MMoE: Enhancing Multimodal Models with Mixtures of Multimodal Interaction Experts

Anonymous ACL submission

Abstract

Advances in multimodal models have greatly improved how interactions relevant to various tasks are modeled. Today's models mainly focus on the correspondence between images and text, using this for tasks like image captioning and image-text retrieval. However, this covers only a subset of real-world interactions. Novel interactions, such as sarcasm expressed through opposing spoken words and gestures or figurative descriptions of images, remain challenging. In this paper, we introduce an approach to enhance multimodal models, which we call **M**ultimodal **M**ixtures **o**f **E**xperts (MMoE). The key idea in MMoE is to train separate expert models for each type of interaction, such as redundancy present in both modalities, uniqueness in one modality, or varying degrees of synergy that emerge when both modalities are fused. On two multimodal sarcasm datasets, we obtain new state-of-the-art results. MMoE also provides the framework to design smaller specialized multimodal experts, and improves the transparency of the modeling process.

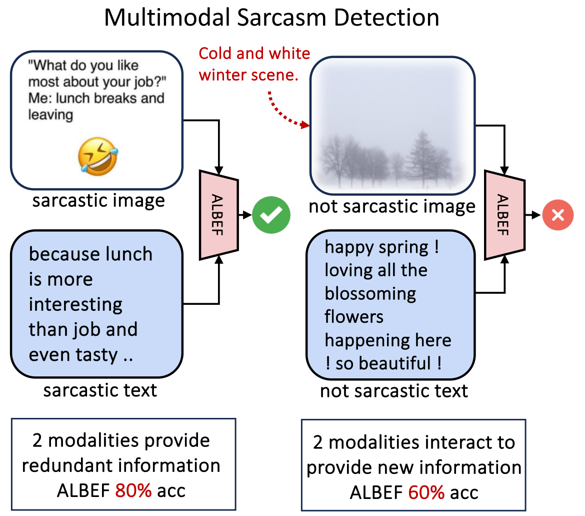


Figure 1: **A single multimodal model cannot handle all types of multimodal interactions well.** For example, ALBEF can handle situations when modalities contain redundant information (e.g., both the text and the image are sarcastic), but struggle when there is synergy between modalities (e.g., the image shows a cold winter scene and the text says it is a happy spring, indicating the user is sarcastic about the weather).

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 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 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 types of interactions 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 only 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 requiring synergistic fusion of image and text. Quantitatively, ALBEF has performance drop of up to 20% on synergistic interactions compared 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 specialized expert models 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 datapoints, MMoE automatically fuses multiple expert models to obtain a final prediction.

MMoE achieves new state-of-the-art results on two multimodal sarcasm datasets we tested on, `MMSarcasm` and `MUSTARD`. Moreover, we show that our approach is easy to implement on different types of models: we used fusion-based vision language models like ALBEF (Li et al., 2021), multimodal language models like BLIP-2 (Li et al., 2023), and image-captioned language models like Qwen2 (qwe, 2024). [1]

2 Related Work

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

Multimodal interactions define 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 have involved semantic definitions based on research in multimedia (Marsh and Domas White, 2003), human (and animal) communication (Partan and Marler, 2005; Flom and Bahrick, 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 (Mazzetto et al., 2021), trained model weights and activations (Sorokina et al., 2008; Tsang et al., 2018, 2019; Hessel and Lee, 2020), feature selection (Ittner 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., 2019; 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 pretrained 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 a wide range of multimodal settings, 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 interactions, which we show to be suboptimal when datasets contain a mix of diverse and complex interactions.

Ensembles and mixture 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 the entire model (Freund et al., 1996). Cheng et al. (2020) utilized 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.,

[1] More information related to the codebase and reproduction of results is available at Appendix §A. We will make the model checkpoints and data public once got accepted.

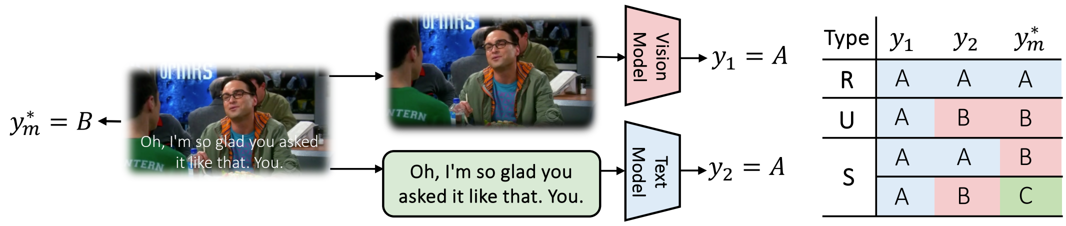


Figure 2: **We classify multimodal datasets into three subsets based on their multimodal interactions**: (1) Redundancy (R), when both modalities agree on the same multimodal label, (2) Uniqueness (U), when modalities disagree and make different predictions, of which one of them is correct, and (3) Synergy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

2021). In settings where it is difficult to define which expert is correct, trainable ensemble functions have been designed to automatically combine multiple experts together in an end-to-end fashion (He et al., 2021; Shazeer et al., 2017; Du et al., 2022). 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 two modalities x_1 and x_2, our goal is to predict the label y using information from both x_1 and x_2. Naturally, the 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 datapoints in a task, our key insight is that each datapoint 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 by (1) *Classifying*: classifying what type of interactions are present in each datapoint in the training set, (2) *Training*: training expert multimodal models to learn each type of interaction, and (3) *Inference*: dynamically ensembling the mixture of expert models during inference on unseen new datapoints. We now explain each of these three steps in detail.

3.1 Classifying 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's predictions. We formalize modality disagreement as follows:

Definition 1. *(Modality disagreement) Given $x_1 \sim \mathcal{X}_1$, $x_2 \sim \mathcal{X}_2$, as well as unimodal classifiers $f_1 : \mathcal{X}_1 \rightarrow \mathcal{Y}$ and $f_2 : \mathcal{X}_2 \rightarrow \mathcal{Y}$, we define modality disagreement as $d(y_1, y_2)$ where $y_1 = f_1(x_1)$, $y_2 = f_2(x_2)$ and $d : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}^{\geq 0}$ is a distance function in label space scoring the disagreement of f_1 and f_2's predictions. Typically, for a multimodal prediction task with a discrete label space $\mathcal{Y}$, the distance function is defined as:*

$$d(y_1, y_2) = \begin{cases} 0, & \text{if } y_1 = y_2 \\ 1, & \text{otherwise} \end{cases} \tag{1}$$

This binary distance function indicates that modalities agree with each other (distance of 0) when f_1 and f_2 produce the same prediction and modalities disagree with each other (distance of 1) when their predictions differ in the discrete label space. It gives us an intuitive way to categorize three types of multimodal interactions:

1. **Redundancy**: when both modalities *agree* with each other on the prediction, and the final multimodal label is the same as each unimodal label, so they contain redundant information.
2. **Uniqueness**: when modalities *disagree* with each other and make different predictions in the label space, of which one of them is the correct multimodal label so that modality contains unique information.
3. **Synergy**: when the multimodal label *disagrees* with either unimodal prediction so there is synergy between modalities that changes the unimodal prediction significantly.

Based on these guidelines, Figure 2 shows an example where we can classify each training datapoint into what type of interaction it exhibits. For each multimodal datapoint (x_1, x_2), we require its true multimodal label $y_m^* = f_m^*(x_1, x_2)$ (labels

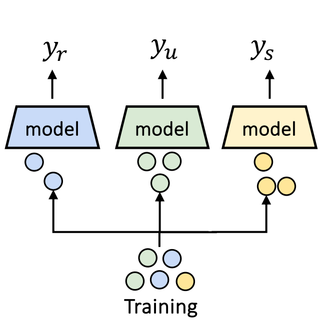


Figure 3: MMoE **training**: Each datapoint is classified based on its multimodal interaction and used to train an expert model tailored only for that interaction.

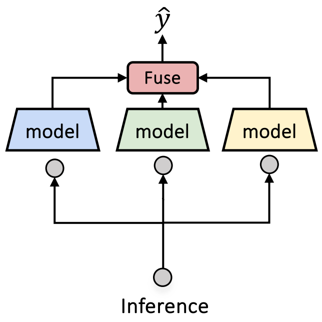


Figure 4: MMoE **inference**: We infer which interaction a test datapoint requires and use a soft weighted fusion over on the outputs from multiple expert models.

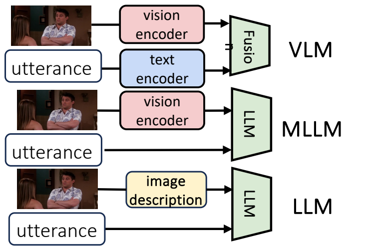


Figure 5: MMoE **applicability**: MMoE can be used as a drop-in layer to multimodal fusion LLMs, multimodal LLMs, and image-captioned LLMs.

are obtained from humans and visible during training), and unimodal predictions $y_1 = f_1(x_1)$ and $y_2 = f_2(x_2)$ obtained from pre-trained unimodal classifiers. Comparing these partial unimodal labels with the ground-truth label enables us to infer the interaction type as follows:

Definition 2. *(**R**edundant, **U**nique, and **S**ynergistic interactions [**RUS**]) Given x_1 and x_2, unimodal partial labels y_1 and y_2, and the ground-truth multimodal label y_m^*. Modalities are redundant when $y_1 = y_2 = y_m^*$, so a measure of redundancy is*

$$R = -d(y_1, y_m^*) - d(y_1, y_2) - d(y_2, y_m^*), \quad (2)$$

Modalities are unique when $y_1 = y_m \neq y_2$ (modality 1 unique) or $y_2 = y_m \neq y_1$ (modality 2 unique), so a measure of uniqueness is

$$U_1 = d(y_2, y_m^*) + d(y_1, y_2) - d(y_1, y_m^*), \quad (3)$$
$$U_2 = d(y_1, y_m^*) + d(y_1, y_2) - d(y_2, y_m^*), \quad (4)$$

Modalities are synergistic when $y_1 = y_2 \neq y_m$ or $y_1 \neq y_2 \neq y_m^*$, so a measure of synergy is

$$S = d(y_1, y_m^*) + d(y_2, y_m^*), \quad (5)$$

In practice, besides the ground-truth multimodal label y_m^*, we obtain unimodal predictions y_1 and y_2 via state-of-the-art unimodal foundation models in the few-shot style for all training datapoints. For vision-only predictions, we utilize vision-language models like CogVLM (Wang et al., 2023) and GPT-4V (Achiam et al., 2023) to obtain them during training by providing only the query and the image. To get text-only predictions, we provide the state-of-the-art language models like CogVLM (Wang et al., 2023) and GPT-4 (Achiam et al., 2023) with the query and the language information so the model answers conditioned only on text for prediction. More information related to the collection of unimodal labels is avaiable at Appendix §F.

3.2 Training one expert model for each multimodal interaction

Given the partitioning of multimodal datasets into subsets each with a similar interaction, this section now describes how we use these interaction-specific datasets to train interaction-specific expert models. Illustrated in Figure 3, there are a total of 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 data distributions they are trained on.

Overall, we first use the estimation process in Section §3.1 to partition each dataset into interaction categories. We then collect all evidences of redundant interactions *across multiple tasks* to train a task-independent redundancy expert f_r. This process is repeated for unique and synergistic interactions, resulting in trained experts f_r, f_u, and f_s. Each expert is trained only on the subset of datapoints that *maximally exhibit that interaction*; this specialization enables experts to be performant at learning that interaction while being smaller in size as it does not have to spend parameters learning other very different interactions. Crucially, *multi-task training* allows us to leverage the power of scale and learn interaction experts that are adaptable to multiple tasks at the same time. For example, the redundancy expert might learn corresponding information between speech and gestures for emotion recognition as well as between images and descriptive captions for image-caption retrieval.

We also note that it is possible to design interaction experts 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). We leave these design explorations for future work.

3.3 Inference with mixture of experts

The conclusion of Section §3.2 yields three expert models each suited for a certain type of multimodal interaction. During inference on unseen test datapoints, we need to select one or more expert models most suitable for that new datapoint. This is a challenge since the categorization of datapoints during training (presented in Section §3.1) requires knowing the ground-truth multimodal label y_m^*, which we have during training but not during inference. One option is to approximate y_m^* with predictions $\hat{y}_m$ from large pre-trained multimodal models, but that is difficult since our goal is to develop a more efficient multimodal model and running state-of-the-art pre-trained models can be slow.

Our key idea is that classifying an interaction is significantly easier than predicting the label itself. Therefore, while pretrained multimodal models might not be able to infer the label y_m^* accurately, they might be able to infer which interaction type that the datapoint belongs to (i.e., predict if modalities have the same or different information, and whether synergistic fusion is required versus actually performing the fusion). Therefore, we *approximately categorize* datapoints 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 datapoint using a pretrained multimodal model (e.g., BLIP-2 in practice). We also test simple baselines like prior constants $\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$ or based on the frequency statistics of each interaction to weight each expert model; see detailed ablation studies on these weights in Section §4.4.

Using these inferred weights, 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 language model architectures and through evaluation on wide range of multimodal tasks with diverse interactions.

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 at Appendix §D.

Models 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). These model categories include:

1. **Fusion-based vision language models (VLMs)** use cross-attention to learn multimodal interactions between all regions of the image with all words in the input text. These models are usually trained from scratch using full-parameter finetuning. Popular examples of such models include ALBEF (Li et al., 2021), LXMERT (Tan and Bansal, 2019) and BLIP (Li et al., 2022).
2. **Multimodal LLMs (MLLMs)** like BLIP-2 (Li et al., 2023) and FROMAGe (Koh et al., 2023) start with an image encoder and a pretrained LLMs as the backbone and only finetune a lightweight transformation from image features to LLM input tokens. Therefore, multimodal extended LLMs are typically trained in a parameter-efficient fine-tuning style.
3. **Image-captioned LLMs (LLMs)** convert images to text using a image captioning model and uses a text-only LLM like Qwen2 (qwe, 2024) on the concatenation of captioned images and text inputs. Examples in this category include Socratic Model (Zeng et al., 2022) and video understanding model (Zhang et al., 2023).

Multimodal prediction tasks We implement both the baselines and our proposed MMoE method on the following two tasks that require learning multimodal interactions between images and text: (1) `MMSarcasm` (Cai et al., 2019) is a multimodal sarcasm detection dataset collected from twitter posts with image-text pairs. It includes 210k image-text pair datapoints annotated for sarcastic and non-sarcastic intents. (2) `MUSTARD` (Castro et al., 2019) is a video-level sarcasm detection dataset including 690 annotated video clips of the

Table 1: MMoE **can beat the state-of-the-art models and be generally applied to any type of model for improvement**. For MUSTARD, Qwen-1.5B out-performs the recent LF-DNN-v1 by 2.25 points. For MMSarcasm, Qwen-1.5B and BLIP-2 reach approximately full marks.

Model	Precision	Recall	F1
	MUSTARD		
MulT	65.51	64.78	64.49
LF-DNN-v1	71.55	71.52	71.08
ALBEF	55.15	49.34	52.08
ALBEF+MMoE	52.20	70.39	59.94
BLIP2	55.45	73.68	63.28
BLIP2+MMoE	56.60	78.95	65.93
Qwen-1.5B	58.40	**91.45**	71.28
Qwen-1.5B+MMoE	**63.46**	86.84	**73.33**
	MMSarcasm		
ALBEF	85.43	86.36	85.90
ALBEF+MMoE	86.81	85.17	85.99
BLIP-2	99.90	99.90	99.90
BLIP-2+MMoE	99.80	100.0	99.90
Qwen-1.5B	**100.0**	**100.0**	**100.0**

TV series. We choose speaker-independent training and testing splits consistent with prior work to avoid potential overlap between speakers.

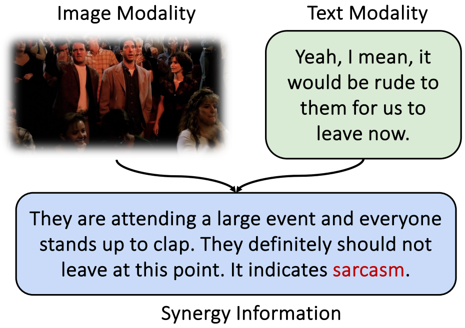
4.2 Main results

We use our results to answer the following research questions. Firstly, we study how the best MMoE model compares to state-of-the-art baselines on the evaluation tasks. Secondly, we study whether MMoE improves performance when applied on top of all three types of base models (multimodal fusion models, multimodal extended LLMs, and image-captioned LLMs).

Overall comparisons with state-of-the-art On both datasets, our best MMoE model substantially improves the state-of-the-art. We beat LF-DNN-v1 (Ding et al., 2022) for MUSTARD with more than 2 points of improvement. Additionally, we find that for MMSarcasm models, the latest models including BLIP-2 and Qwen2-1.5B-Instruct can perfectly answer all the points in the test set correctly. Typically, by comparing MUSTARD and MMSarcasm, we find that MMoE helps gain more improvement on hard dataset (e.g. MUSTARD) that has low F1 but gain less improvement on easy dataset (e.g. MMSarcasm) that already has good performance.

Improvement on various types of multimodal models We first compare performance on traditional cross-attention multimodal fusion models (e.g., ALBEF) with and without MMoE. Based on Table 1, we find that with the help of MMoE, ALBEF performance increases more than 7 points for MUSTARD dataset and around 0.1 point for MMSarcasm dataset. We now apply MMoE to multimodal extended large language models building on top of OPT-2.7b (Zhang et al., 2022). On datasets like MUSTARD, it improves the performance by more than 2 points compared with the baseline. Finally, if we convert the images of MUSTARD and MMSarcasm into image descriptions utilizing GPT-4V and CogVLM, we can use text-only LLMs like Qwen2-7B to conduct experiments. It gains 1.4% improvement on top of image-captioned LLMs.

Figure 6: **Synergy in sarcasm detection**. 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 synergy between modalities when fused together.



4.3 Analysis of MMoE

Given these quantitative results, we further analyze the success of MMoE. We first study the limitations of current models, showing empirical results where one single multimodal model struggles with diverse interactions. We also investigate whether specialized interaction experts can be made smaller, as compared to typically overparameterized models, which can improve efficiency. Finally, we ablate several design decisions in MMoE.

RQ1. What types of multimodal interaction do current models struggle with, and how do expert models perform?

We first show some examples where current methods using a single multimodal model fail to learn specialized interactions, while MMoE can. On the MUSTARD dataset, we classified all data points by their interaction type and found that data points with redundancy, uniqueness, and synergy interaction are highly imbalanced. Redundancy to be 20%, uniqueness to be 50%, and synergy to be 30% in the training data. We find that existing multimodal models including BLIP2 and ALBEF

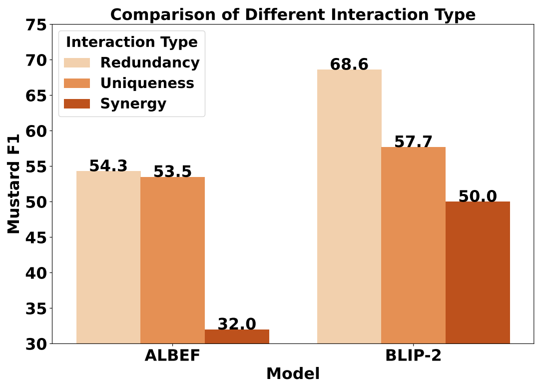


Figure 7: **Multimodal models struggle with synergy much more than redundancy and uniqueness**. Both ALBEF and BLIP-2 showing significantly lower performance on synergistic datapoints compared with redundancy and uniqueness that are split based on ourselves.

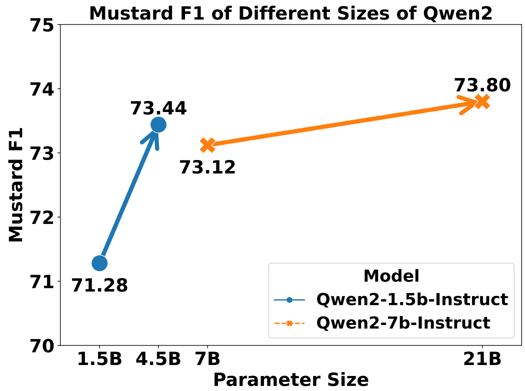


Figure 8: **Qwen2-1.5B-Instruct with MMoE beats single Qwen2-7B-Instruct model**. $3\times$Qwen2-1.5B-Instruct with 3×1.5B parameters beats the Qwen2-7B-Instruct with 7B, indicating MMoE points to a more efficient multimodal architecture.

struggle with synergy multimodal interaction significantly: from Figure 7, we see that they perform at 32% for ALBEF and 50% for BLIP-2, which is significantly lower than for other interactions. We show an example of this failure in Figure 6, where both vision and language contain no clear signal of sarcasm, but when combined, the sarcastic intent is evident. Existing multimodal models fail to learn this interaction between modalities.

While a single large multimodal model may fail, MMoE uses its separate expert models to tackle each type of interaction. Specifically, for `MUSTARD` training with ALBEF, expert training brings improvement from 32.0% to 45.7% on synergy interaction, improvement from 54.32% to 57.95% on redundancy interaction, improvement from 53.5% to 54.4% on uniqueness interaction.

RQ2. How small can expert models be?

It is widely known that neural networks, with enough parameters, are universal approximators of any function. Therefore, sufficiently large multimodal models will eventually be able to approximately learn all interactions, like BLIP-2 can handle all easy interaction cases with one single model for a simple dataset like `MMSarcasm`. However, we hypothesize that expert models that are more specialized for each interaction can be smaller and more efficient while retaining performance.

Overall, to reach the same performance as the traditional finetuning baselines that train one large multimodal model for every interaction, our MMoE approach can be up to 0.36 times smaller in total, and 0.79 times smaller during inference if using only a single expert. Therefore, MMoE presents a path towards more specialized and lightweight multimodal models.

4.4 Ablation studies

Eventually, we test ablations of MMoE components and answer three more research questions including ways of categorizing datapoints based on multimodal interaction types, how we train expert models, and how we fuse expert models.

RQ3. How to categorize datapoints?

We tested the accuracy of categorizing datapoints by their multimodal interactions. Since these interaction values are unknown, we rely on human annotations to provide a gold standard rating using 90 datapoints from the test split of `MUSTARD` (see Appendix §E for more human annotation details). We find that our categorization utilizing unimodal labels and ground-truth labels has an F1 score of 51.26% when compared to human annotation over the 3 interaction types (redundancy, uniqueness, and synergy), indicating that our automatic method is correlated with human judgment. Redundancy interactions are easier to detect, with an accuracy of 64.3%; synergy and uniqueness are harder to detect, with an accuracy of 46.7% and 43.4% respectively. We expect future work on quantifying multimodal interactions to further improve MMoE performance.

To evaluate the effectiveness of data partition for expert training based on interaction categorization, we first ask one research question: *Does class imbalance in data partition hurt expert model training?* Therefore, we design an experiment where we downsample the original RUS partition to make sure each class (redundancy, uniqueness, and synergy) has an equal number of data points. Based on

Table 2: **Ablation study on different data partitioning methods for MMoE**. #R, #U, and #S represent the number of training datapoints for each expert model. We test three settings (1) *RUS partition*: standard interaction classification, (2) *RUS partition (balanced)*: downsample RUS partition to have the same size, and (3) *Random partition*: Keep the partition sizes the same but with random datapoints.

Partition method	#R	#U	#S	MUSTARD F1
RUS partition	57	145	90	**78.65**
RUS partition (balanced)	57	57	57	75.46
Random partition	57	145	90	71.50

Table 3: **Ablation study on various ways of fusing multimodal experts on MUSTARD**. We find that the model-based method is the best, but simple methods like averaging are also enough for strong performance. *Baseline* indicates the performance of existing models (ALBEF, BLIP-2, Qwen2-1.5B) without MMoE.

Fusion function	ALBEF	BLIP-2	Qwen-1.5B
model-based	**59.94**	**65.93**	73.33
average	59.72	63.90	72.87
weighted	57.97	63.31	72.68
cascaded	53.73	63.67	**73.96**
baseline (no fusion)	52.08	63.28	71.28

Table 2, it shows that using as many RUS labeled data points as possible is the most beneficial to MMoE, and downsampling additional data from uniqueness and synergy causes the drop of performance by 3 points.

Then it comes to the second question: *Does the improvement of* MMoE *come just from ensembling expert models?* We would like to discuss whether our improvement is caused by simply ensembling instead of utilizing multimodal interactions. Therefore, we replace our RUS partition data with our randomly selected ones. From Table 2, we find that randomly partition is 7 points worse than our RUS partition, proving that multimodal interaction categorization is crucial for performance gain.

RQ4. How to train expert models?

We ablate whether cross-dataset multitask training helps in training expert models, by pooling together synergy datapoints across multiple datasets including MUSTARD and MMSarcasm to train one synergy expert, similar with redundancy and uniqueness part. While MUSTARD is a small-scale multimodal dataset with only 300+ datapoints for training, MMSarcasm's 190k+ datapoints helps gain an overall improvement of 2.33 points (F1 improves from 61.57 to 63.90). Additionally, synergy experts improve by 2.17 points with the help of an additional 2084 synergy datapoints from MMSarcasm. Therefore, these positive multitask results indicate that multimodal interactions are *universal* properties across all multimodal datasets and expert models that learn specific multimodal features for interaction can be transferred across different datasets.

RQ5. How to fuse expert models?

Finally, we investigate how different choices for the fusion function used to combine multiple expert models together can affect performance. In addition to linear weights, we also test different ways of weighting these experts as inspired by prior literature in MoEs in machine learning and natural language processing (Yuksel et al., 2012).

We find that weights matter a lot for the performance, indicating that different expert models are focusing on different side of multimodal information. Typically, we consider (1) model-based fusion: we train a BLIP-2 model to provide logits that classify test datapoints into redundancy, uniqueness, and synergy type. (2) average fusion: we simply use the average of expert models output logits as the final results. (3) weighted fusion: we pre-define a fixed weight that is 0.2, 0.5, 0.3 based on the approximate proportion of data with those interactions, (4) cascaded fusion: we consider doing inference with the redundancy and uniqueness expert models first; if these two models cannot provide a sufficiently confident decision, we seek help from the synergy expert. Based on Table 3, we find that model-based fusion generally provides the most significant improvement compared with other methods. However, even a simple fusion method through fixed uniform weights provides clear improvements, indicating the robustness of MMoE.

5 Conclusion

This paper proposes a method to enhance multimodal models with a new **M**ultimodal **M**ixtures **of E**xperts structure (MMoE). The key idea is to train separate expert models each tailored to learn a specific type of interaction, which overcomes significant shortcomings of existing multimodal LLMs when diverse types of interactions are simultaneously present. Classifying datapoints into their necessary interactions enables the fusion of expert models during inference, which gives significant boosts to performance and efficiency. MMoE also presents other appealing features of smaller, more efficient specialized experts, and improved transparency of the multimodal modeling process.

Limitations

While we presented 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 (roughly 51% F1 with human annotation) 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 considered modalities that have good unimodal encoders like language models and vision models, 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

There are possible negative societal impacts of our work. Given the framework of our multimodal model based on sarcasm tasks, the improvement and success of our model could allow bad agents to use this technology in a negative manner. Emotion detection models can be used in an inappropriate manner or deployed without proper vetting or understanding in model outputs. Predicting peoples' emotions and using them without consent or consideration can lead to unfair actions and assumptions. We hope to use our paper as a stepping stone for understanding the different noises from modalities from human expression that go into sarcasm and their modeling practices. We do not condone any negative use of these models under any circumstance.

For human evaluation, based on direct communication with our institution's IRB office, this line of research is exempt from IRB, and the information obtained during our study is recorded in such a manner that the identity of the human subjects cannot readily be ascertained, directly or through identifiers linked to the subjects. There is no potential risk to participants and we do not collect any identifiable information from annotators. For the payment, we make sure that our participants are paid with a salary that is higher than the minimum local wage hourly. More details related to human evaluation can be seen in Appendix §E.

References

2024. Qwen2 technical report.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

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. Multi-modal 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). *arXiv preprint arXiv:1906.01815*.

Minhao Cheng, Cho-Jui Hsieh, Inderjit Dhillon, et al. 2020. Voting based ensemble improves robustness of defensive models. *arXiv preprint arXiv: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.

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 arXiv:2303.03378*.

Nan Du, Yanping Huang, Andrew M Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, et al. 2022. Glam: Efficient scaling of language models with mixture-of-experts. In *International Conference on Machine Learning*, pages 5547–5569. PMLR.

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.

Jiaao He, Jiezhong Qiu, Aohan Zeng, Zhilin Yang, Jidong Zhai, and Jie Tang. 2021. Fastmoe: A fast mixture-of-expert training system. *arXiv preprint arXiv: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 *EMNLP*.

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 arXiv:2209.06293*.

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

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 arXiv:2107.12436*.

Andrew Jaegle, Felix Gimeno, Andrew Brock, Andrew Zisserman, Oriol Vinyals, and Joao Carreira. 2021. Perceiver: General perception with iterative attention. *arXiv preprint arXiv:2103.03206*.

Tzyy-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 arXiv: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.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pretraining with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pretraining for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705.

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 arXiv:2306.04539*.

Paul Pu Liang, Yiwei Lyu, Xiang Fan, Jeffrey Tsaw, Yudong Liu, Shentong Mo, Dani Yogatama, Louis-Philippe Morency, and Russ Salakhutdinov. 2022. High-modality multimodal transformer: Quantifying modality & interaction heterogeneity for high-modality representation learning. *Transactions on Machine Learning Research*.

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, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

Sijie Mai, Haifeng Hu, and Songlong Xing. 2019. Divide, 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.

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 Bach. 2021. Semi-supervised aggregation of dependent weak supervision sources with performance guarantees. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, 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. *Advances in neural information processing systems*, 21.

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. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*.

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

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. 2019. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*.

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

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? *Advances in Neural Information Processing Systems*, 33.

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.

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

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

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*, 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 arXiv:2311.03079*.

Paul L Williams and Randall D Beer. 2010. Nonnegative 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. *arXiv preprint arXiv:2012.10210*.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR.

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.

Seniha Esen Yuksel, Joseph N Wilson, and Paul D Gader. 2012. Twenty years of mixture of experts. *IEEE transactions on neural networks and learning systems*, 23(8):1177–1193.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stephane Deny. 2021. Barlow twins: Self-supervised learning via redundancy reduction. In *Proceedings of the 38th International Conference on Machine Learning*, 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 arXiv: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 arXiv: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 (BMVC)*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

A Codebase Link

The anonymous link for the codebase is available **here**. The README file inside provides a detailed guideline on how to run experiments. All the model logits for reproducing the results of the experiments and data split are also available inside.

B 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) and `MMSarcasm` (Cai et al., 2019) for our dataset usage. We use ALBEF (Li et al., 2021), BLIP-2 (Li et al., 2023), Qwen2-1.5B-Instruct (qwe, 2024) and Qwen2-7B-Instruct (qwe, 2024) as our model usage. We show the required information about them and how we follow their requirements when using them.

B.1 Model link and license

ALBEF

Model link: here
License: BSD 3-Clause "New" or "Revised"

BLIP-2

Model link: here
License: BSD 3-Clause "New" or "Revised"

Qwen2-1.5B-Instruct

Model link: here
License: Apache 2.0

Qwen-2-7B-Instruct

Model link: here
License: Apache 2.0

B.2 Data license

`MUSTARD`

Data link: here
License: MIT

`MMSarcasm`

Data link: here
License: MIT

B.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 `MMSarcasm` 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 for the purpose of studying and understanding 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.

Data information

`MUSTARD` This dataset is based on English and mainly collected from TV show clips including Friends, The Big Bang Theory, and so on. Its domain mainly covers daily conversation.

`MMSarcasm` This dataset is based on English and mainly collected from online Twitter content. Its domain mainly covers political, daily life, food, and so on.

C 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.

D Experimental Details

We include all the technical details of our experiments for reproduction.

D.1 Data statistics for experiments

`MUSTARD` contains 690 videos with evenly balanced sarcasm and non-sarcasm labeled points. `MMSarcasm` consists of train, validation, and test sets with sizes of 29040, 2410, and 2409 instances. Images are unique for each instance.

D.2 Model size

We include the size of ALBEF, BLIP-2, Qwen2-1.5B-Instruct, and Qwen2-7B-Instruct model size here. ALBEF has a total size of 3.2GB. BLIP-2-opt-2.7b has a total size of 15.5GB. Qwen2-7B-Instruct has a size of 15.2GB.

D.3 Computational Cost

ALBEF Training

- `MMSarcasm` **Dataset:**
  - 5 A6000 GPUs
  - Baseline training time: 30 minutes
- `MUSTARD` **Dataset:**
  - 5 A6000 GPUs
  - Baseline training time: 5 minutes

BLIP-2 Training

- `MMSarcasm` **Dataset:**
  - 1 A100 GPU
  - Baseline training time: 2 hours
- `MUSTARD` **Dataset:**
  - 1 A100 GPU
  - Baseline training time: 30 minutes

Qwen2-7B-Instruct Training

- `MMSarcasm` **Dataset:**
  - 1 A6000 GPU
  - Baseline training time: 2.5 hours
- `MUSTARD` **Dataset:**
  - 1 A100 GPU
  - Baseline training time: 10 minutes

Qwen2-1.5B-Instruct Training

- `MMSarcasm` **Dataset:**
  - 1 A6000 GPU
  - Baseline training time: 2 hours
- `MUSTARD` **Dataset:**
  - 1 A100 GPU
  - Baseline training time: 6 minutes

D.4 Hyper-parameter

The hyperparameters we tuned for training our models are specified in the paper. We did not tune/do a hyperparameter search across models and kept the same hyperparameters per each unique model we used. We kept our focus on training on different splits of data for each unique model.

D.5 Experimental Statistics

All the available results are based on a single run.

D.6 Parameter for data preprocessing

For `MMSarcasm`, we were 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 19816.

During the training process of ALBEF, images are resized into 384 x 384.

For `MUSTARD`, we had to split the videos into frames for use in our image-text models. We used FFmpeg, where we used 1 frame per second to split into frames. Thus, we created the image modality off on the original video dataset.

D.7 Parameter for evaluation

We used the metrics module from the sci-kit 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 sci-kit learn documentation for the metrics module.

E Human Evaluation Details

In this section, we provide all the technical details for the human evaluation of multimodal interaction classification.

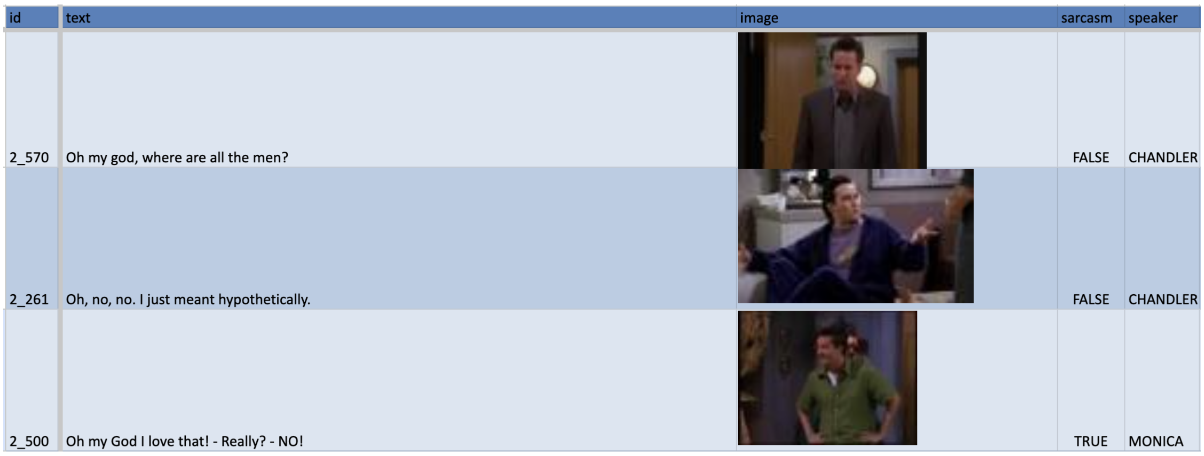
E.1 Human evaluation data

To test whether the model-predicted multimodal interaction type is aligned with human prediction, we select 30 data points that are classified as redundancy by the multimodal model, 30 that are classified as uniqueness, and 30 that are classified as synergy for human evaluation.

E.2 Annotation pipeline

To collect ground-truth labels for the human evaluation data, we implemented a systematic annotation pipeline. Initially, we gathered human-annotated multimodal interaction data for all 90 data points. Each data point was reviewed by multiple participants, and their predictions were aggregated using an ensemble voting method.

In cases where a data point received an equal number of votes for multiple interaction types, these ambiguous points were set aside for further



id	text	image	sarcasm	speaker
2_570	Oh my god, where are all the men?		FALSE	CHANDLER
2_261	Oh, no, no. I just meant hypothetically.		FALSE	CHANDLER
2_500	Oh my God I love that! - Really? - NO!		TRUE	MONICA

Figure 9: **A screenshot of the user interface for human annotation**. The interface is based on Google sheet and users are encouraged to finish one sheet including 90 data points.

review. By the end of the first annotation round, the majority of data points were successfully labeled.

For the remaining uncertain data points, a second round of annotation was conducted. During this phase, we organized a discussion meeting with the group of annotators to deliberate on these ambiguous cases. Through collaborative discussion, the annotators aimed to reach a consensus on the final prediction for each of these data points.

Ultimately, each data point was assigned a single multimodal interaction label based on the majority agreement among annotators. This structured approach ensured that the final dataset was both accurate and representative of diverse perspectives.

E.3 Human instruction

Each human participant, they were told that they needed to provide a multimodal interaction label among redundancy, uniqueness, and synergy for each data point. Typically, in the first step, participants are told that sarcasm refers to content that uses sarcasm, a form of verbal irony where someone says the opposite of what they mean, often for humorous or emphatic effect. Sarcasm can be used to mock or convey contempt, but it can also be used playfully or humorously. Detecting sarcasm in text can be challenging because it relies on context and tone, which are often absent in written communication.

After that, they were asked to see only the text information and only the image information. Based on the text-only information, they provide a yes/no prediction on whether they think the text is expressing sarcastic emotion or not. The same annotation process happens for the image-only side.

After collecting the image-only sarcastic prediction and text-only sarcastic prediction, participants are encouraged to see the ground-truth labels of the data point that indicate whether the ground-truth answer is with sarcasm or without sarcasm. The next step is that they were told redundancy means both image and text modalities provide approximately redundant information about the sarcasm prediction. Uniqueness means that either image or text modalities provide sarcastic information about the prediction. Synergy means that when you combine image and text, your understanding and prediction about the sarcasm prediction switch significantly. They are encouraged to think based on this guidance together with their annotated unimodal labels in the next stage.

Based on all the information and guidelines provided, participants eventually provide an annotation among redundancy, uniqueness, and synergy.

E.4 User interface

The user did the annotation in the Google sheet interface. When doing unimodal side prediction, the other information is hidden. When doing the final redundancy, uniqueness, and synergy prediction, all the information that is available including ground-truth labels, images, and text is available to the participants. Figure 9 shows the UI interface of our annotation.

E.5 Recruitment and Payment

Participants for the annotation tasks were recruited through the authors' networks. We aimed to engage individuals with diverse academic backgrounds to ensure a variety of perspectives in the annotations.

Participants were compensated for their time and effort at a competitive hourly rate. For those residing in the United States, compensation was set above the federal minimum wage. Additionally, one annotator from Switzerland received a payment exceeding the local minimum salary.

E.6 Data consent

Before the process of data collection, we have a consent form selection to ask the participants whether they are willing to have their annotation collected for academic usage.

E.7 IRB approval

Based on direct communication with our institution's IRB office, this line of research is exempt from IRB, and the information obtained during our study is recorded in such a manner that the identity of the human subjects cannot readily be ascertained, directly or through identifiers linked to the subjects. There is no potential risk to participants and we do not collect any identifiable information from annotators.

E.8 Participants details

Four participants participated in our human evaluation experiments for classifying data points based on multimodal interaction. All of them are between the ages of 20-30 and have at least a bachelor's degree in computer science. 3 out of 4 participants are male and 1 left is female. During the experiment, they evaluated 90 sets of multimodal examples related to the Friends TV show and provided predictions on the multimodal interaction type of the data point whether it is redundancy, uniqueness, or synergy based on the provided instruction.

F Unimodal Label Collection

F.1 Vision-only Label Collection

Prompt we used to get zero-shot vision-only prediction for the Mustard dataset with GPT4V:

> **Prompt for Mustard Dataset**
>
> Are the people in the image being sarcastic or not? You need to think based on their figurative language, body language, and facial emotion. Sarcasm often happens when people have intense feelings or emotions. Answer with "Yes" or "No". Follow your initial judgment and explain why.

Prompt we used to get zero-shot vision-only prediction for the MMSarcasm dataset with CogVLM:

> **Prompt for Mustard Dataset**
>
> Think step by step. Does this image contain very obvious sarcasm? Answer yes or no first. Then explain the reason.

F.2 Text-only Label Collection

Prompt we used to get zero-shot text-only prediction for the MMSarcasm dataset with GPT4:

> **Prompt for Mustard Dataset**
>
> Are the people in the image being sarcastic or not? You need to think based on their figurative language, body language, and facial emotion. Sarcasm often happens when people have intense feelings or emotions. Answer with "Yes" or "No". Follow your initial judgment and explain why.

Prompt we used to get zero-shot text-only prediction for the MMSarcasm dataset with CogVLM:

> **Prompt for Mustard Dataset**
>
> Please analyze the text provided below for sarcasm. Begin your response by stating whether the text is sarcastic, answering with a simple 'Yes' or 'No.' Follow your initial judgment with a detailed explanation of your reasoning. Focus on identifying any elements within the text that contribute to a sarcastic tone, such as linguistic cues, context, or contrast between what is said and what may be implied. Text to evaluate:

G Image Description Collection

Prompt we used to get image information for the Mustard dataset with GPT4V:

> **Prompt for Mustard Dataset**
>
> Describe the body language, figurative language, face emotion together with their scenario for characters in the TV show screenshot briefly.

Prompt we used to get image information for the MMSarcasm dataset with CogVLM:

Prompt for Mustard Dataset

Provide a comprehensive description of the image, focusing on its key elements. Include details such as the main subjects, their positions and interactions within the scene, the background setting, and any notable objects or features. Mention the colors, textures, and any text or symbols present. Highlight any action or emotion that is depicted. Also, specify the overall atmosphere or mood of the image, and how these elements collectively contribute to the narrative or message being conveyed.