

MMOE: Enhancing Multimodal Models with Mixtures of Multimodal Interaction Experts

Haofei Yu^{1*}, Zhengyang Qi^{1*}, Lawrence Jang^{1*},
Ruslan Salakhutdinov¹, Louis-Philippe Morency¹, Paul Pu Liang²

¹Carnegie Mellon University, ²Massachusetts Institute of Technology
{haofei, zqi, ljang}@cs.cmu.edu ppli@mit.edu

Abstract

Advances in multimodal models have greatly improved how interactions relevant to various tasks are modeled. Today’s multimodal models mainly focus on the correspondence between images and text, using this for tasks like image-text matching. However, this covers only a subset of real-world interactions. Novel interactions, such as sarcasm expressed through opposing spoken words and gestures or humor expressed through utterances and tone of voice, remain challenging. In this paper, we introduce an approach to enhance multimodal models, which we call **Multimodal Mixtures of Experts (MMOE)**. The key idea in MMOE is to train separate expert models for each type of multimodal interaction, such as redundancy present in both modalities, uniqueness in one modality, or synergy that emerges when both modalities are fused. On a sarcasm detection task (MUSARD) and a humor detection task (URFunny), we obtain new state-of-the-art results. MMOE is also able to be applied to various types of models to gain improvement.

1 Introduction

Recent advances in the design and pretraining of vision-language models have enabled significant progress in capturing the correspondences between images and text (Zhu et al., 2023; Li et al., 2023; Liu et al., 2023). These models have seen successes in image captioning (Xu et al., 2015), text-to-image generation (Saharia et al., 2022), multimodal retrieval (Mithun et al., 2018), multimodal classification (Li et al., 2021), and more. At its core, these methods aim to capture overlaps in semantic content between images and text, making a strong multi-view redundancy assumption (Tian et al., 2020; Liang et al., 2023b; Zbontar et al., 2021). However, redundancy is only one type of interaction seen between two modalities (Williams

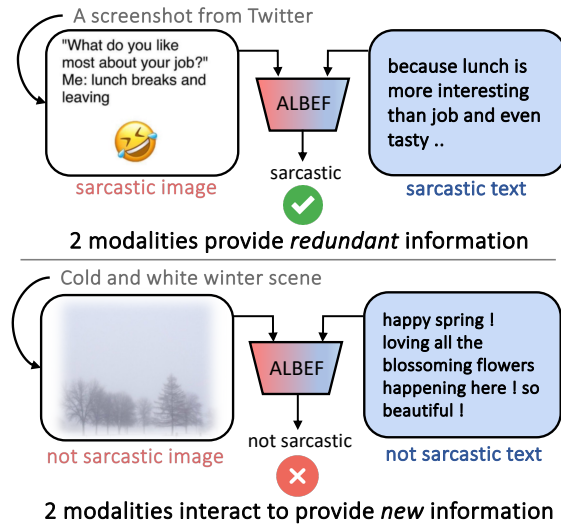


Figure 1: **A single model cannot handle all types of multimodal interactions well for hard multimodal prediction tasks.** For example, to predict sarcasm, ALBEF can have $\sim 89\%$ F1 when modalities contain redundant information (e.g., both the text and the image are sarcastic), but drops to $\sim 24\%$ F1 when there is synergy between modalities (e.g., the image shows a cold winter scene and the text says it is a happy spring, indicating the user’s sarcastic intent about the weather).

and Beer, 2010; Liang et al., 2023a; Marsh and Domas White, 2003). Instead, it might hinge on *unique* details from either modality (e.g. detecting laughter from someone not observed) or the result of *synergistic* fusion of both modalities, producing insights absent when either modality is considered in isolation (e.g., sarcasm and humor discerned from incongruent speech and gestures). Synergy is particularly interesting because it often arises when the predictions from different modalities are *contradicting*, or *incongruent* with one another (Bateman, 2014; Kruk et al., 2019; Zhang et al., 2018).

The diversity of possible real-world multimodal interactions poses a challenge to today’s multimodal models. Empirically, we find that *one single model may not be the most suitable in capturing all*

*Equal Contribution.

types of interaction at the same time. For example, models trained to learn the correspondences between words and image regions (e.g., for retrieval) will struggle when there is unique information in one modality (Liang et al., 2023b; Winterbottom et al., 2020), or when the image and text provide contradicting information that must be contextualized together (Hessel et al., 2022). We show an example of this failure in Figure 1, where ALBEF (Li et al., 2021) can easily detect sarcasm when it is present in both modalities (redundancy), but fails when the sarcastic intent arises from the synergistic fusion of both image and text. Quantitatively, ALBEF has a performance drop of up to 60% on data with synergistic interactions compared with those with redundancy interactions.

To tackle this problem, we propose MMOE, by leveraging the key insight that different interactions require different modeling paradigms. A natural way to model these differences is to use a mixture of multimodal experts with one specialized expert model for each interaction. Each expert model can be specialized based on the unique training data they see or a special training objective. Furthermore, there is evidence that the brain also uses separate expert regions during the multisensory integration process, depending on the types of input modalities and multimodal contexts present during perception (Stein et al., 2020). During inference on unseen data points at testing time, MMOE relies on specific fusion methods to provide weights for each expert model, combine the output of each expert model, and obtain a final prediction.

MMOE achieves new state-of-the-art results on one multimodal sarcasm detection dataset and one multimodal humor detection dataset we tested on, MUSTARD and URFunny. Moreover, we show that our approach is easy to implement on various types of models: fusion-based vision language models like ALBEF (Li et al., 2021), multimodal extended large language models like BLIP2 (Li et al., 2023), and image-captioned large language models like Qwen2 (Yang et al., 2024a) all improve after adding MMOE on top of them.¹

2 Related Work

We cover related work in quantifying and learning multimodal interactions, as well as recent advances in multimodal large language models, ensembling, and mixtures of experts.

Multimodal Interactions defines the degrees of commonality between modalities and the ways they combine to provide new information for a task (Liang et al., 2023d). A core problem lies in understanding the nature of how modalities interact and modeling these interactions using data-driven methods. The study of multimodal interactions has involved semantic definitions based on research in multimedia (Marsh and Domas White, 2003), human (and animal) communication (Partan and Marler, 2005; Flom and Bahrnick, 2007; Ruiz et al., 2006), and human social interactions (Mai et al., 2019; Jung et al., 2018). These have also inspired statistical methods to quantify multimodal interactions from unimodal predictions (Mazetto et al., 2021), trained model weights and activations (Sorokina et al., 2008; Tsang et al., 2018, 2020; Hessel and Lee, 2020), feature selection (Itner et al., 2021; Yu and Liu, 2003, 2004; Auffarth et al., 2010), and information theory (Liang et al., 2023a,c; Williams and Beer, 2010; Bertschinger et al., 2014). Our work builds on this line of work in quantifying multimodal interactions, particularly the statistical definitions that enable accurate estimation from large-scale multimodal datasets.

Multimodal Language Models have revolutionized multimodal learning since representations of images and text can now be fed into large language models for flexible question-answering, reasoning, and multi-turn dialog conditioned on images. Many of these models are built on top of multimodal extensions of the Transformer architecture (Su et al., 2020; Liang et al., 2022; Jaegle et al., 2021; Lu et al., 2019; Tsai et al., 2019; Tan and Bansal, 2019). In addition to training large-scale multimodal transformers *natively* from input modalities, another line of work takes pre-trained language and vision models and aims to learn a small set of *adapter* parameters to align visual and language representations (Koh et al., 2023; Li et al., 2023; Zhu et al., 2023). These approaches have shown strong performance on many multimodal benchmarks, such as in visual question answering (Wang et al., 2022), text-to-video generation (Kondratyuk et al., 2023), robotics tasks (Driess et al., 2023), and biomedical analysis (Acosta et al., 2022). However, these methods train monolithic models that perform the same computation for all types of multimodal interactions, which we show to be suboptimal and inefficient when datasets contain a mix of diverse and complex interactions.

¹Codebase and reproduction guidance are available at <https://github.com/lwaekfjlk/mmo>

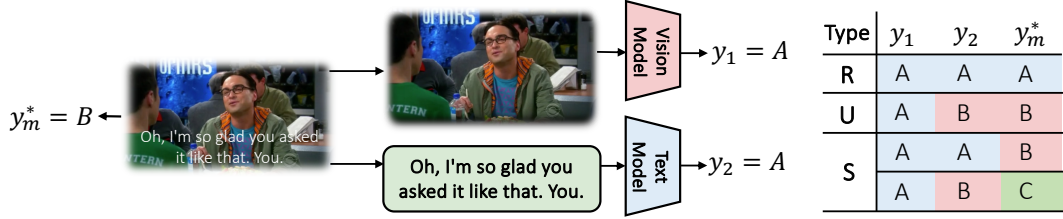


Figure 2: We classify one multimodal dataset into three subsets based on their multimodal interactions: (1) *Redundancy* (R), when both modalities provide the same prediction, (2) *Uniqueness* (U), when two modalities make different predictions, of which one of them is correct, (3) *Synergy* (S), when the ground-truth multimodal labels do not agree with either of unimodal predictions. y_1 represents the prediction based on vision modality, y_2 represents the prediction from text modality, and y_m^* represents the ground-truth label. $\{A, B, C\}$ represents classes.

Ensembles and Mixtures of Experts are commonly used techniques to boost a model’s performance using a collection of expert models each with their specialized expertise but individually weaker than baseline (Freund et al., 1996). Cheng et al. (2020) utilized a voting-based method to ensemble predictions from multiple models to provide more accurate answers. Besides discrete voting, continuous ensembles in logit space have also been proposed (Eigen et al., 2013; Tasci et al., 2021). In settings where it is difficult to define which expert is correct, trainable ensemble functions have been designed to automatically combine multiple experts in an end-to-end fashion (He et al., 2021; Shazeer et al., 2017; Du et al., 2022a). Our work uses these ideas as a foundation to learn different types of multimodal interactions.

3 Multimodal Mixtures of Experts

We focus on multimodal prediction tasks: given feature vectors from two modalities with x_1 and x_2 , our goal is to predict the label y using both x_1 and x_2 . Naturally, task-related information may be contained uniquely in one of the modalities, present redundantly in both, or require synergistically combining of information from both modalities. While prior work has focused on designing a single multimodal model for all data points in a task, our key insight is that each data point may exhibit a different type of interaction and therefore require a different modeling approach. Our method, which we call MMOE, is a natural solution to this problem in three steps (1) *Categorizing*: categorizing multimodal interaction types in each data point for the training set, (2) *Training*: training three expert models to master at each type of interactions (redundancy, uniqueness, and synergy), (3) *Inference*: dynamically ensembling the mixture of expert mod-

els during inference on unseen new data points. We now explain each of these three steps in detail.

3.1 Categorizing Multimodal Interactions

Prior work has provided definitions of *redundant*, *unique*, and *synergistic* interactions using the language of information theory (Williams and Beer, 2010; Liang et al., 2023a). However, estimating information theoretic measures can be challenging for high-dimensional and continuous distributions (Pérez-Cruz, 2008). When these interactions cannot be exactly computed, they can be approximately inferred by considering whether unimodal models trained on each modality *agree* or *disagree* with each other. We can formalize the concept of modality agreement and disagreement with a discrepancy function as follows:

Definition 1 (Prediction Discrepancy Function). Given feature $x_1 \in \mathcal{X}_1$ and $x_2 \in \mathcal{X}_2$, and unimodal classifiers $f_1 : \mathcal{X}_1 \rightarrow \mathcal{Y}$ and $f_2 : \mathcal{X}_2 \rightarrow \mathcal{Y}$, let $y_1 = f_1(x_1)$ and $y_2 = f_2(x_2)$ denote their predictions. We define the prediction discrepancy function $\delta(y_1, y_2)$ as a mapping $\delta : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}_{\geq 0}$ that quantifies the dissimilarity between the predictions of f_1 and f_2 . For tasks with a discrete label space \mathcal{Y} , the discrepancy function is defined as:

$$\delta(y_1, y_2) = \begin{cases} 0, & \text{if } y_1 = y_2, \\ 1, & \text{if } y_1 \neq y_2. \end{cases}$$

The binary discrepancy function indicates that modalities *agree* with each other when $\delta = 0$ and modalities *disagree* with each other when $\delta = 1$. Combining them with multimodal predictions, gives us an intuitive guideline to categorize data points based on three types of interactions:

1. *Redundancy*: when both modalities *agree* with the multimodal prediction, two modalities contain redundant information.

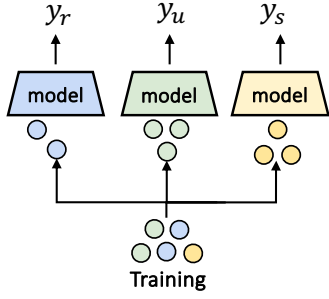


Figure 3: MMOE **training**: Each multimodal datapoint is categorized based on its multimodal interaction and used to train an expert model tailored only for that interaction.

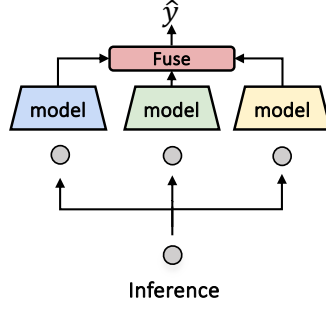


Figure 4: MMOE **inference**: We infer which multimodal interaction a test datapoint requires and use a soft weighted fusion over on the outputs from multiple expert models.

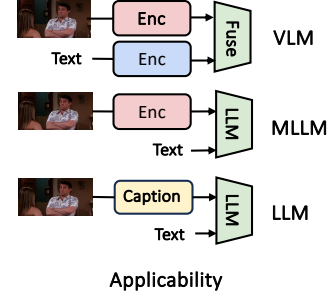


Figure 5: MMOE **applicability**: MMOE can be used as a drop-in method to the training of fusion-based VLMs, multimodal extended LLMs, and image-captioned LLMs.

2. *Uniqueness*: when two modalities *disagree* and one of them is aligned with the multimodal prediction, two modalities contain unique information and one is dominant.
3. *Synergy*: when the multimodal prediction *disagrees* with both unimodal predictions so there is synergistic information generated from two modalities when predicting.

With such intuitive guidelines above, we formally define the categorization process as follows:

Theorem 1 (Multimodal Interaction-based Categorization). *Let y_1 and y_2 denote the predictions from unimodal classifiers, and let y_m represent the multimodal prediction from multimodal models. The interaction discrepancy between the predictions is defined as:*

$$\Delta_{1,2}(y_1, y_2, y_m) = \delta(y_1, y_m) + \delta(y_2, y_m)$$

where $\delta(\cdot, \cdot)$ denotes the discrepancy function between two predictions. The categorization is then described as follows:

- $\Delta_{1,2} = 0$: *Redundancy*, i.e., $y_1 = y_2 = y_m$,
- $\Delta_{1,2} = 1$: *Uniqueness*, i.e., $y_1 = y_m \neq y_2$ or $y_2 = y_m \neq y_1$,
- $\Delta_{1,2} = 2$: *Synergy*, i.e., $y_1 \neq y_m$ and $y_2 \neq y_m$.

To illustrate the categorization rule, Figure 2 shows an example. In practice, obtaining high-quality predictions can be challenging. Labels from multimodal datasets, which are typically generated by humans making multimodal predictions, can be directly used. Also, we obtain high-quality unimodal predictions y_1 and y_2 via state-of-the-art foundation models in the few-shot prompting style

for all training data points. For vision-only predictions, we utilize vision-language models like CogVLM2 (Wang et al., 2023) to obtain them by providing only the query and the image and make sure that generated answers are conditioned only on the vision-side information. To get text-only predictions, we use state-of-the-art language models like Qwen2-72B-Instruct (Yang et al., 2024a) with the query and the language information so the model answers are conditioned only on text for prediction. More information related to the collection of unimodal labels is available in Appendix §F.

3.2 Training Expert Models for Each Multimodal Interaction Type

Given the categorization of multimodal datasets into subsets each with a similar interaction, this section describes how we use these interaction-specific subsets to train interaction-specific expert models. Illustrated in Figure 3, there are three specialized models, which we term f_r , f_u , and f_s for expert models of redundancy, uniqueness, and synergy respectively. While these individual expert models share the same format of inputs with image and text data pairs, their learning outcomes can differ significantly due to the multimodal data distributions they are trained on.

Overall, for expert model training, we collect all high-quality evidence of redundant interactions to train a redundancy expert model f_r . This process is repeated for unique and synergistic interactions, resulting in trained expert models f_r , f_u , and f_s . Each expert is trained only on the subset of data points that *maximally exhibit that interaction*; this specialization enables experts to be performant at learning that specific interaction. More technical details about the training process of expert models

are further discussed in Appendix §G.

We also note that it is possible to design interaction expert models using different modeling architectures and training objectives based on innovations in multimodal machine learning. For example, it has been empirically demonstrated that late fusion models are more suitable when modalities are redundant (Gadzicki et al., 2020), and models with expressive higher-order interactions (e.g., polynomials and tensors) are suitable when there is synergy between modalities (Hou et al., 2019). Moreover, multi-task training allows us to leverage the power of scale and learn interaction expert models adaptable to multiple tasks simultaneously. We leave these explorations for our future work.

3.3 Inference with Mixtures of Expert Models

The conclusion of Section §3.2 yields three expert models each suited for a certain type of multimodal interactions. During inference on unseen test data points, we need to select one or more expert models that are most suitable to get the final prediction. This is a challenge since the categorization of data points during training (presented in Section §3.1) relies on the multimodal prediction y_m , which we have during training but not during inference. Therefore, we need to design a method to provide an accurate estimation of the potential multimodal interactions included in one data point.

Our key assumption is that categorizing multimodal interactions within a data point is an essential sub-task that must be completed before the model can generate a final prediction. The multimodal interaction type captures the information shift between unimodal and multimodal inputs. These interaction-type predictions emphasize more general features compared to those needed for task label prediction. Consequently, even if a multimodal model struggles to accurately predict the task label y^* , it may still be able to determine the interaction type of the data point (e.g., whether two modalities provide similar, distinct, or synergistic information). This distinction becomes particularly relevant when the prediction task involves regression or classification with many classes.

Therefore, we *approximately categorize* data during inference through a *soft mixture of weights*, defined as w_r , w_u , and w_s over the three interaction types. These weights are inferred dynamically for each data point using a finetuned fusion model (e.g., BLIP2 in practice). We also test simple model-free baselines like prior constants

based on the frequency statistics of each interaction to weight each expert model and so on; see detailed ablation studies on these fusion methods in Section §6 and fusion model training details in Appendix §H. Using these inferred weights for each expert model, we obtain a final prediction $\hat{y} = \sum_{i \in \{r,u,s\}} w_i f_i(x_1, x_2)$ as the output of MMOE.

4 Experiments

Our experiments are designed to evaluate the effectiveness of our method when applied to a diverse set of multimodal foundation model architectures and multimodal prediction tasks.

4.1 Experimental Setup

We introduce the models and multimodal prediction tasks that we consider for experiments in this section. More information related to experimental settings is available in Appendix §I.

Model We implement MMOE on top of three categories of multimodal language models to show its widespread applicability on top of many base models (see Figure 5 for an illustration). Detailed model information is available in Appendix §A. These three model categories include:

1. *Fusion-based vision language models (VLM)* uses cross-attention to learn multimodal interactions between all regions of the image with all words in the input text. Examples of such models include ALBEF (Li et al., 2021), LXMERT (Tan and Bansal, 2019) and BLIP (Li et al., 2022).
2. *Multimodal-extended LLMs (MLLM)* includes models like BLIP2 (Li et al., 2023) and FROMAGE (Koh et al., 2023). It starts with an image encoder and an LLM as the backbone of the architecture. Most state-of-the-art models are based on multimodal-extended LLMs.
3. *Image-captioned LLMs (LLM)* convert images to text using an image captioning model and uses a text-only LLM like Qwen2 (Yang et al., 2024a) on the concatenation of captioned images and text inputs. Examples include the Socratic Model (Zeng et al., 2022) and the video understanding model (Zhang et al., 2023).

Multimodal prediction task We implement our method on three multimodal prediction tasks, including two sarcasm detection tasks, which are MUSTARD (Castro et al., 2019) and MMSD2.0 (Qin

et al., 2023), and one humor detection task, which is URFunny (Hasan et al., 2019). These tasks require interaction learning to conduct prediction. Detailed information about dataset statistic information and their preprocessing methods are available in Appendix §C and §D.

4.2 Main Results

In this section, firstly, we study how our best MMoE models compare to state-of-the-art baselines on multiple multimodal prediction tasks. Secondly, we study whether MMoE improves performance when applied on top of all three types of base models mentioned in Section §4.1.

Overall comparison with state-of-the-art In Table 1, we show that MMoE can improve the state-of-the-art performance on both the MUSTARD and URFunny datasets. Specifically, we outperform LFDNN-v1 (Ding et al., 2022) on the MUSTARD dataset, achieving a 1.35-point improvement in F1 score. On the URFunny dataset, our fine-tuned BLIP2 model with MMoE surpasses FDMER (Yang et al., 2022) with a 0.84-point gain in accuracy.

Improvement on various types of models We first compare the performance of 3 types of models with and without MMoE on MUSTARD dataset. As shown in Table 1, all models, including ALBEF, BLIP2, and Qwen2, show improvements in F1 scores. Notably, Qwen2-1.5B achieves an increase of 6.96 points, establishing it as the state-of-the-art model on this task. Additionally, on the MMSD2.0 dataset, both ALBEF and Qwen2 demonstrate performance gains, while BLIP2 remains relatively unchanged. For the URFunny dataset, ALBEF improves accuracy by 1.14 points, and BLIP2 by 0.84 points, whereas Qwen2 experiences a slight decline after applying MMoE. The performance drop on URFunny may be due to the inability of image captioning models to provide useful descriptions relevant to humor detection from the TED talk videos. As a result, text-based models like Qwen2 might struggle to achieve further improvements.

Furthermore, when comparing the performance across the three prediction tasks and three models, we observe a general trend: incorporating the MMoE tends to provide more robust improvements on challenging datasets (e.g., MUSTARD) and weaker models (e.g., ALBEF) with low F1 scores, which initially have lower performance. In contrast, the improvements are less pronounced on easier datasets (e.g., MMSD2.0) or stronger models (e.g., BLIP2), which already exhibit strong performance.

	Model	Acc (↑)	F1 (↑)
MUSTARD	MuT† (Tsai et al., 2019)	-	64.49
	LMF† (Liu et al., 2018)	-	69.92
	LFDNNv1† (Ding et al., 2022)	-	70.99
	ALBEF	54.49 \pm 3.13	48.51 \pm 2.21
	ALBEF+MMoE	<u>54.49</u> \pm 2.85	<u>51.95</u> \pm 2.81
	BLIP2	53.75 \pm 9.33	62.65 \pm 2.67
	BLIP2+MMoE	<u>59.18</u> \pm 2.11	<u>64.74</u> \pm 2.49
	Qwen2-0.5B	54.59 \pm 4.35	58.17 \pm 0.86
	Qwen2-0.5B+MMoE	49.06 \pm 3.00	<u>59.77</u> \pm 0.35
	Qwen2-1.5B	64.79 \pm 4.11	65.38 \pm 5.16
Qwen2-1.5B+MMoE	<u>70.69</u> \pm 3.28	<u>72.34</u> \pm 1.50	
URFunny	MuT† (Tsai et al., 2019)	66.65	-
	FDMER (Yang et al., 2022)	70.43	-
	ALBEF	66.77 \pm 0.86	68.67 \pm 0.18
	ALBEF+MMoE	<u>67.91</u> \pm 0.31	<u>69.85</u> \pm 0.32
	BLIP2	70.43 \pm 0.99	74.31 \pm 0.04
	BLIP2+MMoE	<u>71.27</u> \pm 0.87	<u>74.32</u> \pm 0.05
	Qwen2-0.5B	69.29 \pm 0.81	70.46 \pm 0.14
	Qwen2-0.5B+MMoE	69.19 \pm 0.64	68.38 \pm 1.55
	DynRT-Net (Tian et al., 2023)	71.40	71.34
	MCLIP (Qin et al., 2023)	85.64	84.10
LLaVA-1.5 (Liu et al., 2024)	85.18	85.11	
MMSD2.0	ALBEF	81.79 \pm 0.24	79.33 \pm 0.79
	ALBEF+MMoE	<u>82.30</u> \pm 0.27	<u>80.63</u> \pm 0.52
	BLIP2	84.75 \pm 0.20	83.52 \pm 0.35
	BLIP2+MMoE	<u>84.82</u> \pm 0.30	83.38 \pm 0.36
	Qwen2-0.5B	81.87 \pm 0.54	80.17 \pm 0.14
	Qwen2-0.5B+MMoE	82.27 \pm 0.14	<u>80.67</u> \pm 0.20

Table 1: MMoE can beat state-of-the-art models for MUSTARD and URFunny. It can be applied to any type of model for improvement. The numbers in the table represent the mean values from 3 runs with 3 seeds, with the corresponding standard variance provided. Full results can be found in Appendix §B. † indicates that models utilize all audio, text, and vision information provided in the dataset while ours only utilizes text and vision information for prediction.

5 Analysis

Based on these quantitative results, we further provide a fine-grained analysis of our method. First, we examine the limitations of current multimodal models by presenting empirical evidence where a single model faces challenges in typical types of interactions. We then explore whether specialized multimodal interaction expert models excel in their respective interaction types. Furthermore, we analyze the scaling law of expert models and discuss whether these expert models can be potentially smaller, in contrast to typically overparameterized models. Lastly, we provide additional details on the unimodal predictions and emphasize their important role in the data categorization process.

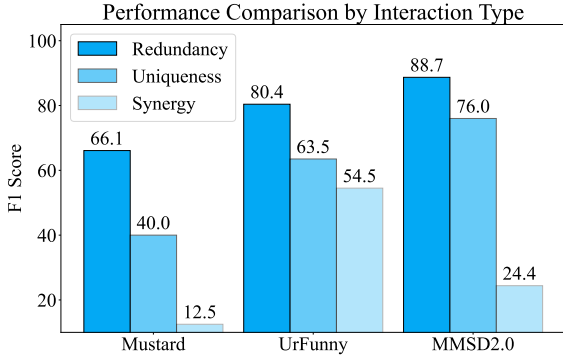


Figure 6: **Multimodal models struggle with synergy much more than redundancy and uniqueness.** ALBEF shows significantly lower performance on synergistic datapoints compared with redundancy and uniqueness that are categorized based on our method.

RQ1. What types of multimodal interaction do current models struggle with?

In Figure 6, we categorize all test data points based on their corresponding interaction type using the method mentioned in Section §3.1. We observe significant performance variations when using the same model to predict across data with different interaction types. Across the three datasets—MUSTARD, URFunny, and MMSD2.0—data points with synergy interactions show markedly lower F1 scores compared to those with uniqueness interactions, with performance gaps of 27.5, 9.0, and 51.6 for MUSTARD, URFunny, and MMSD2.0, respectively. Also, data points with uniqueness interaction perform substantially worse than those with redundancy interaction, with gaps of 26.1, 16.9, and 12.7 for three datasets. These trends are not limited to ALBEF, as we observe similar patterns in BLIP2 and Qwen2, highlighting that data points with strong synergy interactions represent a common challenge across all three types of models.

To better understand why models struggle with synergy-type interactions, we provide a case study in Figure 7 that highlights such failure. In this example, both the visual input (*people watching a show and clapping*) and the language input (*they think they should not leave*) lack clear signals of sarcasm individually. However, when combined, the synergized information (*where "them" refers to a band or show and "now" refers to the beginning or ending point of that*) reveals an evident sarcastic intent that is not present in the original visual or language cues. Despite large-scale pretraining, multimodal models struggle to capture such complex interactions between modalities accurately.

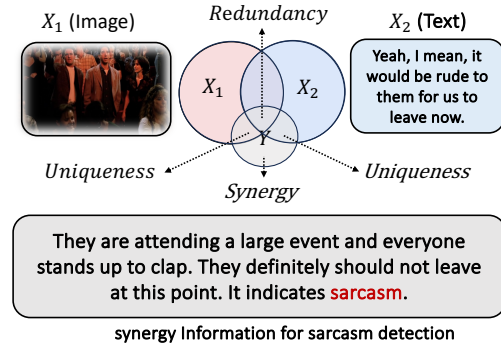


Figure 7: **Case study on synergy interaction.** Existing multimodal models struggle to learn the situation when both text and image modalities alone do not indicate sarcasm, but sarcasm arises due to the synergistic information between modalities when fused together.

RQ2. How do expert models perform on corresponding multimodal interaction data?

While a single large multimodal model may struggle, MMOE leverages specialized expert models to handle each type of interaction. As shown in Table 2, these expert models for redundancy, uniqueness, and synergy outperform test data points with their corresponding interaction types. Notably, expert models for synergy and redundancy show the most significant improvements in MMSD2.0: Qwen2-0.5B gains over 30 F1 points on synergy, and ALBEF improves by around 8 F1 points on redundancy. In contrast, expert models for uniqueness exhibit almost no change across different model settings. This could be because data points with unique interactions are more prevalent in the dataset compared to those with redundancy or synergy (data points with uniqueness account for around 61%). As a result, baseline models tend to focus on learning these features during training, leading to similar performance with expert models.

RQ3. How small can expert models be?

It is well established that neural networks, given enough parameters, are universal function approximators. Therefore, sufficiently large multimodal models should eventually be capable of learning all interaction types. However, we hypothesize that expert models can be smaller and benefit more from MMOE. To explore the scaling law of MMOE, we conducted an empirical study using Qwen2 models of different sizes (0.5B, 1.5B, and 7B). We observed a linear relationship between model size and performance score when plotted on a log-scale x-axis, as shown in Figure 8.

Model	Training	R	U	S
ALBEF	w/o expert train	88.70	76.02	24.39
	w/ expert train	<u>96.66</u>	<u>76.33</u>	<u>28.95</u>
BLIP2	w/o expert train	96.89	<u>80.16</u>	20.56
	w/ expert train	<u>99.10</u>	<u>80.16</u>	<u>48.98</u>
Qwen2-0.5B	w/o expert train	93.71	76.14	21.43
	w/ expert train	<u>96.54</u>	<u>76.16</u>	<u>53.66</u>

Table 2: **Performance of expert models on MMSD2.0.** Expert training based on the corresponding interaction type improves the model’s ability to predict test data points of the same type.

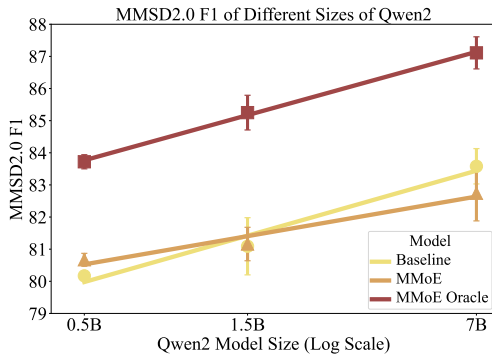


Figure 8: **MMoE gains better improvement on smaller models.** MMoE Oracle means that the model fusion process is based on categorized test datapoints with state-of-the-art unimodal models.

When applying MMoE to the 7B model, its performance worsens compared to the single-model baseline. However, as the model size decreases, the benefits of MMoE become increasingly significant. This scaling law suggests that MMoE is more effective with smaller expert models with worse single model performance, which makes sense since smaller models typically struggle to handle multiple interaction types, and specialized expert models can address this limitation more effectively by training on data with specific types of interactions.

Additionally, we also include the *oracle* performance of MMoE when using an oracle router for classifying interaction types in Figure 8. With such a router, each data point is always directed to the appropriate expert model for inference. In this setting, the mixture of experts achieves significantly higher performance compared to baseline models and shows a steeper slope when scaling to larger models. This finding suggests that the primary bottleneck of MMoE lies in training an accurate router to route data to the correct expert model for each interaction type. Moreover, it highlights that a model’s imbalanced ability to handle different types of multimodal interactions persists regardless

of its size or baseline performance.

RQ4. Is the improvement of MMoE primarily driven by model ensembling?

We investigate whether the performance gains in MMoE are primarily driven by our proposed multimodal interaction-driven data categorization (into redundancy, uniqueness, and synergy) instead of simple multiple model ensembling. To test this hypothesis, we conduct an ablation study using the URFunny dataset. In this experiment, we kept the number of training data points for each expert model unchanged but replaced the corresponding data points with the ones randomly sampled from the dataset. To eliminate any potential influence introduced by the different fusion methods during inference, we calculate the cross-entropy loss from the three expert models with the smallest values and averaged the score on the whole dataset to assess the upper-bound performance of the mixtures of experts. The metric is defined as:

$$CE_{moe} = \frac{1}{N} \sum_{i=1}^N \min_{y \in \{y_r, y_u, y_s\}} CE(y, y^*) \quad (1)$$

where N represents the total number of the dataset, y_r , y_u , and y_s represents the logits from expert models and y^* represents the ground-truth labels. We show that for our multimodal interaction-based categorization, the cross-entropy loss is 0.5853 while for random sampling categorization, the cross-entropy loss is 0.6942 (18.59% increase compared with our proposed categorization). Additionally, the original single model baseline has a loss of 0.8070. It indicates that our methods help build better models for the whole dataset.

RQ5. What do unimodal predictions look like?

The quality of unimodal partial labels is crucial for accurate data categorization, as these labels directly influence the categorization process. As discussed in Section §3.1, we utilize state-of-the-art models to generate unimodal predictions for the training set. Table 2 demonstrates that across all datasets—including MUSTARD, URFunny, and MMSD2.0—there is a clear bias toward the text modality. Text-based predictions are 16 points more accurate than those based on visual information. Moreover, predictions from the visual modality exhibit significantly lower confidence compared to those from the text-based modality, indicating that the visual side offers few reliable features for model predictions.

Dataset	Text Modality			Vision Modality		
	Acc	F1	Conf	Acc	F1	Conf
MUSTARD	<u>66.96</u>	<u>65.45</u>	0.93	50.87	64.72	0.57
URFunny	<u>67.87</u>	61.20	0.97	50.39	<u>62.27</u>	0.48
MMSD2.0	<u>66.88</u>	55.74	0.97	49.58	<u>60.39</u>	0.63

Table 3: **Quality and confidence of unimodal prediction.** *Conf* refers to the confidence of a prediction, calculated as the average of the maximum logits for the tokens "Yes" and "No" from the model’s final output logits over the entire vocabulary.

Fusion Method	MUSTARD	URFunny	MMSD2.0
Baseline	47.90	68.87	78.87
Average fusion	47.16	69.17	80.34
Maximum fusion	47.84	69.55	80.70
Weighted fusion	48.86	69.39	80.25
Model-based fusion	<u>48.97</u>	<u>70.20</u>	<u>80.71</u>
Oracle fusion	56.89	73.36	82.73

Table 4: **Ablation study on various fusion methods on ALBEF.** *Baseline* indicates the single model performance of ALBEF without fusion. *Oracle* refers to fusion performed on the test set that has been categorized using the same method applied to the training data.

6 Ablation Study

We conduct ablation studies on technical details in stages of categorizing and inference.

6.1 Ablation study on data categorization

We find that having *high-quality* categorized data is crucial for effective expert model training. Often, unimodal information alone doesn’t provide enough useful input for predictions, leading expert models to train on noisy data. This issue is particularly pronounced with vision-based predictions, as discussed in Section §5. To ensure expert models are trained on data that reflects unique interaction type, we filter out any data points where $|p(\text{Yes}) - p(\text{No})| < \delta$, with δ being a threshold indicating the confidence of the prediction. In experiments with BLIP2 on URFunny, when $\delta = 0$, meaning all training data is used, we achieve a model-based fusion result of 73.64 F1 score. With $\delta = 0.1$, partial data points are included in the training, and the F1 score improves to 74.65. However, when we increase δ to 0.15, the F1 score drops to 73.99, likely due to the reduction in training data. Therefore, we keep $\delta = 0.1$ for expert data filtering in our main experiments.

Another technique for expert model training is to rebalance the unimodal predictions of the data to

prevent highly imbalanced label distributions after data categorization. Rebalancing helps avoid training collapse in expert models, especially synergy expert models where the training data is few. Further details on data filtering and label rebalancing can be found in Appendix F.3.

6.2 Ablation study on expert model fusion

We also explore how different fusion strategies for combining multiple expert models impact performance. As mentioned in Section §5, fusion methods play a significant role during inference, suggesting that each expert model focuses on different aspects of multimodal information, and mixing them up simply cannot take full use of their prediction ability. The common fusion methods we consider include: (1) *Average Fusion*: where we simply average the softmaxed logits from the expert models to produce the final result. (2) *Maximum Fusion*: where we select the highest logits from all the expert models as the final prediction. (3) *Weighted Fusion*: for each dataset, we assign a fixed weight to each expert model, with the weights reflecting the proportion of each interaction type within the whole dataset. (4) *Model-based Fusion*: where we use a BLIP2-based classifier trained to distinguish between redundancy, uniqueness, and synergy. This classifier dynamically adjusts the weights for each expert model for each data point accordingly. Based on Table 4, we find that model-based fusion generally provides the most significant improvement compared with other model-free methods and single-model baseline. However, even a simple model-free fusion can bring improvement on URFunny and MMSD2.0 datasets, indicating the robustness of our methods.

7 Conclusion

This paper proposes a method to enhance multimodal models with a new **Multimodal Mixtures of Experts** structure (MMOE). The key idea is to train separate expert models each tailored to learn a specific type of multimodal interaction (including redundancy, uniqueness, and synergy), which overcomes significant shortcomings of existing models when diverse types of interactions are simultaneously present. Categorizing data points into their interactions enables the fusion of expert models during inference, which provides improvement to performance. MMOE also presents improved transparency of the multimodal modeling process.

Limitations

While we present a first step towards classifying and learning multimodal interactions, our categorization is still at a rather coarse level with only three interactions. Future work should investigate sub-categorizations of interactions, such as different types of synergy between modalities. This can be used to learn mixtures of interactions at a more fine-grained feature level. Furthermore, even approximate classification of interactions can lead to improved performance, so we expect future improvements in quantifying interactions to further improve MMOE. Future work can also investigate how to better combine multiple interactions in a compositional, multi-step manner to learn more complex higher-order interactions between modalities. Finally, we only consider modalities that have high-quality unimodal encoders like language and vision, future work can extend this direction to novel modalities such as sensors and medical data where unimodal models might have to be learned end-to-end with the multimodal interactions.

Ethics Statement

Multimodal AI systems can revolutionize many areas involving sensing and prediction such as in multimedia, healthcare, affective computing, and education, but there are also potential negative impacts involving monitoring and tracking humans and their states. For example, emotion detection models can be used inappropriately and invade personal privacy. Careful deployment to mitigate potential risks would be important.

Acknowledgement

This material is based upon work partially supported by National Science Foundation awards 1722822 and 1750439, National Institutes of Health awards R01MH125740, R01MH132225, R01MH096951 and R21MH130767, and Meta. PPL is supported in part by a Siebel Scholarship and a Waibel Presidential Fellowship. RS is supported in part by ONR grant N000142312368 and DARPA FA87502321015. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the sponsors, and no official endorsement should be inferred. We thank A100 and H100 GPU support from NetMind.AI² and NVIDIA.

²<https://netmind.ai/home>

References

- Julián N Acosta, Guido J Falcone, Pranav Rajpurkar, and Eric J Topol. 2022. Multimodal biomedical ai. *Nature Medicine*, 28(9):1773–1784.
- Benjamin Auffarth, Maite López, and Jesús Cerquides. 2010. Comparison of redundancy and relevance measures for feature selection in tissue classification of ct images. In *Industrial conference on data mining*, pages 248–262. Springer.
- John Bateman. 2014. *Text and image: A critical introduction to the visual/verbal divide*. Routledge.
- Nils Bertschinger, Johannes Rauh, Eckehard Olbrich, Jürgen Jost, and Nihat Ay. 2014. Quantifying unique information. *Entropy*.
- Yitao Cai, Huiyu Cai, and Xiaojun Wan. 2019. **Multimodal sarcasm detection in Twitter with hierarchical fusion model**. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2506–2515, Florence, Italy. Association for Computational Linguistics.
- Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. 2019. **Towards multimodal sarcasm detection (an _Obviously_ perfect paper)**. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4619–4629, Florence, Italy. Association for Computational Linguistics.
- Minhao Cheng, Cho-Jui Hsieh, Inderjit Dhillon, et al. 2020. **Voting based ensemble improves robustness of defensive models**. *ArXiv preprint*, abs/2011.14031.
- Ning Ding, Sheng-wei Tian, and Long Yu. 2022. A multimodal fusion method for sarcasm detection based on late fusion. *Multimedia Tools and Applications*, 81(6):8597–8616.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. **An image is worth 16x16 words: Transformers for image recognition at scale**. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.
- Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. 2023. **Palm-e: An embodied multimodal language model**. *ArXiv preprint*, abs/2303.03378.
- Nan Du, Yanping Huang, Andrew M. Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, Barret Zoph, Liam Fedus, Maarten P. Bosma, Zongwei Zhou, Tao Wang, Yu Emma Wang, Kellie Webster,

- Marie Pellat, Kevin Robinson, Kathleen S. Meier-Hellstern, Toju Duke, Lucas Dixon, Kun Zhang, Quoc V. Le, Yonghui Wu, Zhifeng Chen, and Claire Cui. 2022a. [Glam: Efficient scaling of language models with mixture-of-experts](#). In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 5547–5569. PMLR.
- Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022b. [Glm: General language model pretraining with autoregressive blank infilling](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.
- David Eigen, Marc’Aurelio Ranzato, and Ilya Sutskever. 2013. Learning factored representations in a deep mixture of experts. *arXiv preprint arXiv:1312.4314*.
- Ross Flom and Lorraine E Bahrick. 2007. The development of infant discrimination of affect in multimodal and unimodal stimulation: The role of intersensory redundancy. *Developmental psychology*, 43(1):238.
- Yoav Freund, Robert E Schapire, et al. 1996. Experiments with a new boosting algorithm. In *icml*, volume 96, pages 148–156. Citeseer.
- Konrad Gadzicki, Razieh Khamsehashari, and Christoph Zetzsche. 2020. Early vs late fusion in multimodal convolutional neural networks. In *2020 IEEE 23rd international conference on information fusion (FUSION)*, pages 1–6. IEEE.
- Md Kamrul Hasan, Wasifur Rahman, AmirAli Bagher Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, and Mohammed (Ehsan) Hoque. 2019. [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, Hong Kong, China. Association for Computational Linguistics.
- Jiaao He, Jiezhong Qiu, Aohan Zeng, Zhilin Yang, Jidong Zhai, and Jie Tang. 2021. [Fastmoe: A fast mixture-of-expert training system](#). *ArXiv preprint*, abs/2103.13262.
- Jack Hessel and Lillian Lee. 2020. [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, Online. Association for Computational Linguistics.
- Jack Hessel, Ana Marasović, Jena D Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. 2022. [Do androids laugh at electric sheep? humor" understanding" benchmarks from the new yorker caption contest](#). *ArXiv preprint*, abs/2209.06293.
- Ming Hou, Jiajia Tang, Jianhai Zhang, Wanzeng Kong, and Qibin Zhao. 2019. [Deep multimodal multilinear fusion with high-order polynomial pooling](#). In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 12113–12122.
- Jan Ittner, Lukasz Bolikowski, Konstantin Hemker, and Ricardo Kennedy. 2021. [Feature synergy, redundancy, and independence in global model explanations using shap vector decomposition](#). *ArXiv preprint*, abs/2107.12436.
- Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and João Carreira. 2021. [Perceiver: General perception with iterative attention](#). In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 4651–4664. PMLR.
- Tzzy-Ping Jung, Terrence J Sejnowski, et al. 2018. Multi-modal approach for affective computing. In *2018 40th annual international conference of the ieee engineering in medicine and biology society (embc)*, pages 291–294. IEEE.
- Jing Yu Koh, Ruslan Salakhutdinov, and Daniel Fried. 2023. [Grounding language models to images for multimodal inputs and outputs](#). In *International Conference on Machine Learning*, pages 17283–17300. PMLR.
- Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Rachel Hornung, Hartwig Adam, Hassan Akbari, Yair Alon, Vighnesh Birodkar, et al. 2023. [Videopoet: A large language model for zero-shot video generation](#). *ArXiv preprint*, abs/2312.14125.
- Julia Kruk, Jonah Lubin, Karan Sikka, Xiao Lin, Dan Jurafsky, and Ajay Divakaran. 2019. [Integrating text and image: Determining multimodal document intent in Instagram posts](#). 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 4622–4632, Hong Kong, China. Association for Computational Linguistics.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. [Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models](#). *ArXiv preprint*, abs/2301.12597.
- Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. 2022. [BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation](#). In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. PMLR.

- Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Gotmare, Shafiq R. Joty, Caiming Xiong, and Steven Chu-Hong Hoi. 2021. [Align before fuse: Vision and language representation learning with momentum distillation](#). In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 9694–9705.
- Paul Pu Liang, Yun Cheng, Xiang Fan, Chun Kai Ling, Suzanne Nie, Richard J Chen, Zihao Deng, Nicholas Allen, Randy Auerbach, Faisal Mahmood, et al. 2023a. Quantifying & modeling multimodal interactions: An information decomposition framework. In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Paul Pu Liang, Zihao Deng, Martin Q Ma, James Zou, Louis-Philippe Morency, and Russ Salakhutdinov. 2023b. Factorized contrastive learning: Going beyond multi-view redundancy. In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Paul Pu Liang, Chun Kai Ling, Yun Cheng, Alex Obolenskiy, Yudong Liu, Rohan Pandey, Alex Wilf, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2023c. [Multimodal learning without labeled multimodal data: Guarantees and applications](#). *ArXiv preprint*, abs/2306.04539.
- Paul Pu Liang, Yiwei Lyu, Xiang Fan, Jeffrey Tsaw, Yudong Liu, Shentong Mo, Dani Yogatama, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2022. [High-modality multimodal transformer: Quantifying modality & interaction heterogeneity for high-modality representation learning](#). *ArXiv preprint*, abs/2203.01311.
- Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2023d. Foundations & trends in multimodal machine learning: Principles, challenges, and open questions. *ACM Computing Surveys*.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. [Visual instruction tuning](#). *ArXiv preprint*, abs/2304.08485.
- Hui Liu, Wenya Wang, and Haoliang Li. 2022. [Towards multi-modal sarcasm detection via hierarchical congruity modeling with knowledge enhancement](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4995–5006, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe Morency. 2018. [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, Melbourne, Australia. Association for Computational Linguistics.
- Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. [Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks](#). In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13–23.
- Sijie Mai, Haifeng Hu, and Songlong Xing. 2019. [Di-vide, conquer and combine: Hierarchical feature fusion network with local and global perspectives for multimodal affective computing](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 481–492, Florence, Italy. Association for Computational Linguistics.
- Emily E Marsh and Marilyn Domas White. 2003. A taxonomy of relationships between images and text. *Journal of documentation*.
- Alessio Mazzetto, Dylan Sam, Andrew Park, Eli Upfal, and Stephen H. Bach. 2021. [Semi-supervised aggregation of dependent weak supervision sources with performance guarantees](#). In *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, April 13-15, 2021, Virtual Event*, volume 130 of *Proceedings of Machine Learning Research*, pages 3196–3204. PMLR.
- Niluthpol Chowdhury Mithun, Juncheng Li, Florian Metze, and Amit K Roy-Chowdhury. 2018. Learning joint embedding with multimodal cues for cross-modal video-text retrieval. In *Proceedings of the 2018 ACM on international conference on multimedia retrieval*, pages 19–27.
- Sarah R Partan and Peter Marler. 2005. Issues in the classification of multimodal communication signals. *The American Naturalist*, 166(2):231–245.
- Fernando Pérez-Cruz. 2008. [Estimation of information theoretic measures for continuous random variables](#). In *Advances in Neural Information Processing Systems 21, Proceedings of the Twenty-Second Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 8-11, 2008*, pages 1257–1264. Curran Associates, Inc.
- Libo Qin, Shijue Huang, Qiguang Chen, Chenran Cai, Yudi Zhang, Bin Liang, Wanxiang Che, and Ruifeng Xu. 2023. [MMSD2.0: Towards a reliable multimodal sarcasm detection system](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10834–10845, Toronto, Canada. Association for Computational Linguistics.
- Natalie Ruiz, Ronnie Taib, and Fang Chen. 2006. Examining the redundancy of multimodal input. In

- Proceedings of the 18th Australia conference on Computer-Human Interaction: Design: Activities, Artefacts and Environments*, pages 389–392.
- Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. 2022. Photo-realistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494.
- Sefik Serengil and Alper Ozpinar. 2024. [A benchmark of facial recognition pipelines and co-usability performances of modules](#). *Journal of Information Technologies*, 17(2):95–107.
- Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, and Jeff Dean. 2017. [Outrageously large neural networks: The sparsely-gated mixture-of-experts layer](#). In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.
- Daria Sorokina, Rich Caruana, Mirek Riedewald, and Daniel Fink. 2008. [Detecting statistical interactions with additive groves of trees](#). In *Machine Learning, Proceedings of the Twenty-Fifth International Conference (ICML 2008), Helsinki, Finland, June 5-9, 2008*, volume 307 of *ACM International Conference Proceeding Series*, pages 1000–1007. ACM.
- Barry E Stein, Terrence R Stanford, and Benjamin A Rowland. 2020. Multisensory integration and the society for neuroscience: Then and now. *Journal of Neuroscience*, 40(1):3–11.
- Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. [VL-BERT: pre-training of generic visual-linguistic representations](#). In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.
- Hao Tan and Mohit Bansal. 2019. [LXMERT: Learning cross-modality encoder representations from transformers](#). 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 5100–5111, Hong Kong, China. Association for Computational Linguistics.
- Binghao Tang, Boda Lin, Haolong Yan, and Si Li. 2024. [Leveraging generative large language models with visual instruction and demonstration retrieval for multimodal sarcasm detection](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1732–1742, Mexico City, Mexico. Association for Computational Linguistics.
- Erdal Tasci, Caner Uluturk, and Aybars Ugur. 2021. A voting-based ensemble deep learning method focusing on image augmentation and preprocessing variations for tuberculosis detection. *Neural Computing and Applications*, 33(22):15541–15555.
- Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. 2020. [What makes for good views for contrastive learning?](#) In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.
- Yuan Tian, Nan Xu, Ruike Zhang, and Wenji Mao. 2023. [Dynamic routing transformer network for multimodal sarcasm detection](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2468–2480, Toronto, Canada. Association for Computational Linguistics.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shrutu Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. [Multimodal transformer for unaligned multimodal language sequences](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6558–6569, Florence, Italy. Association for Computational Linguistics.
- Michael Tsang, Dehua Cheng, Hanpeng Liu, Xue Feng, Eric Zhou, and Yan Liu. 2020. [Feature interaction interpretability: A case for explaining ad-recommendation systems via neural interaction detection](#). In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.
- Michael Tsang, Dehua Cheng, and Yan Liu. 2018. [Detecting statistical interactions from neural network weights](#). In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.
- Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. [OFA: unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework](#). In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 23318–23340. PMLR.
- Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei

- Zhao, Xixuan Song, et al. 2023. [Cogvlm: Visual expert for pretrained language models](#). *ArXiv preprint*, abs/2311.03079.
- Paul L Williams and Randall D Beer. 2010. Non-negative decomposition of multivariate information. *arXiv preprint arXiv:1004.2515*.
- Thomas Winterbottom, Sarah Xiao, Alistair McLean, and Noura Al Moubayed. 2020. [On modality bias in the TVQA dataset](#). In *31st British Machine Vision Conference 2020, BMVC 2020, Virtual Event, UK, September 7-10, 2020*. BMVA Press.
- Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. [Show, attend and tell: Neural image caption generation with visual attention](#). In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 2048–2057. JMLR.org.
- An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024a. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.
- Dingkang Yang, Shuai Huang, Haopeng Kuang, Yangtao Du, and Lihua Zhang. 2022. Disentangled representation learning for multimodal emotion recognition. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1642–1651.
- Dingkang Yang, Dongling Xiao, Ke Li, Yuzheng Wang, Zhaoyu Chen, Jinjie Wei, and Lihua Zhang. 2024b. Towards multimodal human intention understanding debiasing via subject-deconfounding. *arXiv preprint arXiv:2403.05025*.
- Lei Yu and Huan Liu. 2003. Efficiently handling feature redundancy in high-dimensional data. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*.
- Lei Yu and Huan Liu. 2004. Efficient feature selection via analysis of relevance and redundancy. *The Journal of Machine Learning Research*, 5:1205–1224.
- Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. 2021. [Barlow twins: Self-supervised learning via redundancy reduction](#). In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 12310–12320. PMLR.
- Andy Zeng, Maria Attarian, Brian Ichter, Krzysztof Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, et al. 2022. [Socratic models: Composing zero-shot multimodal reasoning with language](#). *ArXiv preprint*, abs/2204.00598.
- Ce Zhang, Taixi Lu, Md Mohaiminul Islam, Ziyang Wang, Shoubin Yu, Mohit Bansal, and Gedas Bertasius. 2023. [A simple llm framework for long-range video question-answering](#). *ArXiv preprint*, abs/2312.17235.
- Mingda Zhang, Rebecca Hwa, and Adriana Kovashka. 2018. [Equal but not the same: Understanding the implicit relationship between persuasive images and text](#). In *British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018*, page 8. BMVA Press.
- Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. [Minigt-4: Enhancing vision-language understanding with advanced large language models](#). *ArXiv preprint*, abs/2304.10592.

A Asset

In this section, we list all the necessary information for our use of models and data. In our paper, we use MUSTARD (Castro et al., 2019), URFunny (Hasan et al., 2019), MMSD2.0 (Qin et al., 2023) and MMSD (Cai et al., 2019) for our dataset usage. We use ALBEF (Li et al., 2021), BLIP2-OPT-2.7B (Li et al., 2023), Qwen2-0.5B-Instruct (Yang et al., 2024a), Qwen2-1.5B-Instruct, Qwen2-7B-Instruct, Qwen2-72B-Instruct, CogVLM2-LLaMA3-chat-19B (Wang et al., 2023) for our model usage. We show the required information about them and how we follow their requirements when using them.

A.1 Model and Data License

ALBEF ([download link](#))

License: BSD 3-Clause "New" or "Revised"

BLIP2-OPT-2.7B ([download link](#))

License: BSD 3-Clause "New" or "Revised"

Qwen2-0.5B-Instruct ([download link](#))

License: Apache 2.0

Qwen2-1.5B-Instruct ([download link](#))

License: Apache 2.0

Qwen2-7B-Instruct ([download link](#))

License: Apache 2.0

Qwen2-72B-Instruct ([download link](#))

License: Apache 2.0

CogVLM2-LLaMA3-chat-19B ([download link](#))

License: Apache 2.0

A.2 Data License

MUSTARD ([download link](#))

License: MIT

MMSD ([download link](#))

License: Open source, license not specified

MMSD2.0 ([download link](#))

License: Open source, license not specified

URFunny ([download link](#))

License: Open source, license not specified

A.3 Model and Data Use

Personally identifiable information All of the used datasets in this paper are derived from public sources. Therefore, there is no exposure of any personally identifiable information that requires informed consent from those individuals. The used dataset relates to people insofar as it draws text from public sources that relate to people, or people created, obeying related licenses.

Offensive content claim All the used datasets including MUSTARD and MMSD are already public and

widely used. While these datasets may contain instances of offensive content, our work does not aim to generate or amplify such content. Instead, we employ these datasets to study and understand the nature of sarcasm in text. Our use of these datasets follows ethical guidelines, and we do not endorse or support any offensive material contained within them. Moreover, we have implemented measures to mitigate the propagation of offensive content within our research.

B Additional Experimental Results

Besides the models listed in our main sections, we test under more experimental settings with more models. Additionally, we include more baselines for comparison. We also include metrics of precision and recall besides F1 and accuracy that have already been included in the main section. Table 6 shows comprehensive experimental results on all the settings that we run and compare.

B.1 Model Details

Model Name To simplify the terminology in our paper, we use short names for our models. For instance, when we mention BLIP2, we are referring to BLIP2-OPT-2.7B. Similarly, when we refer to Qwen2-0.5B/1.5B/7B/72B, this corresponds to Qwen-2-0.5-0.5B/1.5B/7B/72B-Instruct. Lastly, CogVLM2 refers to CogVLM2-LLaMA3-chat-19B.

Model Size ALBEF consists of a BERT base model with 123.7 million parameters and a ViT-B/16 with 85.8 million parameters, bringing the total to 209.5 million. BLIP2, on the other hand, includes a 2.7 billion-parameter OPT model, a Q-Former, and a ViT. Since the Q-Former and ViT are relatively small compared to OPT, the total size of BLIP2 is approximately 2.7 billion parameters. For the Qwen models, the number of parameters corresponds to the model names: 0.5B, 1.5B, 7B, and 72B. Lastly, CogVLM2 includes a ViT-style vision encoder and a 19 billion-parameter LLaMA3-chat checkpoint. Since the vision encoder and projection parameters are much smaller than LLaMA3-chat, the total size of CogVLM2 is around 19 billion parameters.

C Dataset Details

Statistical information for the splits of 4 multi-modal datasets included in our experiments is

Table 5: **Statistical information for 4 multimodal datasets that we use in our experiments.** † indicates that the validation split is not provided in the original dataset and is conducted by randomly sampling from training data by ourselves.

Dataset	#Train	#Valid	#Test
MUSTARD	251†	83†	356
MMSD	29,040	2,410	2,409
MMSD2.0	19,816	2,410	2,409
URFunny	7,614	980	992

shown in Table 5. We introduce the basic information for each dataset in the following.

MUSTARD contains 690 videos with evenly balanced sarcasm and non-sarcasm labeled points. This dataset is based on English and mainly collected from TV show clips including *Friends*, *The Big Bang Theory*, *The Golden Girls*, and *Sarcasmaholics Anonymous*. Its domain mainly covers daily conversation. The annotation is conducted by two graduate students in two steps: annotating *The Big Bang Theory* first and annotating the remaining ones. Additionally, we use the speaker-independent training and testing splits to make sure that there is no overlap between speakers in the training and testing sets to avoid potential bias.

MMSD collects English tweets containing a picture and some special hashtag (e.g., #sarcasm, etc.) as positive examples (i.e. sarcastic) and collects English tweets with images but without such hashtags as negative examples (i.e. not sarcastic). Furthermore, it excludes tweets with keywords like sarcasm, sarcasm, irony, and irony. Moreover, it discards tweets containing URLs to avoid introducing additional information and discards tweets with words that frequently co-occur with sarcastic tweets and thus may express sarcasm, for instance, jokes, humor, and engagement.

MMSD2.0 is a polished version of MMSD. It removes the spurious cues (e.g., sarcasm word) from the text in the MMSD, which encourages the model to truly capture the relationship across different modalities rather than just memorize the spurious correlation. Additionally, it re-annotates the unreasonable data in MMSD. Therefore, the text information in MMSD2.0 is slightly different from MMSD and part of the labels are re-annotated.

URFunny is a collection of 1866 TED talks, as well as their transcripts, including 1,741 speakers and 417 topics that include speakers from different backgrounds and nationalities and topics from scientific discoveries to everyday ordinary events. The laughter markup is used to filter out 8,257 humorous punchlines from the transcripts. The last sentence is assumed a punchline and similar to the positive instances, the context is chosen.

D Dataset Preprocessing Details

Different multimodal datasets require different preprocessing methods before conducting model training. We include the details of our preprocessing in this section.

MMSD and MMSD2.0 We are only able to extract a total of 24635 images from the released dataset and thus filtered the dataset by the existence of corresponding image IDs. The sizes of validation and test sets are unaffected, while the number of training instances drops to 19,816.

MUSTARD and URFunny There are no existing keyframes in the original dataset. We had to split the videos into frames for use in our image-text models. Typically, we find that key frames matter a lot for the multimodal prediction. Therefore, we used FFmpeg, where we used 1 frame per second to split into frames. Out of the frames extracted per video, we choose the most representative frame by conducting facial expression recognition by DeepFace (Serengil and Ozpinar, 2024) and selecting the frame with the highest emotion intensity score. We thus created the image modality off on the original video dataset.

E Image Description Details

To allow the applicability of our method to pure text-based LLMs, we convert each image into detailed descriptions that include task-related information. We include the details about the process of using CogVLM2-LLaMA3-chat-19B to achieve this in Table 7, 8, and 9.

F Data Categorization Details

To achieve the dataset categorization based on three types of multimodal interaction including redundancy, uniqueness, and synergy. We need to finish this in multiple steps: (1) vision-based prediction collection (2) text-based prediction collection (3) multimodal data categorization. In the following

section, we include the technical details for each of them.

F.1 Vision-based Prediction Collection

We utilize CogVLM2-LLaMA3-chat-19B as our base model vision-only prediction collection. Typically, even though CogVLM2-LLaMA3-chat-19B is a multimodal model, we only include image-side information and only add task-related queries like "Is the image sarcastic or not?" as the input to make sure the model does not utilize text-side information from the multimodal dataset to do the prediction. We conduct few-shot prompting on the train and validation split of all multimodal models. Since MMSD and MMSD2.0 share the same set of images and conduct the same multimodal prediction task, we show three prompts that are used for 4 multimodal datasets in Table 10, 11, and 12.

F.2 Text-based Prediction Collection

We utilize Qwen2-72B-Instruct as our base model for text-only prediction. Typically, we only include text-side information and task-related queries like "Is this image sarcastic or not?" as the input to make multimodal predictions. Even though MMSD and MMSD2.0 do not share the same setting, most of their data is similar. Therefore, we utilize the same prompts for them. We show three few-shot prompts that are used for 4 multimodal datasets in Table 13, 14, 15.

F.3 Multimodal Data Categorization

After collecting unimodal predictions for all multimodal datasets, we conduct our algorithm for categorizing each data point into different multimodal interaction types (redundancy, uniqueness, and synergy) to make sure our categorized data is suitable for training. Typically, to achieve more robust and effective data categorization, we design filtering and rebalancing stages as part of categorization.

Filtering After collecting prediction logits with the model CogVLM2-LLaMA3-chat-19B, we collect the output logits for predictions, which reflect the model’s confidence for “Yes” or “No” responses. To finalize the predictions in multimodal tasks, we apply a softmax operation on these logits to convert them into probabilities. We observed that relying only on vision-related information might lead to inaccurate or uncertain predictions. Therefore, to enhance the reliability of the training data, we remove any data points where the prediction confidence is

below 0.55. These low-confidence predictions are seen as lacking clear patterns and could introduce noise into the training process. By filtering out these training data points, we aim to improve the overall quality and accuracy of the model’s predictions.

Rebalancing Filtering guarantees a high-quality set of pseudo labels. However, the bias from a single modality might cause significant bias in the prediction. The extremely imbalanced distribution of the pseudo labels might lead to the model overfitting to the majority class. To address this issue, we rebalance the dataset by undersampling the majority class. We rank the probability of the prediction from high to low. If the minority of the class is more than 20% of the overall dataset number, we keep the prediction as it is. Otherwise, we consider it as an extremely unbalanced case and make sure that the minority class is at least 20% of the overall dataset to avoid extreme imbalance in the dataset. This helps us avoid expert models (including redundancy, uniqueness, and synergy) overfitting to the majority class and ensures that the model is trained on a balanced dataset.

Categorizing Upon finishing the filtering and rebalancing stage, we have groups of high-quality and balanced unimodal predictions. Therefore, combining it with the ground-truth labels, we categorize the dataset into redundancy, uniqueness, and synergy separately on train, validation, and test splits. The algorithm that is used to conduct the categorization is below.

Algorithm 1 Multimodal Categorization

Require: Text-based label y_1 , Vision-based label y_2 , Ground-truth label y

Ensure: Interaction category: R , U , or S

```
1: if  $y_1 = y_2$  and  $y_1 = y$  then  
2:   return  $R$   
3: else if  $y_1 = y_2$  then  
4:   return  $S$   
5: else if  $y_1 = y$  or  $y_2 = y$  then  
6:   return  $U$   
7: else  
8:   return  $S$   
9: end if
```

G Expert Model Training Details

To improve expert training, we find that instead of starting from the initial pre-trained model check-

point, it’s more effective to initialize the expert training phase using fine-tuned baseline models. This approach leads to faster training and better overall results. The reasoning behind this decision is that continuing training from an already fine-tuned model allows the model to build on its learned features while still maintaining strong performance across the entire dataset. Preserving this capability is essential during inference because the fusion process might assign incorrect nodes to the wrong expert models, and maintaining some general competency helps mitigate such errors from the fusion model and achieve better general performance.

H Fusion Model Training Details

To conduct a model-based fusion, we need to train a fusion model. We use BLIP2 for classifying multimodal interactions and focus on three key categories: redundancy, uniqueness, and synergy. However, these categories are often imbalanced in datasets such as MUSTARD, URFunny, and MMSD2.0, with certain types being underrepresented. To address this imbalance problem, we adopt focal loss as the optimization target:

$$FL(p_t) = -\alpha_t(1 - p_t)^\gamma \log(p_t)$$

where we set $\alpha = 1$ and $\gamma = 2$.

I Experimental Details

We include all the technical details of our experiments including computational requirements and hyper-parameter settings.

I.1 Computational Costs

We utilize $5 \times A6000$ or $1 \times A100$ to run baseline experiments. Expert model training approximately requires 1.5 times longer than baseline training since we need to train redundancy, uniqueness, and synergy models separately. The fusion model training includes a similar training configuration with baselines but just trains under a 3-class classification.

I.2 Hyper-parameter Settings

We use different sets of hyperparameters for the various training settings, including baseline training, expert model training, and fusion model training. We do not perform hyperparameter searches but instead tune the parameters based on the validation set. For LoRA-based fine-tuning, we generally set

the maximum sequence length to 512, rank to 16, scaling factor to 32, and dropout rate to 0.05.

For baseline training, the specific hyperparameters are as follows:

- **For MUSTARD:** We use 10 epochs, a learning rate of $4e-5$, evaluation steps every 100 iterations, and a batch size of 40 for ALBEF and BLIP2. For Qwen2 models, the number of epochs is also set to 10, evaluating at the end of each epoch, and the batch size is 1.
- **For URFunny:** We use 4 epochs, a learning rate of $5e-5$, evaluation steps every 100 iterations, and a batch size of 10 for ALBEF, BLIP2. We use a batch size of 1 for Qwen2, evaluating at the end of each epoch, and the number of epochs is also set to 10.
- **For MMSD2.0:** We use 4 epochs, a learning rate of $5e-5$, evaluation steps every 100 iterations, and a batch size of 10 for ALBEF, BLIP2. We use 5 epochs and a batch size of 1, also evaluating at the end of each epoch, for Qwen2.

For expert model training, we increase the number of epochs to 10, while keeping the other hyperparameters unchanged, to ensure sufficient training.

For fusion model training, the hyperparameters vary across datasets when training BLIP2 on them:

- **For MUSTARD:** We use 50 epochs, a learning rate of $1e-4$, evaluation steps every 20 iterations, and a batch size of 50.
- **For URFunny:** We use 50 epochs, a learning rate of $1e-4$, evaluation steps every 70 iterations, and a batch size of 50.
- **For MMSD2.0:** We use 20 epochs, a learning rate of $1e-4$, evaluation steps every 200 iterations, and a batch size of 50.

I.3 Model Selection Details

In our experiments, which include baseline training, expert model training, and fusion model training, we consistently use the F1 score on the validation set as the metric for model selection. For baseline training, we select the model checkpoint with the highest F1 score on the entire development set. During expert model training, we choose the best expert model checkpoint based on the highest F1 score on the specific subset of the development set

that corresponds to the relevant type of multimodal interaction. For fusion model training, we select the model that has the highest 3-class F1 score.

I.4 Evaluation Details

We used the metrics module from the scikit-learn package for evaluating our prediction tasks. Since our tasks are binary prediction tasks, we chose the binary averaging strategy for precision, recall, and f1. Additional details can be found in the scikit-learn documentation for the metrics module.

I.5 Experimental Statistics

All the available results are based on three different random seeds, with both the mean and standard deviation reported. Typically, F1 results where adding MMOE leads to a statistically significant change (p-value < 0.05) are marked with a * in Table 6. F1 results have a p-value < 0.1 and are marked with a ** in Table 6.

J AI Assistance

We did use ChatGPT as the writing assistant to help us write part of the paper. Additionally, we utilize the power of CodePilot to help us code faster. However, all the AI-generated writing and coding components assisted by AI are manually checked and modified. There is no full AI-generated content in the paper.

Table 6: **Comprehensive results on all types of models and different datasets.** The numbers in the table represent the mean values from 3 runs with 3 seeds, with the corresponding standard variance provided. † indicates that the results include information from audio modality while ours does not.

Model	Acc (†)	Precision (†)	Recall (†)	F1 (†)
MUSTARD				
MuT† (Tsai et al., 2019)	-	65.51	64.78	64.49
LMF† (Liu et al., 2018)	-	70.46	70.34	69.92
LF-DNN-v2† (Ding et al., 2022)	-	65.95	63.88	62.30
LMF (Liu et al., 2018)	-	70.73	70.90	70.68
LF-DNN-v1† (Ding et al., 2022)	-	71.55	71.52	70.99
ALBEF	54.49 \pm 3.13	47.08 \pm 3.03	50.22 \pm 3.62	48.51 \pm 2.21
ALBEF+MMoE	54.49 \pm 2.85	47.36 \pm 2.72	57.68 \pm 4.76	51.95 \pm 2.81
BLIP2	53.75 \pm 9.33	48.46 \pm 4.94	90.13 \pm 9.21	62.65 \pm 2.67
BLIP2+MMoE	59.18 \pm 2.11	51.26 \pm 1.38	87.94 \pm 6.11	64.74 \pm 2.49
Qwen2-0.5B	54.59 \pm 4.35	48.35 \pm 3.31	74.12 \pm 9.40	58.17 \pm 0.86
Qwen2-0.5B+MMoE	49.06 \pm 3.00	45.16 \pm 1.39	88.60 \pm 3.97	59.77 \pm 0.35
Qwen2-1.5B	64.79 \pm 4.11	56.45 \pm 3.53	78.73 \pm 13.18	65.38 \pm 5.16
Qwen2-1.5B+MMoE	70.69 \pm 3.28	60.86 \pm 3.58	89.47 \pm 3.29	72.34 \pm 1.50
Qwen2-7B	72.75 \pm 0.74	63.27 \pm 1.56	86.62 \pm 4.38	72.91 \pm 0.74
Qwen2-7B+MMoE	70.41 \pm 3.23	60.64 \pm 3.57	89.04 \pm 3.38	71.78 \pm 1.47
URFunny				
MuT† (Tsai et al., 2019)	66.65	-	-	-
FDMER (Yang et al., 2022)	70.43	-	-	-
MMIM+SuCl† (Yang et al., 2024b)	70.92	-	-	-
FDMER† (Yang et al., 2022)	71.87	-	-	-
ALBEF	66.77 \pm 0.24	64.29 \pm 1.08	73.74 \pm 2.90	68.67 \pm 0.79
ALBEF+MMoE	67.91 \pm 0.27	65.17 \pm 0.30	75.24 \pm 1.53	69.85 \pm 0.52
BLIP2	70.43 \pm 0.20	65.14 \pm 0.23	86.60 \pm 1.07	74.31 \pm 0.35
BLIP2+MMoE	71.27 \pm 0.30	66.60 \pm 1.23	84.15 \pm 1.95	74.32 \pm 0.36
Qwen2-0.5B	69.29 \pm 0.54	67.16 \pm 1.70	74.15 \pm 1.85	70.46 \pm 0.14
Qwen2-0.5B+MMoE	69.19 \pm 0.14	69.36 \pm 0.07	67.55 \pm 0.44	68.38 \pm 0.20
Qwen2-1.5B	70.43 \pm 0.53	66.03 \pm 0.41	83.13 \pm 2.27	73.51 \pm 0.89
Qwen2-1.5B+MMoE	68.25 \pm 0.81	64.40 \pm 1.87	80.07 \pm 1.37	71.34 \pm 0.52
Qwen2-7B	72.41 \pm 0.52	68.14 \pm 0.65	82.93 \pm 0.43	74.80 \pm 0.55
Qwen2-7B+MMoE	71.88 \pm 0.51	69.18 \pm 0.67	78.16 \pm 2.56	73.29 \pm 0.86
MMSD2.0				
HKE (Liu et al., 2022)	76.50	73.48	71.07	72.25
ViT (Dosovitskiy et al., 2021)	72.02	65.26	74.83	69.72
DynRT-Net (Tian et al., 2023)	71.40	71.80	72.17	71.34
Multi-view CLIP (Qin et al., 2023)	85.64	80.33	88.24	84.10
ChatGLM2-6B (Du et al., 2022b)	80.08	80.52	81.04	80.04
LLaMA2-7B (Touvron et al., 2023)	84.68	84.40	84.94	84.53
LLaVA1.5-7B (Liu et al., 2024)	85.18	85.89	85.20	85.11
LLaVA1.5-7B+DemoRetrieval (Tang et al., 2024)	86.43	87.00	86.30	86.34
ALBEF	81.79 \pm 0.86	77.58 \pm 1.35	81.23 \pm 1.59	79.33 \pm 0.18
ALBEF+MMoE	82.30 \pm 0.31	76.24 \pm 0.24	85.57 \pm 0.42	80.63 \pm 0.32
BLIP2	84.75 \pm 0.99	78.08 \pm 1.65	89.78 \pm 2.71	83.52 \pm 0.04
BLIP2+MMoE	84.82 \pm 0.87	78.87 \pm 1.60	88.49 \pm 2.36	83.38 \pm 0.05
Qwen2-0.5B	81.87 \pm 0.81	75.83 \pm 1.47	85.09 \pm 1.59	80.17 \pm 0.14
Qwen2-0.5B+MMoE	82.27 \pm 0.64	76.02 \pm 1.37	85.92 \pm 4.07	80.67 \pm 1.55
Qwen2-1.5B	83.24 \pm 1.32	78.81 \pm 2.33	83.54 \pm 4.51	81.10 \pm 0.94
Qwen2-1.5B+MMoE	82.76 \pm 0.63	76.70 \pm 1.28	86.21 \pm 3.37	81.16 \pm 0.76
Qwen2-7B	85.28 \pm 1.17	80.38 \pm 0.87	87.05 \pm 2.43	83.58 \pm 1.29
Qwen2-7B+MMoE	84.35 \pm 1.05	78.74 \pm 2.65	87.21 \pm 4.34	82.74 \pm 0.43

Table 7: Prompt for generating image description of MUSTARD

Role	Content
System	<p>Describe the image in detail.</p> <p>If there are people, focus on their emotions, postures, facial expressions, body language, and interactions. Based on this information, infer what event is going on.</p> <p>If there are no people, analyze the event or scene, considering background elements and overall context to infer what event is going on.</p> <p>Provide evidence to predict if the situation is humorous.</p> <p>Ensure the description is between 15 to 100 words.</p>

Table 8: Prompt for generating image description of MMSD and MMSD2.0

Role	Content
System	<p>Describe the image in detail.</p> <p>If there are people, focus on their emotions, postures, facial expressions, body language, and interactions. Based on this information, infer what is the event going on.</p> <p>If there are no people, analyze the event or scene, considering background elements and overall context to infer what is the event going on.</p> <p>Provide evidence to predict if the situation is sarcastic.</p> <p>Ensure the description is between 15 to 100 words.</p>

Table 9: Prompt for generating image description of URFunny

Role	Content
System	<p>Describe the image in detail.</p> <p>If there are people, focus on their emotions, postures, facial expressions, body language, and interactions. Based on this information, infer what is the event going on.</p> <p>If there are no people, analyze the event or scene, considering background elements and overall context to infer what is the event going on.</p> <p>Provide evidence to predict if the situation is sarcastic.</p> <p>Ensure the description is between 15 to 100 words.</p>

Table 10: Prompt for generating image-only prediction of MUSTARD

Role	Content
System	<p>Please analyze the image provided for sarcastic or not. The image is a screenshot of a TV show.</p> <p>If you think the image includes exaggerated emotions (like laughing or looking angry or raising eyebrows) or exaggerated posture (like stretching hands), please answer 'Yes'.</p> <p>If you think the image shows people discussing serious things and just daily routines, please answer 'No'.</p> <p>You need to think about what is the potential event going on in the image.</p> <p>Please make sure that your answer is based on the image itself, not on the context or your knowledge.</p> <p>There are only two options: 'Yes' or 'No'.</p> <p>If you are not sure, please provide your best guess and do not say that you are not sure.</p> <p>You should only make No judgment when you are very sure that the image is not sarcastic. As long as you think potentially it is sarcastic, you should say Yes.</p>

Table 11: Prompt for generating image-only prediction of MMSD and MMSD2.0

Role	Content
System	<p>Please analyze the image provided for sarcastic or not. The image is a screenshot of the image on Twitter. It might include a lot of text, so you need to combine the information of the text in the image.</p> <p>If you think the image includes exaggerated emotions (like laughing or looking angry or raising eyebrows) or exaggerated posture (like stretching hands), please answer 'Yes'.</p> <p>If you think the image includes text that is sarcastic or exaggerated, please answer 'Yes'.</p> <p>If you think the image shows people discussing serious things and just daily routines, please answer 'No'.</p> <p>You need to think about what is the potential event going on in the image.</p> <p>Please make sure that your answer is based on the image itself, not on the context or your knowledge.</p> <p>There are only two options: 'Yes' or 'No'.</p> <p>If you are not sure, please provide your best guess and do not say that you are not sure.</p> <p>You should only make No judgment when you are very sure that the text is not sarcastic. As long as you think potentially it is sarcastic, you should say Yes.</p>

Table 12: Prompt for generating image-only prediction of URFunny

Role	Content
System	<p>You are looking at a screenshot of a TED talk. It is part of the talk and it can be a slide or a speaker.</p> <p>Please analyze the image provided to show whether the image is part of a talk that is showing serious content or trying to show some potentially funny content that can make the audience laugh.</p> <p>If you are looking at a slide, please think about the content of the slide.</p> <p>If the slide is showing some very interesting and informal things, we believe the speaker is trying to make some jokes, and please answer 'Yes'.</p> <p>If the slide is showing some very serious and formal things, we believe the speaker is trying to show some serious content and please answer 'No'.</p> <p>If you are looking at a speaker, please think about the speaker's facial expression and body language.</p> <p>If you think the image includes exaggerated emotions or its body language is exaggerated, we believe the speaker is talking about some informal things and please answer 'Yes'.</p> <p>If you think the speaker in the image looks very serious and formal, they are trying to convey their key points and please answer 'No'.</p> <p>Please make sure that your answer is based on the image itself, not on the context or your knowledge.</p> <p>There are only two options: 'Yes' or 'No'.</p> <p>If you are not sure, please provide your best guess and do not say that you are not sure.</p>

Table 13: Prompt for generating text-only prediction of MUSTARD

Role	Content
System	<p>Please analyze the text provided below for sarcasm. If you think the text includes an exaggerated description or includes strong emotion or its real meaning is not aligned with the original one, please answer 'Yes'. If you think the text is neutral or its true meaning is not different from its original one, please answer 'No'. Please make sure that your answer is based on the text itself, not on the context or your knowledge. There are only two options: 'Yes' or 'No'. If you are not sure, please provide your best guess and do not say that you are not sure. You should only make Yes judgment when you are very sure that the text is sarcastic.</p>
User	TEXT: Yes yes it is! In Prison!!
Assistant	Yes. It expresses the speaker's strong emotion about the situation which indicates that the speaker is sarcastic.
User	TEXT: And then and then you clicked it again, she's dressed. She is a businesswoman, she is walking down the street and oh oh oh she's naked.
Assistant	No. It is a neutral statement.

Table 14: Prompt for generating text-only prediction of MMSD and MMSD2.0

Role	Content
System	<p>Please analyze the text provided below for sarcasm. If you think the text includes an exaggerated description or its real meaning is not aligned with the original one, please answer 'Yes'. If you think the text is neutral or its true meaning is not different from its original one, please answer 'No'. Please make sure that your answer is based on the text itself, not on the context or your knowledge. There are only two options: 'Yes' or 'No'. If you are not sure, please provide your best guess and do not say that you are not sure.</p>
User	TEXT: because lunch is more interesting than job and even tasty...
Assistant	Yes. It expresses the speaker's preference for lunch over the job by using the word 'tasty'.
User	TEXT: gameday ready'
Assistant	No. It is a neutral statement.

Table 15: Prompt for generating text-only prediction of URFunny

Role	Content
System	<p>Please analyze the text provided below for humor or not. If you think the text includes an exaggerated description or it is expressing sarcastic meaning, please answer 'Yes'. If you think the text is neutral or just common meaning, please answer 'No'. Please make sure that your answer is based on the text itself, not on the context or your knowledge. There are only two options: 'Yes' or 'No'. If you are not sure, please provide your best guess and do not say that you are not sure. You should only make No judgment when you are very sure that the text is not funny. As long as you think potentially it is funny, you should say Yes.</p>
User	TEXT: why invite men they are the problem
Assistant	Yes. It expresses that men can be problematic and the speaker is sarcastic to make people laugh.
User	TEXT: we all feel the same things.
Assistant	No. It is a neutral statement.