Interpretable Tensor Fusion

Saurabh Varshneya[1], **Antoine Ledent**[2], **Philipp Liznerski**[1], **Andriy Balinskyy**[1], **Purvanshi Mehta**[3], **Waleed Mustafa**[1] and **Marius Kloft**[1]
[1]RPTU Kaiserslautern-Landau
[2]Singapore Management University
[3]Lica World
{varshneya, liznerski, balinskyy, mustafa, kloft}@cs.uni-kl.de, purvanshi@lica.world, aledent@smu.edu.sg

Abstract

Conventional machine learning methods are predominantly designed to predict outcomes based on a single data type. However, practical applications may encompass data of diverse types, such as text, images, and audio. We introduce interpretable tensor fusion (InTense), a multimodal learning method for training neural networks to simultaneously learn multimodal data representations and their interpretable fusion. InTense can separately capture both linear combinations and multiplicative interactions of diverse data types, thereby disentangling higher-order interactions from the individual effects of each modality. InTense provides interpretability out of the box by assigning relevance scores to modalities and their associations. The approach is theoretically grounded and yields meaningful relevance scores on multiple synthetic and real-world datasets. Experiments on six real-world datasets show that InTense outperforms existing state-of-the-art multimodal interpretable approaches in terms of accuracy and interpretability.

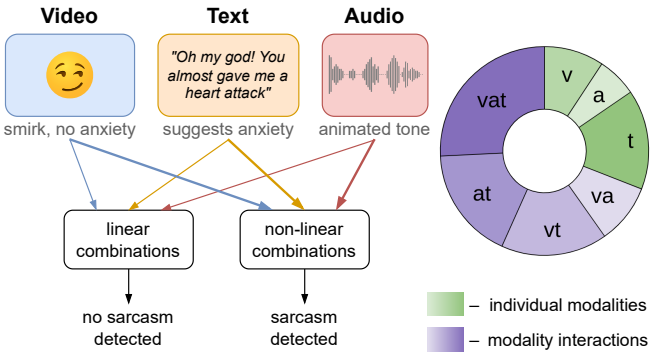

Figure 1: Left is an excerpt of the MUStARD dataset on sarcasm detection, where the proposed InTense method sets a new state-of-the-art. (See Section 4 for details.) A linear combination of modalities fails here because the expressions of happiness and anxiety combine to something neutral rather than sarcasm. To detect sarcasm, the interactions among modalities are crucial. InTense captures these interactions and assigns them with interpretable relevance scores, shown in the pie chart. Scores for individual modalities and their interactions are colored green and blue, respectively. InTense reveals that interactions are crucial for successful sarcasm detection.

1 Introduction

The vast majority of machine learning systems are designed to predict outcomes based on a single datatype or "modality". However, in various applications—spanning fields from biology and medicine to engineering and multimedia—multiple modalities are frequently in play [He *et al.*, 2020; Lunghi *et al.*, 2019]. The main challenge in multimodal learning is how to effectively fuse these diverse modalities. The most common approach is to combine the modalities in an additive way [Poria *et al.*, 2015; Chagas *et al.*, 2020]. Such linear combinations suffice in some cases. However, numerous applications necessitate capturing non-linear interactions between modalities. One such instance is sarcasm detection, described in Figure 1 [Hessel and Lee, 2020]. The arguably most popular approach to capture non-linear interactions of modalities is "Tensor fusion" [Zadeh *et al.*, 2017; Tsai *et al.*, 2019; Liang *et al.*, 2021]. The main idea in tensor fusion is to concatenate modalities via tensor products in a neural network.

A substantial drawback of tensor fusion is its inherent lack of interpretability, which can significantly hinder its application in real-world scenarios. Interpretable multimodal models may reveal the relative importance of modalities [Büchel *et al.*, 1998; Hessel and Lee, 2020], unveiling spurious modalities and social biases in the data. Identifying interactions among modalities is the main goal in several application domains. For instance, in statistical genetics, it is crucial to identify the interactions among Single Nucleotide Polymorphisms (SNPs) that contribute to the inheritance of a disease [Behravan *et al.*, 2018; Elgart *et al.*, 2022]. Although some interpretable multimodal methods exist, they are limited to linear combinations or require resource-intensive post hoc algorithms for interpretation.

In this paper, we introduce *interpretable tensor fusion* (InTense), which jointly learns multimodal neural representations and their interpretable fusion. InTense provides out-of-the-box interpretability by assigning relevance scores to all modalities and their interactions. Our approach is inspired by

Full paper with technical appendix and code is available at: https://arxiv.org/abs/2405.04671

multiple kernel learning [Kloft *et al.*, 2011], a classic kernel-based approach to interpretable multimodal learning, which we generalize to deep neural networks and term *Multiple Neural Learning* (MNL). While both MNL and InTense provide relevance scores for the modalities, InTense additionally produces scores for the interactions of modalities. These interaction scores are made possible through a novel interpretation of neural network weight matrices: We show that neural networks tend to favor higher-order tensor products, leading to spurious interpretations (i.e., overstating high-order interactions between modalities). We resolve this issue by deriving a theoretically well-founded normalization approach. In the theoretical analysis, we prove that this produces genuine relevance scores, avoiding spurious interpretations. In extensive experiments, we empirically validate the relevance scores on data and show that InTense outperforms existing state-of-the-art multimodal interpretable approaches in terms of accuracy and interpretability.

In summary, our contributions are:

- We introduce *Multiple Neural Learning* (MNL), a theoretically guided adaptation of the established Multiple Kernel Learning algorithm to deep learning.
- We introduce InTense, an extension of MNL and tensor fusion designed to capture non-linear interactions among modalities in an interpretable manner.
- We provide a rigorous theoretical analysis that provides evidence of the correct disentanglement within our fusion framework.
- We validate our approach through extensive experiments, where we meet the state-of-the-art classification accuracy while providing robust interpretability.

2 Related Work

We now review existing multimodal learning methods that produce interpretability scores for the modalities.

Interpretable Methods for Learning *Linear* Combinations of Modalities. The vast majority of interpretable multimodal learning methods consider linear combinations of modalities. The arguably most popular instance is Multiple Kernel Learning (MKL), where kernels from different modalities are combined linearly. Here, a weight is learned for each kernel determining its importance in the resulting linear combination of kernels [Kloft *et al.*, 2011; Rakotomamonjy *et al.*, 2008]. However, the performance of MKL is limited by the quality of the kernels. Finding adequate kernels can be especially problematic for structured high-dimensional data, such as text or images. Addressing this, several authors have studied combining multiple modalities using neural networks in a linear manner [Poria *et al.*, 2015; Chen *et al.*, 2014; Arabacı *et al.*, 2021]. However, these representations are independently learned to form basis kernels and later combined in a second step through an SVM or another shallow learning method. Such independently learned representations cannot properly capture modality interactions.

Methods for Learning *Non-linear* Combinations of Modalities. Hessel and Lee [2020] map neural representations to a space defined by a linear combination of the modalities. While they quantify the overall importance of non-linear interactions, they do not provide scores for individual modality interactions. Tsai *et al.* [2020] introduce multimodal routing, which is based on dynamic routing [Sabour *et al.*, 2017], to calculate scores for the modality interactions. These scores depend on the similarity of a modality's representation to so-called concept vectors, where one such vector is defined for each label. However, routing does not distinguish between linear and non-linear combinations and is thus misled by partially redundant information in the combinations. Indeed, we show through experiments (see Section 4) that the non-linear combinations learned by routing are incorrectly overestimated. Gat *et al.* [2021] propose a method to obtain modality relevances by computing differences of accuracies on a test set and a permuted test set. However, this method has limited interpretability and requires multiple forward passes through the trained network to obtain relevance scores. Wörtwein *et al.* [2022] learn an aggregated representation for unimodal, bimodal, and trimodal interactions, respectively. However, their method does not learn fine-grained relevance scores for the various combinations of modalities. Alongside methods offering limited interpretability, there exist methods that non-linearly combine modalities without adding any interpretability [Zhang *et al.*, 2023; Liang *et al.*, 2021; Tan and Bansal, 2019].

In summary, none of these methods learns proper relevance scores of interactions between modalities.

Post-hoc Explanation Methods. There exist several methods for post-hoc explanation of multimodal learning methods [Gat *et al.*, 2021; Chandrasekaran *et al.*, 2018; Park *et al.*, 2018; Kanehira *et al.*, 2019; Cao *et al.*, 2020; Frank *et al.*, 2021]. These methods consist of two steps: first, training a multimodal model that is not inherently interpretable, followed by the calculation of relevance scores in hindsight. However, their two-step nature makes these methods challenging to analyze theoretically. Moreover, since the initial model disregards interpretability, it may lead to inherent limitations in the explanatory process. Additionally, these methods come with the added computational burden of producing relevance scores. Another limitation is their applicability, which is confined to specific types of modalities.

3 Methodology

In the following sections, we introduce several components comprising our approach. First, we review the classical L_p-norm Multiple Kernel Learning (MKL) framework [Kloft *et al.*, 2011], which we extend to Multiple Neural Learning. Subsequently, we propose Interpretable Tensor Fusion (InTense), which captures non-linear modality interactions. Furthermore, we show how InTense learns disentangled neural representations, thereby computing correct relevance scores.

3.1 Preliminaries

We consider a dataset $\{(x_i, y_i)\}_{i=1}^n$ with labels $y_i \in \{-1, 1\}$. The inputs have M modalities, where $x_i^m \in \mathcal{X}^m$ for $m \in \{1, \ldots, M\}$ denotes the m^{th} modality of the datapoint x_i, and $\mathcal{X}^m$ is the input space associated with the modality m. In MKL, one considers kernel mixtures of the form $k(u, v) =$

$\sum_{m=1}^{M} \beta_m k_m(u,v)$, where $k_m(u,v)$ is a base kernel and $\beta_m \geq 0$ for all m. Imposing an L_p-norm constraint on the vector $\beta \in \mathbb{R}^M$ gives rise to the following classic optimization problem:

$$\underset{\substack{w_1,w_2,\cdots,w_L,\beta \\ \beta\in\mathbb{R}^M,\beta\geq 0,\|\beta\|_p\leq 1}}{\text{minimize}} \left(\sum_{\text{i=1}}^{\text{n}} \ell\Big(\sum_{m=1}^{M} \sqrt{\beta_m}\langle w_m, \Psi_m(x_i^m)\rangle_{\mathcal{H}^m} + b, y_i\Big) + \frac{\lambda}{2}\sum_{m=1}^{M} \|w_m\|^2_{L^2(\mathcal{H}^m)}\right), \quad (1)$$

where ℓ is a loss function, and $\Psi_m : \mathcal{X}^m \to \mathcal{H}^m$ are feature maps from the input space $\mathcal{X}^m$ to the Hilbert space $\mathcal{H}^m$ associated with kernel k_m such that for each $m \in \{1, 2, \ldots, M\}$ and $u, v \in \mathcal{X}^m$, $k_m(u,v) = \langle \Psi_m(u), \Psi_m(v)\rangle_{\mathcal{H}^m}$, where $\langle\cdot,\cdot\rangle^m_{\mathcal{H}}$ denotes the inner-product associated with the Hilbert space $\mathcal{H}^m$. The base kernels k_m are assumed to be fixed functions. This is in sharp contrast to Multiple Neural Learning (MNL) introduced in the next section, where each feature map $\Psi_m(x)$ is learned from the data.

3.2 Multiple Neural Learning

In this section, we propose Multiple Neural Learning (MNL), **an interpretable method for linear combination of modalities**. In MNL, we train a neural network composed of two components: 1) modality subnetworks that output a neural representation for each modality and 2) a linear fusion layer that combines the representations in an interpretable manner. We define the optimization problem as:

$$\underset{\substack{w_L^1,\cdots,w_L^M,\beta, \\ W^1,W^2,\cdots,W^M, \\ \beta\in\mathbb{R}^M,\beta\geq 0,\|\beta\|_p\leq 1}}{\text{minimize}} \left(\sum_{i=1}^{n} \ell\Big(\sum_{m=1}^{M} \sqrt{\beta_m}\langle w_L^m, f^m(x_i^m)\rangle + b, y_i\Big) + \Lambda \sum_{l=1}^{L}\sum_{m=1}^{M} \|w_l^m\|_2^2\right), \quad (2)$$

where f^m is the m^{th} modality's subnetwork composed of $L-1$ layers with weights $W^m = \{w_1^m, \ldots, w_{L-1}^m\}$. A representation for the i^{th} data point's m^{th} modality is obtained by $f^m(x_i^m)$. The fusion layer L with weights $w_L^1, \ldots, w_L^M$ learns a linear combination of the modality representations. Λ and p $(1 \leq p < \infty)$ are hyperparameters and $\ell(t,y) = -\log\left(\frac{\exp(ty)}{1+\exp(ty)}\right)$ is the cross-entropy loss function. This setup can also be seen as additive fusion because it represents a linear combination of the modalities with weights $\sqrt{\beta_m}$.

Notably, $\sqrt{\beta_m}$ is a positive weight for the m'th modality, indicating its relevance score. The vector β is simultaneously optimized with the network weights. However, the constraints on β introduce an increased difficulty in optimizing equation 2. The following theorem presents a simplified optimization problem by eliminating β from equation 2 along with a method to retrieve β from the learned weights.

Theorem 1. *The optimization problem in equation 2 is equivalent to the following problem, where the parameters β are no longer present:*

$$\underset{\substack{w_L^1,w_L^2,\ldots,w_L^M, \\ W^1,W^2,\cdots,W^M}}{\text{minimize}} \sum_{i=1}^{n} \ell\left(\sum_{m=1}^{M}\langle w_L^m, f^m(x_i^m)\rangle + b, y_i\right) + \Lambda\sum_{l=1}^{L-1}\sum_{m=1}^{M}\|w_l^m\|_2^2 + \Lambda\left(\sum_{m=1}^{M}\|w_L^m\|_2^q\right)^{\frac{2}{q}}, \quad (3)$$

where $q = \frac{2p}{p+1}$ (and therefore $1 \leq q \leq 2$). The corresponding values of relevance score β can be recovered after the optimization as:

$$\beta_m = \frac{\|w_L^m\|_2^{\frac{2}{p+1}}}{\left(\sum_{\tilde{m}=1}^{M}\|w_L^{\tilde{m}}\|_2^{\frac{2p}{p+1}}\right)^{\frac{1}{p}}}. \quad (4)$$

The theorem states that the relevance of a modality in our jointly trained network can be obtained by applying a suitable $p-$norm over the weights of the fusion layer L. A detailed proof of the theorem can be found in Appendix A. The central idea is observing that the parameters β can be absorbed into the weights w_L^m, pushing β into the regularization term. Subsequently, by showing that β can be minimized independently from the weights, the optimal value is attained through equation 4. Absorbing β in fusion weights in turn introduces the additional block L^q norm regularization term $\Lambda(\sum_{m=1}^{M}\|w_L^m\|_2^q)^{\frac{2}{q}}$.

Correct Relevance Scores Through Normalization. In pre-experiments (see Appendix F.1) we found that the relevance scores can be misleading, especially when the network outputs high activation values for some modalities. We address this issue with proper normalization techniques. We propose an adaptation of the standard Batch Normalization (batch norm) [Ioffe and Szegedy, 2015], which we call Vector-wise Batch Normalization (VBN). VBN ensures that the L_2-norm of the activation values, *averaged over a mini-batch*, is constant. Let B be a mini-batch of datapoints' indices. We define VBN as:

$$\tilde{f}^m(x_i^m) = \frac{f^m(x_i^m) - \mu_{B,m}}{\sigma_{B,m}}, \text{where} \quad (5)$$

$$\mu_{B,m} = \frac{\sum_{i\in B} f^m(x_i^m)}{|B|}, \text{and} \quad (6)$$

$$\sigma^2_{B,m} = \frac{\sum_{i\in B}\|f^m(x_i^m) - \mu_{B,m}\|_2^2}{|B|}. \quad (7)$$

The mean $\mu_{B,m}$ is computed element-wise as in the standard batch norm. However, in equation 7, instead of computing the variance element-wise, we calculate the average of the squared L_2-norm of a modality representation across the mini-batch. Unlike batch norm, we do not shift and scale the representations element-wise after the normalization step. Using VBN, the loss in equation 3 changes to:

$$\sum_{i=1}^{n}\ell\left(\sum_{m=1}^{M}\left\langle w_L^m, \tilde{f}^m(x_i^m)\right\rangle + b, y_i\right).$$

Note that VBN is applied *after* the activation function to obtain $\tilde{f}^m(x^m)$. We found empirically that proper normalization is crucial for MNL to achieve competitive performance.

3.3 Interpretable Tensor Fusion

In this section, we propose Interpretable Tensor Fusion (InTense), an extension of MNL that additionally produces scores for interactions (non-linear combinations) of modalities. InTense is based on tensor fusion, which captures multiplicative interactions among modalities by computing a tensor product over the individual modality representations [Zadeh *et al.*, 2017]. InTense operates as follows. For a dataset with M modalities, we consider interactions up to a given order of D, where $D \leq M$. An order of D implies interaction among D modalities. A multiplicative interaction of modalities is defined by a subset $I \in \mathcal{I}$, where $\mathcal{I} = \{J \subset \{1, \ldots, M\} : |J| \leq D\}$, and a tensor product $f^I(x) := f^{I_1} \otimes f^{I_2} \otimes \ldots f^{I_{|I|}}$, where f^{I_m} is the representation of modality I_m, and $\otimes$ denotes the tensor product operator. Analogously to equation 2, we obtain a new objective:

$$\underset{\substack{w_L^I, \beta_I : I \in \mathcal{I}, \\ \|\beta\|_p \leq 1, \beta_I \geq 0 \\ W^1, W^2, \cdots, W^M}}{\text{minimize}} \left(\sum_{i=1}^{n} \ell\Big(\sum_{I \subset \mathcal{I}} \sqrt{\beta_I} \langle w_L^I, f^I(x) \rangle + b, y_i \Big) + \Lambda \sum_{l=1}^{L} \sum_{m=1}^{M} \|w_l^m\|_2^2 \right) \quad (8)$$

This optimization problem can be seen as a special case of MNL, where the multiplicative interactions are treated as separate modalities. Therefore, in combination with equation 8, Theorem 1 computes the relevance scores for all modalities and their interactions.

What Can Go Wrong?

In our experiments with synthetic multimodal datasets (Section 4.1), we found that relevance scores of higher-order interactions are greatly overestimated. Scores can be high even when no true interactions exist in the data. We call this phenomenon *higher-order interaction bias*. The bias is caused by higher-order tensor products corresponding to very large function classes, which approximately include the function classes corresponding to lower-order tensors as subsets.

Indeed, it is possible that a linear combination of the components of a tensor product learns the same functions as a linear combination of the individual-modality representations. For instance, consider two modalities (m_u, m_v) and their representation vectors as $\mathbf{u}, \mathbf{v} \in \mathbb{R}^3$. Assume the first component of the learned representations is constant (e.g., 1), i.e., $\mathbf{u} = [1, u_2, u_3]^\intercal$ and $\mathbf{v} = [1, v_2, v_3]^\intercal$. In such a scenario, the linear combination $\alpha_1 u_2 + \alpha_2 v_2$ (for $\alpha_1, \alpha_2 \in \mathbb{R}$) can also be represented as $\alpha_1(\mathbf{u} \otimes \mathbf{v})_{2,1} + \alpha_2(\mathbf{u} \otimes \mathbf{v})_{1,2}$. Using the MNL algorithm, the relevance scores for modalities m_u and m_v are α_1 and α_2 respectively. However, the relevance score for the modality with tensor product $(m_{u \otimes v})$ is $\sqrt{(\alpha_1)^2 + (\alpha_2)^2}$. Here, the L^p-norm regularization with any $p < 2$ will favor the representation $m_{u \otimes v}$. Therefore, if the dimensions of the modality representations $\mathbf{u}$ and $\mathbf{v}$ are strictly greater than required to represent the ground truth (which is usually the case in modern networks), lower-order functions will be preferably represented inside the higher-order products by learning a constant in the representations of each modality. Our experiments show that a trained network typically exhibits such behavior. We propose a solution to this problem of higher-order interaction bias in the rest of this section.

Correct Bi-modal Interactions. We now address the problem of higher-order interaction bias. The key idea is to introduce a normalization scheme that downweights higher-order interactions. Our normalization scheme is a sophisticated generalization of the Vector-wise Batch Normalization (VBN) scheme described in equation 5. Let m_1 and m_2 be two modalities and define their representations as f^{m_1} and f^{m_2}. The representation of the bi-modal interaction is defined as $f^{\{m_1, m_2\}} = f^{m_1} \otimes f^{m_2}$. In this simple bi-modal case, our solution can be summarized as follows: we apply VBN to f^{m_1} and f^{m_2} before taking the product, and finally apply VBN again to the result. Formally, normalize each modality representation according to equation 5 as:

$$\tilde{f}^m(x_i) = \frac{f^m(x_i) - \mathbb{E}(f^m(x_i))}{\sqrt{\mathbb{E}\left(\|f^m(x_i) - \mathbb{E}(f^m(x_i))\|_2^2\right)}},$$

then compute the tensor-product as:

$$\hat{f}^{\{m_1, m_2\}}(x_i) = \tilde{f}^{m_1} \otimes \tilde{f}^{m_2},$$

and similarly apply VBN to the tensor-product to obtain $\tilde{f}^{\{m_1, m_2\}}$.

The centering step of normalization is applied element-wise over a mini-batch. Thus, if a few components were to be non-zero constants in a mini-batch, they would become zero after the normalization. This ensures that $\tilde{f}^{m_1, m_2}$ cannot easily access lower degree information contained in $\tilde{f}^{m_1}$, because elements in $\tilde{f}^{m_2}$ cannot be a non-zero constant, and vice versa. The normalization could seemingly be trivially extended to more than two modalities by applying the normalization iteratively up to an n-order tensor product. However, such an extension may still lead to a high-interaction bias. We illustrate why such a trivial extension may not work for more than two modalities and later generalize the normalization scheme for any number of modalities.

Generalization Over M-modal Interactions. Extending the aforementioned normalization to cases where $D > 2$ is not straightforward. This complexity arises because, when fusing more than two modalities, potentially, the representations of a subset of M modalities conspire to produce a constant even though each individual modality representation is non-constant. For instance, consider three modalities with one-dimensional representations and apply VBN to the representations f^1, f^2, f^3, then to $f^{1,2}$, and finally to $f^{1,2,3} = \tilde{f}^{1,2} \otimes \tilde{f}^3$, it is still possible that representation f^1 is learned in the higher-order tensor product. For instance, assume the components of f^2 and f^3 are learned to satisfy the following for each datapoint: (1) $f^2 = f^3$; (2) we have that f^2 is a Rademacher variable ($P[f^2 = 1] = P[f^2 = -1] = 0.5$); and (3) f^2 is independent of f^1. Since $f^{1,2,3} = f^1 f^2 f^3$ and $f^2 f^3 = 1$ for all datapoints, we actually have $f^{1,2,3} = f^1$,

and this seemingly higher order combination can still recover the first modality.

We address this issue by carefully normalizing the modalities features. The key is to prevent the combination of features of one or more modalities from resulting in a constant value. Similar to the bi-modal scenario, we need to ensure that the contribution of a subset of modalities to the larger fusion set is, on average, zero. This guarantees that no constant value, other than zero, is multiplied by the product of the complement of that subset within the original fusion set. Formally, for each $I \subset \{1, 2, \ldots, M\}$, the centering step of our batch norm procedure is defined as follows, where we first assume each modality is one-dimensional to simplify the exposition:

$$\hat{f}^I = \sum_{\ell=0}^{|I|} (-1)^\ell \sum_{\substack{\emptyset \neq S^1, \ldots, S^\ell \subset I \\ S^1, \ldots, S^\ell \text{disjoint}}} \prod_{m \in I \setminus (\cup S^k)} f^m \prod_{k \in \{1,2,\ldots,\ell\}} \mathbb{E}\left(\prod_{m \in S^k} f^m \right). \tag{9}$$

When the modalities are multi-dimensional, the above operation is applied independently to each multi-index component [1]. After preforming the centering step above for each multi-index component, we perform the generalized normalization step as follows

$$\tilde{f}^I = \frac{\hat{f}^I}{\sqrt{\mathbb{E}\|\hat{f}^I\|^2_{\text{Fr}}}}. \tag{10}$$

While the solution can no longer be easily interpreted as a composition of standard batch norm operations, it is, in fact, possible to show that lower-order fusion can not be represented by a linear combination of their higher-order counterparts. Theorem 2 formalizes this result.

Theorem 2. *The centering step described in equation 9 can be represented as the multi variate polynomial:*

$$\sum_{J \subset I} \mathcal{G}_J \prod_{m \in J} f^m,$$

for some real coefficients $\mathcal{G}_J$. Furthermore, the expected contribution of a subset of modalities J in the fusion of the set of modalities I, where $J \subsetneq I$ is zero. That is, we have for any $J \subsetneq I$ (including the empty set),

$$\mathbb{E}\left(\sum_{K : J \subset K \subset I} \mathcal{G}_K \prod_{m \in K \setminus J} f^m \right) = 0. \tag{11}$$

The theorem states that the expected value of the contribution of any subset $J \subsetneq I$ of modalities is zero in the fusion of I. Thus, $\tilde{f}^I$ (higher-order) can not learn a linear combination of the $\tilde{f}^J$ (lower-order). Appendix B contains a comprehensive proof of the theorem.

To make the exposition clearer, we provide as an example the case $I = \{1, 2, 3\}$ and the individual representations f^m are standardized using VBN. In this case, we have, using the notation $\overline{f^1 f^2} = \mathbb{E}(f^1 f^2)$:

$$\hat{f}^{1,2,3} = f^1 \times f^2 \times f^3 - \overline{f^1 \times f^2} \times f^3 - \overline{f^1 \times f^3} \times f^2 - \overline{f^2 \times f^3} \times f^1 - \overline{f^1 \times f^2 \times f^3}.$$

An elaborated centering step, without normalizing the individual modality representations, and strictly following equation 9 is described in Appendix B.1.

In this section, we introduced *Iterative Batch Normalization (IterBN)*, a normalization scheme addressing higher-order interaction bias in multimodal learning. In the next section, we show the effectiveness of our method on synthetic and real-world datasets.

4 Experiments

First, we experiment on synthetic data, where we control the amount of relevant information in the modalities, and compare InTense's relevance scores to the established ground truth. Second, we compare the predictive performance of InTense with popular multimodal fusion methods on six real-world multimodal datasets.

4.1 Evaluating the Relevance Scores

We created a multimodal dataset where each modality of a datapoint is a sequence of letters chosen randomly from a predefined set. For each datapoint x and modality m, an informative subsequence is inserted at a random position with a probability of p_m. We call our dataset SYNTHGENE. More details about it can be found in Appendix C.

We perform two experiments to determine the correctness of the relevance scores obtained from InTense. First, we construct a binary classification dataset with labels that ensure the modalities are independent and do not interact. Second, we generate another set of labels that can only be predicted using non-linear interactions among the modalities.

InTense Assigns Correct Relevance Scores to *Independent* Modalities

In this set of experiments, we create a synthetic dataset with independent modalities (i.e., without interactions). As a base-

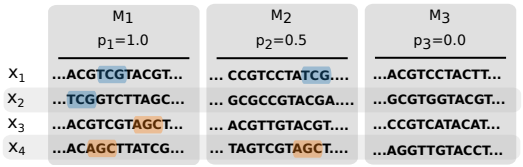

Figure 2: An excerpt of three modalities of SYNTHGENE, our self-curated binary classification dataset, where each sequence is made from a set of letters {A,C,G,T}. A positive class-sequence "TCG" and a negative class-sequence "AGC" is added according to the probability p_m.

[1] In particular, the multivariate case could be expressed with a similar formula as equation 9 with the products replaced by outer tensor products, but this would require a different reordering of the components for each term of the sum.

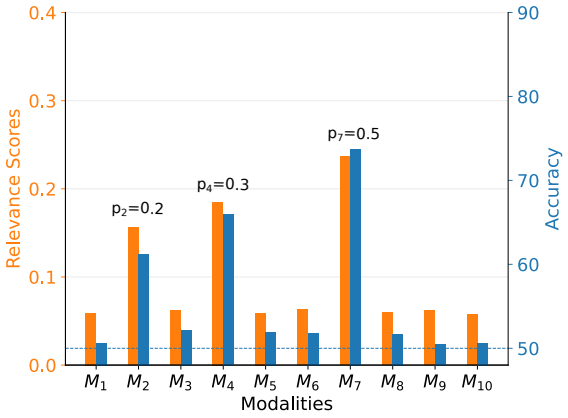

Figure 3: The figure shows a high correlation of InTense' relevance scores and accuracies of unimodal models on SynthGene. The modalities M_2, M_4, M_7 achieve high relevance scores and high accuracy as they contain class-specific information. Other modalities contain no class-specific information, which leads to a very low relevance score and an accuracy of around 50% (equivalent to random guessing).

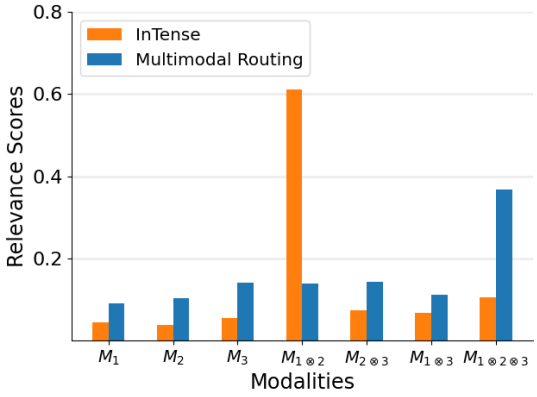

Figure 4: Illustration of the relevance scores calculated by the proposed InTense and the MultiRoute baseline when higher-order modality interactions are involved in the ground truth. MultiRoute leads to biased results (blue bars), where the relevance scores are concentrated toward higher-order interactions $M_{1\otimes2\otimes3}$. In contrast, InTense (orange bars) correctly assigns a high relevance score only to the interaction $M_{1\otimes2}$, which contains all class-specific signals.

line, we train one model for each modality and then compare the accuracies of those models with the relevance scores of InTense trained on all modalities together.

Dataset. We create the SYNTHGENE dataset, where for each modality m, each datapoint with a positive/negative class label contains a class-specific sequence with a probability p_m independent of other modalities. A high value of p_m indicates that most sequences in the modality m contain a class-specific sequence. Thus, the higher the value of p_m, the more relevant the modality m becomes. The labels for all datapoints are uniformly distributed between the two classes. Figure 2 shows how probability p_m affects modality relevance. We use 10 modalities where the informative subsequence is inserted into modalities M_2, M_4, and M_7. There is no discriminative information present in other modalities.

Results. Figure 3 shows the relevance scores calculated by InTense on SynthGene. InTense assigns the correct relevance scores as they align with human intuition. The higher the probability p_m, the more informative signal is contained in modality m, and the higher the predicted relevance score. We further validate the correctness of InTense's interpretability by comparing it with the accuracies obtained from unimodal models trained on each modality separately. Again, InTense's relevance scores correlate with the unimodal accuracies.

InTense Assigns Correct Relevance Scores to *Interacting* Modalities

We now turn to a situation where the label depends on a non-linear interaction among the modalities by design.

Dataset. We also create the SYNTHGENE-TRI dataset, a tri-modal version of SYNTHGENE. However, this time, the informative subsequence is not class-specific. Instead, the label is defined by an exclusive-or (XOR) relationship between the first two modalities (M_1 and M_2). The label is 0 if both modalities contain the subsequence or none of them does, and the label is 1 otherwise (i.e., when one of the modalities contains the subsequence). Note that modality M_3 does not contain any informative subsequence; thus, it is irrelevant. As before, we generate a balanced dataset with 50% of the samples being positive and 50% negative.

Results. Figure 4 shows the results. We observe that the global relevance scores calculated by the MultiRoute baseline [Tsai *et al.*, 2020] are biased toward high-order interactions. This occurs even when the interactions do not add any useful information and thus should have been discarded by MultiRoute. We further see that the proposed InTense method does avoid this bias and correctly assigns a high relevance score solely to the interaction of M_1 and M_2. This shows that InTense can ensure the correctness of relevance scores even when higher-order interactions are involved.

4.2 InTense Performs SOTA in Real-World Applications

We demonstrate the effectiveness of InTense in providing interpretability without compromising predictive performance across a range of real-world applications. In order to compare performance and ensure reproducibility, we followed the experimental setup (e.g., data preprocessing, encodings of different modalities) of the MultiBench [Liang *et al.*, 2021] benchmark for all the experiments.

Sentiment analysis. In sentiment analysis, also known as opinion mining, the target is to identify the emotional tone or feeling underlying the data. Initially confined to text data, sentiment analysis has evolved to encompass multiple modalities. The task becomes challenging due to the intricate interactions of modalities, which play a significant role in expressing sentiments. Understanding sentiments is crucial in business intelligence, customer feedback analysis, and social media monitoring. To evaluate InTense in sentiment analysis, we employed CMU-MOSEI [Bagher Zadeh *et al.*, 2018], the

	Baselines		Ours	
	MultiRoute	MRO	MNL	InTense
MUStARD	65.9	66.5	67.4	**69.6**
CMU-MOSI	76.8	75.8	**80.8**	79.7
UR-FUNNY	63.6	63.4	63.4	**65.1**
CMU-MOSEI	80.2	79.7	80.5	**81.5**
AV-MNIST	71.8	72.0	72.4	**72.8**
ENRICO	46.7	49.2	47.1	**50.8**

Table 1: Accuracies for different baselines on the test fold. Each experiment is carried out ten times to compute the statistics.

largest dataset of sentence-level sentiment analysis for real-world online videos, and CMU-MOSI [Zadeh *et al.*, 2016], a collection of annotated opinion video clips.

Humor and Sarcasm detection. Humor detectors identify elements that evoke amusement or comedy, while sarcasm detection aims to discern whether a sentence is presented in a sarcastic or sincere manner. Sarcasm and humor are often situational. Successfully detecting them requires a comprehensive understanding of various information sources, encompassing the utterance, contextual intricacies of the conversation, and background of the involved entities. As this information extends beyond textual cues, the challenge lies in learning the complex interactions among the available modalities. To assess our approach's effectiveness in these tasks, we utilized UR-FUNNY [Hasan *et al.*, 2019] for humor detection and MUStARD [Castro *et al.*, 2019] for sarcasm detection.

Layout Design Categorization. Layout design categorization is about classifying graphical user interfaces into predefined categories. Automizing this task can support designers in optimizing the arrangement of interactive elements, ensuring the creation of interfaces that are not only visually appealing but also functional and user-centric. Classifiers can, e.g., assign semantic captions to elements, enable smart tutorials, or be the foundation for advanced search engines. For this paper, we considered the ENRICO [Leiva *et al.*, 2020] dataset as an example for layout design categorization. ENRICO comprises 20 design categories and 1460 user interfaces with five modalities, including screenshots, wireframe images, semantic annotations, DOM-like tree structures, and app metadata.

Digit Recognition. We also include results for Audiovision-MNIST (AV-MNIST) [Vielzeuf *et al.*, 2018], a multimodal dataset comprising images of handwritten and recordings of spoken digits. Despite its apparent lack of immediate real-world application, the dataset's significance lies in its establishment as a standard multimodal benchmark. It allows us to situate our research within the broader context of previous research [Pérez-Rúa *et al.*, 2019; LeCun *et al.*, 1998].

Baselines. We compare the classification performance of *InTense* and MNL to the following state-of-the-art interpretable multimodal learning baselines: 1) Multimodal Residual Optimization (MRO) [Wörtwein *et al.*, 2022] and 2) Multimodal Routing (MultiRoute) [Tsai *et al.*, 2020]. Additionally, we consider three non-interpretable baselines: 1) *LF-Concat*, 2) *TF Network* [Zadeh *et al.*, 2017], and 3) multimodal transformer (*MulT*) [Tsai *et al.*, 2019]. These baselines have been identified as leading in the independent comparison conducted by Liang *et al.* [2021]. The multimodal transformer was particularly highlighted for consistently reaching some of the highest accuracy levels.

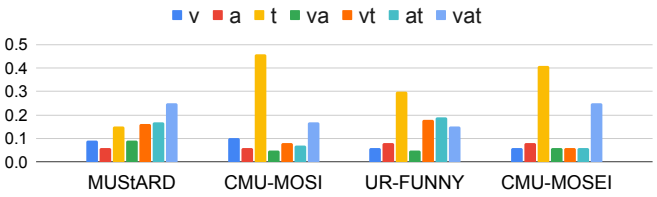

Figure 5: Relevance scores from InTense for audio (a), vision (v), text (t), and all their possible interactions.

Results. The results are shown in Table 1. We observe that our proposed models, MNL and InTense, surpass all interpretable baselines in terms of classification accuracy. InTense achieves the highest performance across all datasets except for CMU-MOSI, where MNL excels. CMU-MOSI is the smallest dataset in our analysis, a factor that may contribute positively to MNL's performance. Compared with non-interpretable multimodal learning methods (see Table 1 in Appendix E), MNL and especially InTense demonstrate impressive performance, almost meeting the accuracy of the non-interpretable Multimodal Transformer (MulT). The performance of our proposed models is within a narrow 2% error margin (and frequently much lower) compared to the MulT baseline.

Figure 5 shows the interpretable relevance scores that InTense assigns to the various modalities. Notably, in three of the four datasets analyzed, text emerges as the most significant modality. We identify two plausible explanations for this phenomenon. First, several studies have reported a strong correlation of text with sentiment [Gat *et al.*, 2021]. Second, the predominance of the text modality may be attributed to the availability of sophisticated word embeddings obtained from large pre-trained foundation models. However, we find an exception in the interpretability scores for the sarcasm detection dataset (MUStARD). Sarcasm detection requires information from multiple modalities, making sole reliance on one, especially text, insufficient for accuracy.

5 Conclusion

We introduced InTense, a novel interpretable approach to multimodal learning that offers reliable relevance scores for modalities and their interactions. InTense achieves state-of-the-art performance in several challenging applications, from sentiment analysis and humor detection to layout design categorization and multimedia. We proved theoretically and validated empirically that InTense correctly disentangles higher-order interactions from the individual effects of each modality. The full transparency of InTense makes it suitable for future application in safety-critical domains.

Ethical Statement

As an interpretable approach, the proposed methodology naturally aids in making multimodal learning more transparent. By attributing importance scores to different modalities and their interactions, InTense may reveal biases in decision-making and improve trustworthiness. For instance, consider a system tasked with classifying loan suitability. Our approach may expose social biases when relevance scores for certain modalities, such as gender extracted from vision, are disproportionately high. Moreover, unlike existing approaches, InTense has no higher-order interaction bias (see Section 3.3). That is, it does not incorrectly assign large relevance scores to higher-order interactions, which can create the false impression of a social bias. The full transparency of InTense prevents the deployment of a harmful classification model, contributing to the ethical use of AI in sensitive domains.

Acknowledgements

SV, PL, WM, and MK acknowledge support by the Carl-Zeiss Foundation, the DFG awards KL 2698/2-1, KL 2698/5-1, KL 2698/6-1 and KL 2698/7-1, and the BMBF awards 03|B0770E, and 01|S21010C.

References

[Arabacı *et al.*, 2021] Mehmet Ali Arabacı, Fatih Özkan, Elif Surer, Peter Jančovič, and Alptekin Temizel. Multi-modal egocentric activity recognition using multi-kernel learning. *Multimedia Tools and Applications*, 80(11):16299–16328, 2021.

[Bagher Zadeh *et al.*, 2018] 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, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[Behravan *et al.*, 2018] Hamid Behravan, Jaana M Hartikainen, Maria Tengström, Katri Pylkäs, Robert Winqvist, Veli-Matti Kosma, and Arto Mannermaa. Machine learning identifies interacting genetic variants contributing to breast cancer risk: A case study in finnish cases and controls. *Scientific reports*, 8(1):13149, 2018.

[Büchel *et al.*, 1998] Christian Büchel, Cathy Price, and Karl Friston. A multimodal language region in the ventral visual pathway. *Nature*, 394(6690):274–277, 1998.

[Cao *et al.*, 2020] Jize Cao, Zhe Gan, Yu Cheng, Licheng Yu, Yen-Chun Chen, and Jingjing Liu. Behind the scene: Revealing the secrets of pre-trained vision-and-language models. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 565–580. Springer, 2020.

[Castro *et al.*, 2019] Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. Towards multimodal sarcasm detection, 2019.

[Chagas *et al.*, 2020] Paulo Chagas, Luiz Souza, Ikaro Araújo, Nayze Aldeman, Angelo Duarte, Michele Angelo, Washington LC Dos-Santos, and Luciano Oliveira. Classification of glomerular hypercellularity using convolutional features and support vector machine. *Artificial intelligence in medicine*, 103:101808, 2020.

[Chandrasekaran *et al.*, 2018] Arjun Chandrasekaran, Viraj Prabhu, Deshraj Yadav, Prithvijit Chattopadhyay, and Devi Parikh. Do explanations make vqa models more predictable to a human? *arXiv preprint arXiv:1810.12366*, 2018.

[Chen *et al.*, 2014] JunKai Chen, Zenghai Chen, Zheru Chi, and Hong Fu. Emotion recognition in the wild with feature fusion and multiple kernel learning. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 508–513, 2014.

[Elgart *et al.*, 2022] Michael Elgart, Genevieve Lyons, Santiago Romero-Brufau, Nuzulul Kurniansyah, Jennifer A Brody, Xiuqing Guo, Henry J Lin, Laura Raffield, Yan Gao, Han Chen, et al. Non-linear machine learning models incorporating snps and prs improve polygenic prediction in diverse human populations. *Communications Biology*, 5(1):856, 2022.

[Frank *et al.*, 2021] Stella Frank, Emanuele Bugliarello, and Desmond Elliott. Vision-and-language or vision-for-language? on cross-modal influence in multimodal transformers. *arXiv preprint arXiv:2109.04448*, 2021.

[Gat *et al.*, 2021] Itai Gat, Idan Schwartz, and Alex Schwing. Perceptual score: What data modalities does your model perceive? *Advances in Neural Information Processing Systems*, 34:21630–21643, 2021.

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

[He *et al.*, 2020] Zhipeng He, Zina Li, Fuzhou Yang, Lei Wang, Jingcong Li, Chengju Zhou, and Jiahui Pan. Advances in multimodal emotion recognition based on brain–computer interfaces. *Brain sciences*, 10(10):687, 2020.

[Hessel and Lee, 2020] Jack Hessel and Lillian Lee. Does my multimodal model learn cross-modal interactions? it's harder to tell than you might think! In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 861–877, 2020.

[Ioffe and Szegedy, 2015] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr, 2015.

[Kanehira *et al.*, 2019] Atsushi Kanehira, Kentaro Takemoto, Sho Inayoshi, and Tatsuya Harada. Multimodal explanations by predicting counterfactuality in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8594–8602, 2019.

[Kloft *et al.*, 2011] Marius Kloft, Ulf Brefeld, Sören Sonnenburg, and Alexander Zien. Lp-norm multiple kernel learning. *The Journal of Machine Learning Research*, 12:953–997, 2011.

[LeCun *et al.*, 1998] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[Leiva *et al.*, 2020] 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.

[Liang *et al.*, 2021] 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.

[Lunghi *et al.*, 2019] Giacomo Lunghi, Raul Marin, Mario Di Castro, Alessandro Masi, and Pedro J Sanz. Multimodal human-robot interface for accessible remote robotic interventions in hazardous environments. *IEEE Access*, 7:127290–127319, 2019.

[Park *et al.*, 2018] Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. Multimodal explanations: Justifying decisions and pointing to the evidence. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8779–8788, 2018.

[Pérez-Rúa *et al.*, 2019] Juan-Manuel Pérez-Rúa, Valentin Vielzeuf, Stéphane Pateux, Moez Baccouche, and Frédéric Jurie. Mfas: Multimodal fusion architecture search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6966–6975, 2019.

[Poria *et al.*, 2015] Soujanya Poria, Erik Cambria, and Alexander Gelbukh. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2539–2544, 2015.

[Rakotomamonjy *et al.*, 2008] Alain Rakotomamonjy, Francis Bach, Stéphane Canu, and Yves Grandvalet. Simplemkl. *Journal of Machine Learning Research*, 9:2491–2521, 2008.

[Sabour *et al.*, 2017] Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. Dynamic routing between capsules. *Advances in neural information processing systems*, 30, 2017.

[Tan and Bansal, 2019] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.

[Tsai *et al.*, 2019] 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 conference. Association for Computational Linguistics. Meeting*, volume 2019, page 6558. NIH Public Access, 2019.

[Tsai *et al.*, 2020] Yao-Hung Hubert Tsai, Martin Q Ma, Muqiao Yang, Ruslan Salakhutdinov, and Louis-Philippe Morency. Multimodal routing: Improving local and global interpretability of multimodal language analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2020, page 1823. NIH Public Access, 2020.

[Vielzeuf *et al.*, 2018] Valentin Vielzeuf, Alexis Lechervy, Stéphane Pateux, and Frédéric Jurie. Centralnet: a multilayer approach for multimodal fusion. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018.

[Wörtwein *et al.*, 2022] Torsten Wörtwein, Lisa Sheeber, Nicholas Allen, Jeffrey Cohn, and Louis-Philippe Morency. Beyond additive fusion: Learning non-additive multimodal interactions. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4681–4696, 2022.

[Zadeh *et al.*, 2016] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Mosi: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos, 2016.

[Zadeh *et al.*, 2017] 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.

[Zhang *et al.*, 2023] Yiyuan Zhang, Kaixiong Gong, Kaipeng Zhang, Hongsheng Li, Yu Qiao, Wanli Ouyang, and Xiangyu Yue. Meta-transformer: A unified framework for multimodal learning. *arXiv preprint arXiv:2307.10802*, 2023.