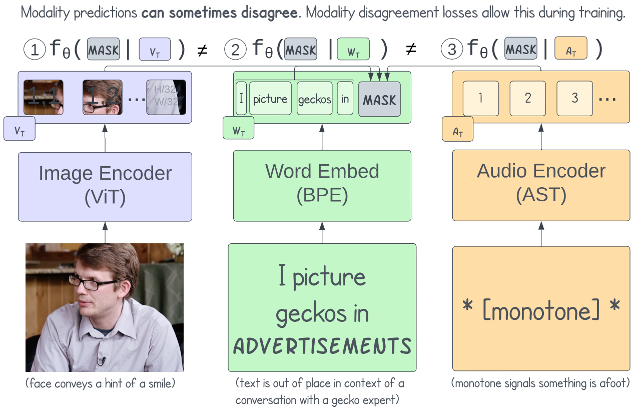
Modality predictions can sometimes disagree. Modality disagreement losses allow this during training.
① f$_\theta$(MASK | V$_T$) ≠ ② f$_\theta$(MASK | W$_T$) ≠ ③ f$_\theta$(MASK | A$_T$)
1,1
1,2
H/32, W/32
V$_T$
I
picture
geckos
in
MASK
W$_T$
1
2
3
A$_T$
Image Encoder (ViT)
Word Embed (BPE)
Audio Encoder (AST)
I picture geckos in ADVERTISEMENTS
* [monotone] *
(face conveys a hint of a smile)
(text is out of place in context of a conversation with a gecko expert)
(monotone signals something is afoot)

Figure 4: Masked predictions do not always agree across modalities, as shown in this example from the Social-IQ dataset [93]. Adding a slack term enabling pre-training with modality disagreement yields strong performance improvement over baselines.

Method: We build upon masked prediction that is popular in self-supervised pre-training: given multimodal data of the form $(x_1, x_2) \sim p(x_1, x_2)$ (e.g., x_1 = caption and x_2 = image), first mask out some words (x_1') before using the remaining words ($x_1 \backslash x_1'$) to predict the masked words via learning $f_\theta(x_1' | x_1 \backslash x_1')$, as well as the image x_2 to predict the masked words via learning $f_\theta(x_1' | x_2)$ [68, 95]. In other words, maximizing agreement between $f_\theta(x_1' | x_1 \backslash x_1')$ and $f_\theta(x_1' | x_2)$ in predicting x_1':

$$\mathcal{L}_{\text{agree}} = d(f_\theta(x_1' | x_1 \backslash x_1'), x_1') + d(f_\theta(x_1' | x_2), x_1') \tag{10}$$

for a distance d such as cross-entropy loss for discrete word tokens. To account for disagreement, *we allow predictions on the masked tokens x_1' from two different modalities i, j to disagree by a slack variable λ_{ij}*. We modify the objective such that each term only incurs a loss penalty if each distance $d(x, y)$ is larger than λ as measured by a margin distance $d_\lambda(x, y) = \max(0, d(x, y) - \lambda)$:

$$\mathcal{L}_{\text{disagree}} = \mathcal{L}_{\text{agree}} + \sum_{1 \le i < j \le 2} d_{\lambda_{ij}}(f_\theta(x_1' | x_i), f_\theta(x_1' | x_j)) \tag{11}$$

These λ terms are hyperparameters, quantifying the amount of disagreement we tolerate between each pair of modalities during cross-modal masked pretraining ($\lambda = 0$ recovers full agreement). We show this visually in Figure 4 by applying it to masked pre-training on text, video, and audio using MERLOT Reserve [95], and also apply it to FLAVA [68] for images and text experiments (see extensions to 3 modalities and details in Appendix D).

Table 4: Allowing for disagreement during self-supervised masked pre-training yields performance improvements on these datasets. Over 10 runs, improvements that are statistically significant are shown in bold ($p < 0.05$).

	SOCIAL-IQ	UR-FUNNY	MUSTARD	CARTOON
FLAVA [68], MERLOT Reserve [95]	70.6 ± 0.6	80.0 ± 0.7	77.4 ± 0.8	38.6 ± 0.6
+ disagreement	$\mathbf{71.1} \pm \mathbf{0.5}$	$\mathbf{80.7} \pm \mathbf{0.5}$	$\mathbf{78.1} \pm \mathbf{1.1}$	39.3 ± 0.5

Setup: We choose four settings with natural disagreement: (1) UR-FUNNY: humor detection from $16,000$ TED talk videos [35], (2) MUSTARD: 690 videos for sarcasm detection from TV shows [12], (3) SOCIAL IQ: $1,250$ multi-party videos testing social intelligence knowledge [93], and (4) CARTOON: matching 704 cartoon images and captions [38].

Results: From Table 4, allowing for disagreement yields improvements on these datasets, with those on SOCIAL IQ, UR-FUNNY, MUSTARD being statistically significant (p-value < 0.05 over 10 runs). By analyzing the value of λ resulting in the best validation performance through hyperparameter search, we can analyze when disagreement helps for which datasets, datapoints, and modalities. On a dataset level, we find that disagreement helps for video/audio and video/text, improving accuracy by up to 0.6% but hurts for text/audio, decreasing the accuracy by up to 1%. This is in line with intuition, where spoken text is transcribed directly from audio for these monologue and dialog videos, but video can have vastly different information. In addition, we find more disagreement between text/audio for SOCIAL IQ, which we believe is because it comes from natural videos while the others are scripted TV shows with more agreement between speakers and transcripts.

We further analyze individual datapoints with disagreement On UR-FUNNY, the moments when the camera jumps from the speaker to their presentation slides are followed by an increase in agreement since the video aligns better with the speech. In MUSTARD, we observe disagreement between vision and text when the speaker's face expresses the sarcastic nature of a phrase. This changes the meaning of the phrase, which cannot be inferred from text only, and leads to synergy. We include more qualitative examples including those on the CARTOON captioning dataset in Appendix D.

5 Related Work

Multivariate information theory: The extension of information theory to 3 or more variables [86, 29, 72, 60, 74, 30] remains on open problem. Partial information decomposition (PID) [87] was proposed as a potential solution that satisfies several appealing properties [10, 33, 83, 87]. Today, PID has primarily found applications in cryptography [59, 42], neuroscience [63], physics [26], complex systems [69], and biology [16], but its application towards machine learning, in particular multimodality, is an exciting but untapped research direction. To the best of our knowledge, our work is the first to provide formal estimates of synergy in the context of unlabeled or unpaired multimodal data which is common in today's self-supervised paradigm [55, 64, 68, 95].

Understanding multimodal models: Information theory is useful for understanding co-training [11, 5, 13], multi-view learning [77, 80, 76, 71], and feature selection [89], where redundancy is an important concept. Prior research has also studied multimodal models via additive or non-additive interactions [27, 70, 37], gradient-based approaches [81], or visualization tools [85]. This goal of quantifying and modeling multimodal interactions [57] has also motivated many successful learning algorithms, such as contrastive learning [46, 64], agreement and alignment [21, 54], factorized representations [79], as well as tensors and multiplicative interactions [92, 52, 43].

Disagreement-based learning has been used to estimate performance from unlabeled data [4, 44], active learning [19, 34], and guiding exploration in reinforcement learning [62, 66]. In multimodal learning, however, approaches have been primarily based on encouraging agreement in prediction [11, 21, 24, 71] or feature space [64, 61] in order to capture shared information. Our work has arrived at similar conclusions regarding the benefits of disagreement-based learning, albeit from different mathematical motivations and applications.

6 Conclusion

We proposed estimators of multimodal interactions when observing only *labeled unimodal data* and some *unlabeled multimodal data*, a general setting that encompasses many real-world constraints involving partially observable modalities, limited labels, and privacy concerns. Our key results draw new connections between multimodal interactions, the disagreement of unimodal classifiers, and min-entropy couplings. **Future work** should investigate more applications of multivariate information theory in designing self-supervised models, predicting multimodal performance, and other tasks involving feature interactions such as privacy-preserving and fair representation learning.

References

[1] Armen Aghajanyan, Lili Yu, Alexis Conneau, Wei-Ning Hsu, Karen Hambardzumyan, Susan Zhang, Stephen Roller, Naman Goyal, Omer Levy, and Luke Zettlemoyer. Scaling laws for generative mixed-modal language models. *arXiv preprint arXiv:2301.03728*, 2023.

[2] Paul D Allison. Testing for interaction in multiple regression. *American journal of sociology*, 83(1): 144–153, 1977.

[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.

[4] Christina Baek, Yiding Jiang, Aditi Raghunathan, and J Zico Kolter. Agreement-on-the-line: Predicting the performance of neural networks under distribution shift. *Advances in Neural Information Processing Systems*, 35:19274–19289, 2022.

[5] Maria-Florina Balcan, Avrim Blum, and Ke Yang. Co-training and expansion: Towards bridging theory and practice. *Advances in neural information processing systems*, 17, 2004.

[6] 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.

[7] Reuben M Baron and David A Kenny. The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of personality and social psychology*, 51(6):1173, 1986.

[8] Anthony J Bell. The co-information lattice. In *Proceedings of the fifth international workshop on independent component analysis and blind signal separation: ICA*, volume 2003, 2003.

[9] Samy Bengio and Hervé Bourlard. *Machine learning for multimodal interaction*. Springer, 2005.

[10] Nils Bertschinger, Johannes Rauh, Eckehard Olbrich, Jürgen Jost, and Nihat Ay. Quantifying unique information. *Entropy*, 2014.

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

[12] Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. Towards multimodal sarcasm detection (an _obviously_ perfect paper). In *ACL*, pages 4619–4629, 2019.

[13] C Mario Christoudias, Raquel Urtasun, and Trevor Darrell. Multi-view learning in the presence of view disagreement. In *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*, 2008.

[14] Ferdinando Cicalese and Ugo Vaccaro. Supermodularity and subadditivity properties of the entropy on the majorization lattice. *IEEE Transactions on Information Theory*, 48(4):933–938, 2002.

[15] Ferdinando Cicalese, Luisa Gargano, and Ugo Vaccaro. How to find a joint probability distribution of minimum entropy (almost) given the marginals. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 2173–2177. IEEE, 2017.

[16] Nigel Colenbier, Frederik Van de Steen, Lucina Q Uddin, Russell A Poldrack, Vince D Calhoun, and Daniele Marinazzo. Disambiguating the role of blood flow and global signal with partial information decomposition. *Neuroimage*, 213:116699, 2020.

[17] Spencer Compton. A tighter approximation guarantee for greedy minimum entropy coupling. In *2022 IEEE International Symposium on Information Theory (ISIT)*, pages 168–173. IEEE, 2022.

[18] Spencer Compton, Dmitriy Katz, Benjamin Qi, Kristjan Greenewald, and Murat Kocaoglu. Minimum-entropy coupling approximation guarantees beyond the majorization barrier. In *International Conference on Artificial Intelligence and Statistics*, pages 10445–10469. PMLR, 2023.

[19] Corinna Cortes, Giulia DeSalvo, Mehryar Mohri, Ningshan Zhang, and Claudio Gentile. Active learning with disagreement graphs. In *International Conference on Machine Learning*, pages 1379–1387. PMLR, 2019.

[20] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[21] Daisy Yi Ding, Shuangning Li, Balasubramanian Narasimhan, and Robert Tibshirani. Cooperative learning for multiview analysis. *Proceedings of the National Academy of Sciences*, 119(38):e2202113119, 2022.

[22] Shimon Even, Alon Itai, and Adi Shamir. On the complexity of time table and multi-commodity flow problems. In *16th annual symposium on foundations of computer science (sfcs 1975)*, pages 184–193. IEEE, 1975.

[23] Robert M Fano. *Transmission of information: a statistical theory of communications*. Mit Press, 1968.

[24] Jason Farquhar, David Hardoon, Hongying Meng, John Shawe-Taylor, and Sandor Szedmak. Two view learning: Svm-2k, theory and practice. *NeurIPS*, 18, 2005.

[25] Meir Feder and Neri Merhav. Relations between entropy and error probability. *IEEE Transactions on Information theory*, 40(1):259–266, 1994.

[26] Benjamin Flecker, Wesley Alford, John M Beggs, Paul L Williams, and Randall D Beer. Partial information decomposition as a spatiotemporal filter. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 2011.

[27] Jerome H Friedman and Bogdan E Popescu. Predictive learning via rule ensembles. *The annals of applied statistics*, 2(3):916–954, 2008.

[28] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *Conference on Empirical Methods in Natural Language Processing*. ACL, 2016.

[29] Wendell R Garner. Uncertainty and structure as psychological concepts. 1962.

[30] Timothy J Gawne and Barry J Richmond. How independent are the messages carried by adjacent inferior temporal cortical neurons? *Journal of Neuroscience*, 13(7):2758–2771, 1993.

[31] AmirEmad Ghassami and Negar Kiyavash. Interaction information for causal inference: The case of directed triangle. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 1326–1330. IEEE, 2017.

[32] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913, 2017.

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

[34] Steve Hanneke et al. Theory of disagreement-based active learning. *Foundations and Trends® in Machine Learning*, 7(2-3):131–309, 2014.

[35] Md Kamrul Hasan, Wasifur Rahman, AmirAli Bagher Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, and Mohammed Ehsan Hoque. Ur-funny: A multimodal language dataset for understanding humor. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2046–2056, 2019.

[36] Trevor Hastie and Robert Tibshirani. Generalized additive models: some applications. *Journal of the American Statistical Association*, 1987.

[37] Jack Hessel and Lillian Lee. Does my multimodal model learn cross-modal interactions? it's harder to tell than you might think! In *EMNLP*, 2020.

[38] Jack Hessel, Ana Marasović, Jena D Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. Do androids laugh at electric sheep? humor" understanding" benchmarks from the new yorker caption contest. *arXiv preprint arXiv:2209.06293*, 2022.

[39] Ming Hou, Jiajia Tang, Jianhai Zhang, Wanzeng Kong, and Qibin Zhao. Deep multimodal multilinear fusion with high-order polynomial pooling. *Advances in Neural Information Processing Systems*, 32: 12136–12145, 2019.

[40] Tzu-Ming Harry Hsu, Wei-Hung Weng, Willie Boag, Matthew McDermott, and Peter Szolovits. Unsupervised multimodal representation learning across medical images and reports. *arXiv preprint arXiv:1811.08615*, 2018.

[41] 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.

[42] Ryan G James, Jeffrey Emenheiser, and James P Crutchfield. Unique information and secret key agreement. *Entropy*, 21(1):12, 2018.

[43] Siddhant M. Jayakumar, Wojciech M. Czarnecki, Jacob Menick, Jonathan Schwarz, Jack Rae, Simon Osindero, Yee Whye Teh, Tim Harley, and Razvan Pascanu. Multiplicative interactions and where to find them. In *International Conference on Learning Representations*, 2020.

[44] Yiding Jiang, Vaishnavh Nagarajan, Christina Baek, and J Zico Kolter. Assessing generalization of sgd via disagreement. In *International Conference on Learning Representations*, 2022.

[45] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

[46] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021.

[47] Murat Kocaoglu, Alexandros Dimakis, Sriram Vishwanath, and Babak Hassibi. Entropic causal inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

[48] Mladen Kovačević, Ivan Stanojević, and Vojin Šenk. On the entropy of couplings. *Information and Computation*, 242:369–382, 2015.

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

[50] Luis A Leiva, Asutosh Hota, and Antti Oulasvirta. Enrico: A dataset for topic modeling of mobile ui designs. In *22nd International Conference on Human-Computer Interaction with Mobile Devices and Services*, pages 1–4, 2020.

[51] Paul Pu Liang, Ruslan Salakhutdinov, and Louis-Philippe Morency. Computational modeling of human multimodal language: The mosei dataset and interpretable dynamic fusion.

[52] Paul Pu Liang, Zhun Liu, Yao-Hung Hubert Tsai, Qibin Zhao, Ruslan Salakhutdinov, and Louis-Philippe Morency. Learning representations from imperfect time series data via tensor rank regularization. In *ACL*, 2019.

[53] Paul Pu Liang, Yiwei Lyu, Xiang Fan, Zetian Wu, Yun Cheng, Jason Wu, Leslie Yufan Chen, Peter Wu, Michelle A Lee, Yuke Zhu, et al. Multibench: Multiscale benchmarks for multimodal representation learning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021.

[54] Paul Pu Liang, Peter Wu, Liu Ziyin, Louis-Philippe Morency, and Ruslan Salakhutdinov. Cross-modal generalization: Learning in low resource modalities via meta-alignment. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2680–2689, 2021.

[55] Paul Pu Liang, Yiwei Lyu, Xiang Fan, Shengtong Mo, Dani Yogatama, Louis-Philippe Morency, and Ruslan Salakhutdinov. Highmmt: Towards modality and task generalization for high-modality representation learning. *arXiv preprint arXiv:2203.01311*, 2022.

[56] Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Foundations and recent trends in multimodal machine learning: Principles, challenges, and open questions. *arXiv preprint arXiv:2209.03430*, 2022.

[57] 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.

[58] Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe Morency. Efficient low-rank multimodal fusion with modality-specific factors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2247–2256, 2018.

[59] Ueli M Maurer and Stefan Wolf. Unconditionally secure key agreement and the intrinsic conditional information. *IEEE Transactions on Information Theory*, 45(2):499–514, 1999.

[60] William McGill. Multivariate information transmission. *Transactions of the IRE Professional Group on Information Theory*, 4(4):93–111, 1954.

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

[62] Deepak Pathak, Dhiraj Gandhi, and Abhinav Gupta. Self-supervised exploration via disagreement. In *International conference on machine learning*, pages 5062–5071. PMLR, 2019.

[63] Giuseppe Pica, Eugenio Piasini, Houman Safaai, Caroline Runyan, Christopher Harvey, Mathew Diamond, Christoph Kayser, Tommaso Fellin, and Stefano Panzeri. Quantifying how much sensory information in a neural code is relevant for behavior. *Advances in Neural Information Processing Systems*, 30, 2017.

[64] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[65] Massimiliano Rossi. Greedy additive approximation algorithms for minimum-entropy coupling problem. In *2019 IEEE International Symposium on Information Theory (ISIT)*, pages 1127–1131. IEEE, 2019.

[66] Ramanan Sekar, Oleh Rybkin, Kostas Daniilidis, Pieter Abbeel, Danijar Hafner, and Deepak Pathak. Planning to explore via self-supervised world models. In *International Conference on Machine Learning*, pages 8583–8592. PMLR, 2020.

[67] Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27 (3):379–423, 1948.

[68] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15638–15650, 2022.

[69] Sten Sootla, Dirk Oliver Theis, and Raul Vicente. Analyzing information distribution in complex systems. *Entropy*, 19(12):636, 2017.

[70] Daria Sorokina, Rich Caruana, Mirek Riedewald, and Daniel Fink. Detecting statistical interactions with additive groves of trees. In *Proceedings of the 25th international conference on Machine learning*, pages 1000–1007, 2008.

[71] Karthik Sridharan and Sham M Kakade. An information theoretic framework for multi-view learning. In *Conference on Learning Theory*, 2008.

[72] Milan Studenỳ and Jirina Vejnarová. The multiinformation function as a tool for measuring stochastic dependence. In *Learning in graphical models*, pages 261–297. Springer, 1998.

[73] Masahiro Suzuki, Kotaro Nakayama, and Yutaka Matsuo. Joint multimodal learning with deep generative models. *arXiv preprint arXiv:1611.01891*, 2016.

[74] Han Te Sun. Multiple mutual informations and multiple interactions in frequency data. *Inf. Control*, 46: 26–45, 1980.

[75] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 776–794. Springer, 2020.

[76] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? *Advances in Neural Information Processing Systems*, 33, 2020.

[77] Christopher Tosh, Akshay Krishnamurthy, and Daniel Hsu. Contrastive learning, multi-view redundancy, and linear models. In *Algorithmic Learning Theory*, pages 1179–1206. PMLR, 2021.

[78] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6558–6569, 2019.

[79] Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. Learning factorized multimodal representations. In *International Conference on Learning Representations*, 2019.

[80] Yao-Hung Hubert Tsai, Yue Wu, Ruslan Salakhutdinov, and Louis-Philippe Morency. Self-supervised learning from a multi-view perspective. In *International Conference on Learning Representations*, 2020.

[81] Michael Tsang, Dehua Cheng, and Yan Liu. Detecting statistical interactions from neural network weights. In *International Conference on Learning Representations*, 2018.

[82] Michael Tsang, Dehua Cheng, Hanpeng Liu, Xue Feng, Eric Zhou, and Yan Liu. Feature interaction interpretability: A case for explaining ad-recommendation systems via neural interaction detection. In *International Conference on Learning Representations*, 2019.

[83] Praveen Venkatesh and Gabriel Schamberg. Partial information decomposition via deficiency for multivariate gaussians. In *2022 IEEE International Symposium on Information Theory (ISIT)*, pages 2892–2897. IEEE, 2022.

[84] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018.

[85] Xingbo Wang, Jianben He, Zhihua Jin, Muqiao Yang, Yong Wang, and Huamin Qu. M2lens: Visualizing and explaining multimodal models for sentiment analysis. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):802–812, 2021.

[86] Satosi Watanabe. Information theoretical analysis of multivariate correlation. *IBM Journal of research and development*, 1960.

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

[88] Shaowei Yao and Xiaojun Wan. Multimodal transformer for multimodal machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.400. URL https://www.aclweb.org/anthology/2020.acl-main.400.

[89] Lei Yu and Huan Liu. Efficiently handling feature redundancy in high-dimensional data. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2003.

[90] Lei Yu and Huan Liu. Efficient feature selection via analysis of relevance and redundancy. *The Journal of Machine Learning Research*, 5:1205–1224, 2004.

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

[92] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1103–1114, 2017.

[93] Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. Social-iq: A question answering benchmark for artificial social intelligence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8807–8817, 2019.

[94] 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.

[95] Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. Merlot reserve: Neural script knowledge through vision and language and sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16375–16387, 2022.