Text Takes Over: A Study of Modality Bias in Multimodal Intent Detection

[1]**Ankan Mullick**   [2]**Saransh Sharma**   [3]**Abhik Jana**   [1]**Pawan Goyal**

ankanm@kgpian.iitkgp.ac.in   sarsharma@adobe.com
abhikjana@iitbbs.ac.in   pawang@cse.iitkgp.ac.in

[1]IIT Kharagpur, India   [2]Adobe Research, India
[3]IIT Bhubaneswar, India

Abstract

The rise of multimodal data, integrating text, audio, and visuals, has created new opportunities for studying multimodal tasks such as intent detection. This work investigates the effectiveness of Large Language Models (LLMs) and non-LLMs, including text-only and multimodal models, in the multimodal intent detection task. Our study reveals that Mistral-7B, a text-only LLM, outperforms most competitive multimodal models by approximately 9% on MIntRec-1 and 4% on MIntRec2.0 dataset. This performance advantage comes from a strong textual bias in these datasets, where over 90% of the samples require textual input, either alone or in combination with other modalities, for correct classification. We confirm the modality bias of these datasets via human evaluation, too. Next, we propose a framework to debias the datasets, and upon debiasing, more than 70% of the samples in MIntRec-1 and more than 50% in MIntRec2.0 get removed, resulting in significant performance degradation across all models, with smaller multimodal fusion models being the most affected with an accuracy drop of over 50 - 60%. Further, we analyze the context-specific relevance of different modalities through empirical analysis. Our findings highlight the challenges posed by modality bias in multimodal intent datasets and emphasize the need for unbiased datasets to evaluate multimodal models effectively. We release both the code and the dataset used for this work.[1]

1 Introduction

Intent detection refers to classifying user queries into predefined categories and is a key component of systems such as chatbots, search engines, virtual assistants, and robots. While traditional approaches mainly rely on text, multimodal intent detection extends this by incorporating audio and visual inputs.
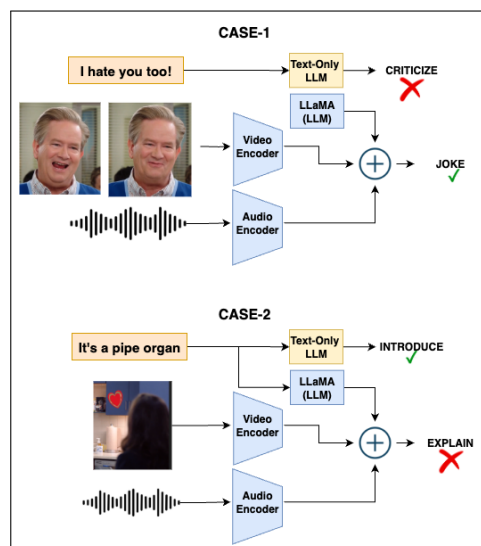

Figure 1: Role of different modalities in Multimodal Intent Detection Task

By considering speech, tone, and facial expressions, these systems capture both what is said and how it is conveyed, making them essential for building more accurate, natural, and context-sensitive applications.

For example, the phrase **"I hate you too!"** may seem negative if judged only by the text and be classified as *Criticize*. However, if the tone or expression is playful, the actual intent is a *Joke*, which can only be identified by considering non-verbal signals (Case-1 in Fig. 1). On the other hand, there are cases where multimodal inputs may confuse the model. In Case-2, the phrase **"It's a pipe organ"** is correctly interpreted as *Introduce* using text alone, but the multimodal model misclassifies it as *Explain*. Hence, a multimodal intent detection model must dynamically balance multiple modalities and prioritize which one or ensemble approach to utilize. To evaluate such a model, we would require gold-standard datasets having no bias towards a particular modality, and preparing such a dataset from scratch is a cumbersome job.

In this paper, we pick existing gold-standard datasets, detect the bias towards any specific modal-

---

[1]https://github.com/Text-Takes-Over-EMNLP-2025/MultiModal-Intent-EMNLP-2025

ity and finally propose a framework to debias the datasets. To validate our claim of bias in existing datasets, we conduct a comprehensive analysis on the performance of state-of-the-art models for the intent detection task using both multimodal and unimodal architectures on both biased and unbiased datasets. We evaluate large language models (LLMs) such as LLaMA, GPT-4o, and Claude, as well as non-LLMs like SDIF, BERT, and MAG-BERT, to assess their ability to capture nuanced intents from text, audio, and visual inputs. Our study includes models ranging from 100 million to 15 billion parameters and explores different training strategies, including full training, parameter-efficient fine-tuning, and zero-/few-shot learning. Our results show that **Mistral-7B**, a text-based LLM, achieves the best performance on both datasets, even surpassing GPT-4o. We evaluate the models on two benchmark multimodal intent detection datasets: MIntRec-1 (Zhang et al., 2022) (**M-1**) and MIntRec2.0 (Zhang et al., 2024) (**M-2.0**). Our findings reveal a strong bias towards the textual modality: **70%** of the samples in M-1 and about **55%** of the samples in M-2.0 can be classified correctly using text alone, with text being necessary for correct classification in more than **90%** of cases. We then re-evaluate the models using the debiased datasets, effectively removing heavily biased intent categories (6 in M-1 and 3 in M-2.0) and reducing the proportion of biased samples.

Re-evaluation of models' performance reveals that the initial results were **heavily inflated** due to textual bias in the dataset. After debiasing, all models struggle significantly on the task, with 60% performance drop observed in smaller multimodal models and more than 50% performance drop in textual models. Human annotation also verifies that around **80%** samples in both datasets are textually biased, which supports our findings.

To summarize, our contributions are as follows: (1) We perform a comprehensive evaluation of models with various sizes and fusion techniques for multimodal intent detection and find that the Mistral-7B textual model outperforms others, despite the task's multimodal nature; (2) We identify the presence and impact of modality bias in existing datasets, analyzing how it affects model performance and capabilities. (3) We propose a debiasing framework that systematically detects and removes biased samples, enabling the construction of more balanced datasets for bias-free intent detection. (4) Through re-evaluation after debiasing, we identify

a significant gap between the expected and actual performance of models, highlighting the need for robust multimodal intent detection frameworks.

2 Related Works

The task of intent detection has been explored through various approaches discussed below:

Generalized Intent Detection Approaches: Researchers investigate several methodologies for intent detection, including few-shot techniques (Xia et al., 2021), zero-shot approaches (Xia et al., 2018), and clustering-based frameworks (Mullick et al., 2022d). Other studies (Wang et al., 2020; Mu et al., 2017b,a) focus on identifying new intents, often formulated as an outlier detection problem.

Intent Detection Applications: There are various domain-specific applications of intent detection task in specialized contexts (Mullick, 2022; Mullick et al., 2022a,b,d; Mullick, 2023a,b; Mullick et al., 2024a). For instance, (Mullick et al., 2023, 2024c) and (Bharti et al., 2020)(MedBot) focus on medical queries and conversational systems, while (Patwa et al., 2021) identifies fake and hostile posts related to COVID-19 in multilingual context. Unlike these works, our study investigates the comparative roles of unimodal and multimodal frameworks in intent detection.

Existing Datasets: Several datasets for intent detection tasks have been proposed. Text-only datasets include SNIPS (Coucke et al., 2018), ATIS (Tur et al., 2010), HWU (Liu et al., 2021), and BANKING (Casanueva et al., 2020). For multimodal contexts, (Kruk et al., 2019) curate MDID dataset with binary labels for Instagram posts' images, while (Singh et al., 2023) proposes the EmoInt dataset featuring 32 emotions and 15 empathetic intents, although it is not publicly available. MIntRec-1 (Zhang et al., 2022) and MIntRec2.0 (Zhang et al., 2024) datasets focus on multimodal intent detection, offering diverse contexts for evaluating models that integrate text, audio, and visual information. MIntRec2.0 focuses more on out-of-scope intents. These two datasets are used in our experiments.

Modality Bias: Modality bias happens when models rely too much on one input, like image or video, while ignoring others. It is common in VQA and video reasoning tasks (Guo et al., 2023; Gao et al., 2024). (Vosoughi et al., 2024) uses counterfactual reasoning to reduce both visual and language shortcuts, while video-dropout training helps with

missing frames but can cause audio bias (Dai et al., 2024). Multimodal intent detection introduces additional challenges, as the dominant modality can vary, and there is no ground-truth modality supervision. But, the impact of modality bias in this setting remains largely unexplored.

3 Dataset

We use two publicly available benchmark multimodal intent detection datasets: MIntRec-1 (Zhang et al., 2022) **[M-1]** and MIntRec2.0 (Zhang et al., 2024) **[M-2.0]**, both under fine-grained settings with text, video, and audio modalities. M-1 contains more than 2000 annotated samples with 20 intent labels and was the first benchmark dataset for multimodal intent classification. M-2.0 is a larger extension, consisting of more than 9000 samples across 30 intent categories, making it the most comprehensive publicly available dataset in this domain. Further dataset details are provided in Table 1 and Appendix section 9.2.

Dataset	Train	Dev	Test	# Intents
M-1	1334	445	445	20
M-2.0	6165	1106	2033	30

Table 1: Statistics of Multimodal Intent Datasets

4 Approach

We consider a diverse set of models spanning different architectures and methodologies to ensure a comprehensive evaluation of multimodal intent detection. We include both state-of-the-art textual and multimodal models, covering both LLM-based and non-LLM approaches. We experiment with smaller-size Large Language Models (LLMs) such as LLaMA-2-7B [L2-7B] and LLaMA-2-13B [L2-13B] (Touvron et al., 2023), LLaMA-3-8B [L3-8B] (Dubey et al., 2024), Qwen2.5-7B [Q-7B] (Yang et al., 2024) and Mistral-7B [M-7B] (Jiang et al., 2023) with fine-tuning, as well as larger-sized LLMs like Claude[2] and GPT-4o[3] using prompting for intent-detection task. We compare against state-of-the-art (SOTA) multimodal models such as MulT (Tsai et al., 2019), MAG-BERT (Rahman et al., 2020), and MISA (Hazarika et al., 2020). To systematically analyze model performance, we categorize the selected approaches into four groups:

(A) Textual Non-LLM Approach: This approach employs BERT (Kenton and Toutanova, 2019)

[bert-base-uncased] as a text encoder to detect intents using textual features.

(B) Textual LLM Approach: We utilize several small-sized Large Language Models (LLMs), including L2-7B, L2-13B, L3-8B, Q-7B and M-7B.

(C) Multimodal Non-LLM Approach: We examine multiple state-of-the-art non-LLM multimodal frameworks, including MISA, MulT, MAG-BERT and SDIF-DA (Huang et al., 2024), which integrates the Shallow-to-Deep Interaction Framework with Data Augmentation (SDIF). BERT, MISA, MulT, MAG-BERT are tested by the authors of M-1 and M-2.0. SDIF, a newer method, outperforms others on M-1 and is thus used in this study.

(D) Multimodal LLM (MLLM) Approach: This approach leverages advanced multimodal LLMs, including: (i) Video-ChatGPT (Maaz et al., 2024) [VChatGPT], which combines a CLIP-like visual encoder with Vicuna LLM; (ii) Video-LLaVA (Lin et al., 2024) [VLLaVA] which uses an encoder-decoder architecture, where LanguageBind encoders align image and video features into a unified textual space for the LLM decoder (e.g., Vicuna) to process; (iii) GPT-4o and Claude-3 in two settings: text-only (*ClaudeT, GPT-4T*) and text with six uniformly sampled video frames (*ClaudeV, GPT-4V*); (iv) Video-LLaMA (Zhang et al., 2023) [VLLaMA], which utilizes separate vision (Video Q-former for spatiotemporal visual features) and audio branches (pre-trained ImageBind and Q-former for audio representations).

The training strategy of a model is determined by its size and resource requirements. Smaller models,

Model [Approach]	M-1 Acc	M-1 F1	M-2.0 Acc	M-2.0 F1
BERT [A]	70.8	67.4	57.1	49.3
M-7B [B]	**82.9**	**82.5**	**65.2**	**64.4**
L2-7B [B]	79.3	79.5	57.1	56.7
Q-7B [B]	72.4	64.6	61.6	62.4
L3-8B [B]	77.3	77.1	61.3	59.9
L2-13B [B]	80.7	80.4	54.8	54.3
Mult [C]	71.5	67.9	58.4	51.5
MAG [C]	72.7	68.6	58.2	49.4
MISA [C]	71.8	69.1	57.8	51.9
SDIF [C]	72.8	71.6	58.6	52.5
ClaudeT [D]	57.7	56.4	39.7	38.1
ClaudeV [D]	59.1	55.9	40.9	39.2
GPT-4T [D]	60.4	59.8	42.1	41.8
GPT-4V [D]	59.5	58.8	41.8	41.4
VLLaMA [D]	23.6	23.4	11.8	11.0
VLLaVA [D]	32.4	30.5	18.8	17.5
VChatGPT [D]	33.3	33.0	18.7	16.8

Table 2: Performance of different models on original dataset (Acc and F1 in %)

	M-1 Performance														
	BERT	**M-7B**	**L2-7B**	**Q-7B**	**L3-8B**	**L2-13B**	**MulT**	**MAG**	**MISA**	**SDIF**	**ClaudeT**	**ClaudeV**	**GPT-4T**	**GPT-4V**	**VChatGPT**
Apologize	96.3	96.3	**100.0**	96.3	92.6	92.6	96.3	97.3	97.4	96.3	**100.0**	**100.0**	85.2	96.3	92.6
Flaunt	48.0	50.0	70.0	**80.0**	50.0	70.0	46.0	42.0	54.0	60.0	40.0	40.0	60.0	60.0	58.3
Joke	26.0	70.0	80.0	20.0	70.0	**90.0**	33.0	32.0	40.0	60.0	30.0	30.0	50.0	50.0	11.8
Taunt	16.7	8.3	16.7	25.0	16.7	25.0	16.7	11.1	26.7	16.7	16.7	8.3	33.3	41.7	**84.0**
Thank	98.0	**100.0**	**100.0**	88.0	96.0	**100.0**	97.6	99.5	99.6	96.0	**100.0**	84.0	84.0	92.0	60.0
	M-2.0 Performance														
Apologize	92.9	97.0	94.0	**100.0**	93.8	96.9	94.5	94.5	93.2	93.2	98.5	98.5	81.5	84.6	81.5
Flaunt	9.1	36.4	9.1	18.2	27.3	9.1	15.4	8.2	21.8	20.0	36.4	36.4	**73.7**	**73.7**	7.4
Joke	4.8	10.3	17.2	6.9	6.9	3.4	4.1	0.0	15.2	12.4	28.6	28.6	**48.1**	37.0	0.0
Taunt	14.5	31.0	18.9	34.5	20.7	37.9	25.2	4.8	23.8	23.1	20.7	20.7	**37.0**	**37.0**	26.3
Thank	96.5	**100.0**	98.3	94.8	**100.0**	98.3	96.5	96.5	96.2	96.5	94.8	94.8	81.0	81.0	67.6
Refuse	14.1	58.8	41.2	0.0	29.4	35.3	12.9	10.6	23.5	22.3	47.1	52.9	52.9	52.9	**84.5**

Table 3: MIntRec intent-wise results (Intent Accuracy in %); Models grouped based on approach (A to D in order). Due to space constraints extended tables for M-1 and M-2.0 are in Appendix Tables 14, 16, 17 and 18

such as those based on BERT, are fully trained due to their manageable size ($\sim$ few hundred MBs). For textual LLMs, fine-tuning is performed using memory-efficient techniques, including quantization to 4-bit precision and low-rank adapter (LoRA) fine-tuning, to optimize resource usage. Class A and B models utilize only text modality, and classes C and D (except GPT-4T/V, and ClaudeT/V) utilize audio, video, and text modality. GPT-4V and ClaudeV utilize text and image modality. GPT-4T and ClaudeT utilize text only.

Fine-tuning MLLMs is challenging due to size, alignment issues, and modality-specific noise, making accuracy metrics across models non-comparable. This study focuses on analyzing multimodal intent detection methods rather than optimizing performance. Detailed configurations are provided in Appendix Section 9.3.

5 Experiment and Results

Our experiment consists of three phases. In **Phase-1**, we analyze overall and intent-wise model performance. **Phase-2** then finds the smallest set of modalities needed to classify each example correctly. Finally, **Phase-3** checks for modality biases in datasets and measures model performance after de-bias techniques.

5.1 Phase-1: Initial Performance Analysis

We evaluate various strategies to handle multiple modalities on two datasets with overall and intent-wise accuracy and F1-score. To enhance multimodal LLM performance, we apply in-context learning by adding a few example input-output pairs in the prompt. The model's predictions are then compared with the true intent labels. In cases where the model generates an explanation instead of a label (about 6% in M-1 and 15% in M-2.0), we select the label with the highest BERTScore similarity to the output.

a) Overall Results: We compare several methods across four model categories on the M-1 and M-2.0 datasets using accuracy and F1-score, as shown in Table 2. Text-only LLMs, especially M-7B, perform best overall with accuracy of 82.9% on M-1, and 65.2% on M-2.0. Traditional multimodal non-LLM models (C) like Mult, MAG, MISA, and SDIF do reasonably well, with SDIF being the strongest in this group. These models outperform multimodal LLMs (D), but still fall short of text-only LLMs (B). Among multimodal LLMs, GPT-4T performs the best, but remains lower than others. This is likely due to their general-purpose chatbot type design, which limits their ability to handle structured multimodal intent tasks. Note that we do not focus on achieving state-of-the-art results; instead, our goal is to understand per-class behavior and how different types of models perform under resource constraints. Further analysis is in the Appendix section 9.5.

b) Intent-wise Performance: Performances of different models across different intents and evaluations (accuracy) are shown in Table 3 for both MIntRec-1 (Extended Appendix Table 14) and 2.0 datasets (Extended Appendix Table 16). Textual LLMs excel in certain intents like *Agree*, *Thank*, and *Greet*, achieving 100% accuracy with the textual modality alone. However, for intents such as *Taunt*, *Flaunt*, and *Warn*, multimodal models outperform textual models due to the additional contextual information from visual or other modalities. Small-scale multimodal models perform better on intents like *Care*, *Apologize*, and *Ask for help*, while larger multimodal LLMs excel in more complex intents like *Taunt*, *Praise*, *Invite*, and *Complain*. On the other hand, naive multimodal fusion techniques, which treat all modalities with equal importance, can degrade performance for intents where textual cues are dominant. Overall (Table 2), textual LLMs and small-scale multimodal models

offer practical solutions for resource-constrained settings, while larger multimodal models deliver superior performance on more nuanced intents.

5.2 Phase-2: Analysis of Modality Importance

Given that the textual model M-7B achieves the best overall performance on both datasets, and often surpasses their multimodal counterparts, we investigate the relative contribution of each modality by examining whether samples require multimodal or unimodal information for correct classification.

a) Optimal Modality Combination Selection: We conduct modality-wise ablation using *SDIF* for M-1 and *MulT* for M-2.0, as these are the best-performing Class-C models for their respective datasets. This analysis helps to determine which combination of modalities contributes most to accurate classification. To measure the importance of each modality, we perform ablation by masking individual modalities. Each masked modality's feature vector is replaced with a zero tensor. The models are evaluated in 7 different settings: Text (T), Video (V), Audio (A), Text+Video (T+V), Text+Audio (T+A), Video+Audio (V+A), and Text+Video+Audio (T+V+A).

b) Automated Annotation Framework: We train *SDIF* and *MulT* under different masking setups to identify the minimal combination of modalities required for accurate classification. For each sample, the models produce seven output probabilities corresponding to the seven modality combinations. For example, the probability corresponding to the Text+Audio setting represents the model's output probability for the ground label when provided with text and audio while masking video. Using these seven probability features, we determine the smallest modality combination that correctly classifies the sample or emits the highest probability for the ground label. This analysis provides insight into the optimal modality fusion required for accurate classification in multimodal intent detection.

c) Key Findings: As shown in Table 4, *Text* is the most dominant modality for intent detection, followed by the combination of *Text+Audio+Video*. When combining all modality combinations that include text, approximately 90% of the samples of both datasets require text for correct classification. This observation suggests that the dataset is not only textually biased, given that a high fraction of samples can be correctly classified using text alone, but also that text serves as the primary and most significant modality. This finding underscores the cru-

Modality Combination	M-1 % samples	M-2.0 % samples
Text (T)	**69.66**	**57.85**
Video (V)	3.60	0.00
Audio (A)	4.72	4.53
T+V	1.35	7.18
T+A	3.37	0.54
V+A	0.67	1.48
T+V+A	16.63	28.43
Σ T	**91.01**	**94.00**
Σ V	22.25	37.09
Σ A	25.39	34.97

Table 4: % of samples for distribution of modality combinations in the M-1 and M-2.0 datasets. Σ T denotes the total percentage of samples that require text for correct classification, i.e., the combined share of samples in the **T**, **T+V**, **T+V+A**, and **T+A** categories.

	T+V+A	T	V	A
Acc	**72.8**	69.9	17.8	28.5
Agree	**100.0**	63.2	61.4	40.4
Apologize	**96.3**	81.4	11.6	48.8
Arrange	68.2	**96.3**	7.4	63.0
Ask Help	50.0	**100.0**	0.0	64.0
Care	**84.2**	39.1	0.0	13.0
Complain	66.7	**91.7**	0.0	8.3
Greet	**83.3**	40.0	0.0	10.0
Inform	70.2	**72.2**	0.0	16.7
Introduce	57.1	**89.5**	0.0	5.3
Joke	60.0	**61.4**	61.4	40.4
Praise	**79.1**	66.7	0.0	4.8
Thank	**96.0**	66.7	16.7	50.0

Table 5: Ablation analysis for M-1. SDIF is the best-performing small multimodal model

cial role of text in intent recognition, establishing it as the foundational modality in multimodal models (relative to MIntRec datasets). Further exploratory analysis details are in the Appendix Section 9.6.

5.3 Phase-3: Debiasing MIntRec Datasets

Given that over 90% of samples in both datasets rely on text for classification, we further analyze the nature of these samples. Specifically, we characterize textually biased examples, those correctly classified using only text, and those benefiting from additional modalities. This sheds light on the dataset's modality sensitivity and helps identify the true utility of multimodal learning in this task.

a) Defining Textual Bias: Text plays a dominant role in determining intents, with textual models achieving 100% accuracy on specific intents such as *Thank*, *Greet*, and *Apologize*. Further details are in Appendix Fig. 10. Table 5 (Extended Appendix Tables 24 and 25) indicates that models relying solely on text perform significantly better than those using other modalities alone. This sug-

gests the presence of textual bias in the dataset. Upon manual inspection, we identify two types of bias: explicit and implicit. Explicit bias occurs when a statement directly contains words associated with a specific intent. For instance, phrases such as "I am sorry" or "Thank you" contain words like "sorry" and "thank", which are directly linked to corresponding intent labels (Appendix Fig. 5). Implicit bias refers to cases where the intent is not explicitly stated but can be inferred from the context. For example, the sentence "Hannah, can you update the SIM card on the robot?" does not contain words directly tied to an intent category but can be inferred as a request based on the model's understanding of the context.

b) Automated Debiasing Framework: We propose a systematic framework for debiasing the dataset by identifying a subset of samples with reduced textual bias. Our goal is to focus on statements where textual models may struggle due to the absence of other modalities. To achieve this, we select three models: BERT (text-only), L2-7B (LLM), and SDIF (a small multimodal model). BERT is chosen for its strong textual understanding, L2-7B is selected instead of a more capable model like M-7B to ensure that even a relatively basic LLM can identify obvious textual biases, and SDIF evaluates the impact of multimodal inputs with masked audio and video cues.

The identification of textually biased samples follows a structured process. Since both datasets maintain a training, development, and test set ratio of approximately 3:1:1, we divide the training dataset into three equal parts randomly. A round-robin approach is applied where the first three parts form the training set, the fourth part serves as the development set, and the last part is used as the test set. This process is repeated for all five splits. Using these sets, three models are trained: LLaMA-2-7B is fine-tuned with QLoRA, BERT is fully trained on the training set, and the small multimodal model is trained with masked audio and video inputs. During training, models are evaluated on the development set, and early stopping is used to prevent overfitting. Once trained, they predict labels for the test set. To identify textually biased samples, any test sample that is correctly classified by the majority of these three models is labeled as textually biased. This iterative process ensures a systematic and controlled identification of textual bias while maintaining a balanced dataset for multimodal intent detection.

Model [Approach]	M-1		M-2.0	
	Acc	F1	Acc	F1
BERT [A]	16.7	23.6	18.5	22.6
M-7B [B]	32.4	38.7	**33.6**	34.7
L2-7B [B]	30.6	33.8	30.1	33.4
Q-7B [B]	35.2	35.2	31.3	30.9
L3-8B [B]	25.0	30.0	32.2	**35.2**
L2-13B [B]	37.0	38.0	31.5	34.1
Mult [C]	27.8	29.4	22.7	25.4
MAG [C]	19.4	25.8	21.0	24.1
MISA [C]	22.2	26.6	21.4	24.7
SDIF [C]	22.2	25.6	19.4	23.1
ClaudeT [D]	34.3	32.9	24.5	22.6
ClaudeV [D]	29.6	27.6	25.8	23.6
GPT-4T [D]	44.9	40.2	26.0	26.1
GPT-4V [D]	**52.9**	**54.1**	30.5	31.0
VLLaMA [D]	9.3	7.4	4.8	4.8
VLLaVA [D]	11.1	9.6	9.9	8.8
VChatGPT [D]	18.0	19.7	12.3	15.5

Table 6: Results of different models after Phase-3 i.e. Debiasing (Acc and F1 in %)

After debiasing, we observe that approximately 70% of the samples in M-1 and 50% in M-2.0 exhibit textual bias (more details in Appendix Fig. 11). Intent categories such as *Agree*, *Apologize*, *Care*, *Greet*, *Praise*, *Thank* in M-1, and *Apologize*, *Thank*, *Greet* in M-2.0 (more details in Appendix Tables 26 and 27) show over 85% biased samples. In contrast, categories such as *Taunt* and *Flaunt* exhibit less than 25% bias, indicating that these statements cannot be inferred by textual models alone. Further analysis in Appendix Section 9.5.

c) Debiased Dataset Creation and Filtering: To obtain the debiased datasets, all textually biased samples are removed. However, after this step, the six intent categories (*Agree*, *Apologize*, *Care*, *Greet*, *Praise*, *Thank*) of M-1 and the three of M-2.0 (*Apologize*, *Thank*, *Greet*) with high bias are left with very few samples across different splits. Due to this limited representation, we exclude these categories. The final debiased dataset consists of 14 intent categories in M-1 and 27 in M-2.0.

The debiased dataset for M-1 consists of 290 train, 107 dev, and 108 test samples. For M-2.0, the dataset includes 2,826 train, 475 dev, and 891 test samples (Fig. 11). To validate the consistency of the observations, prior experiments are re-conducted on the debiased datasets.

d) Performance Analysis on Debiased Datasets: Table 6 shows model performance after debiasing, with all models dropping compared to Table 2. 1) On M-1, GPT-4V performs best and outperforms GPT-4T on both datasets, showing the need for visual cues. 2) Textual LLMs show a sharp de-

					M-1 Performance After Debiasing										
	BERT	M-7B	L2-7B	Q-7B	L3-8B	L2-13B	MulT	MAG	MISA	SDIF	ClaudeT	ClaudeV	GPT-4T	GPT-4V	VChatGPT
Arrange	0.0	0.0	14.3	**42.9**	28.6	14.3	28.6	0.0	0.0	14.3	14.3	28.6	20.0	14.3	0.0
Flaunt	0.0	0.0	**50.0**	16.7	33.3	33.3	16.7	0.0	0.0	16.7	**50.0**	33.3	33.3	**50.0**	40.0
Joke	33.3	33.3	**83.3**	66.7	50.0	50.0	33.3	16.7	16.7	33.3	50.0	16.7	50.0	62.5	16.7
Taunt	0.0	0.0	30.0	30.0	0.0	0.0	0.0	20.0	20.0	10.0	10.0	10.0	**100.0**	**100.0**	14.3
Apologize, Thank	Removed														
					M-2.0 Performance After Debiasing										
Arrange	5.9	5.9	5.9	11.8	0.0	**23.5**	5.9	17.7	17.7	5.9	6.3	6.3	5.9	5.1	0.0
Flaunt	0.0	38.9	27.8	**44.4**	5.6	33.3	11.1	0.0	5.6	11.1	27.8	27.8	25.5	30.6	0.0
Joke	12.0	28.0	12.0	24.0	4.0	8.0	4.0	4.0	12.0	8.0	29.2	29.2	40.0	**44.0**	0.0
Taunt	8.7	43.5	8.7	54.4	23.9	23.9	17.4	28.3	15.2	17.4	17.4	17.4	66.7	**72.7**	7.7
Refuse	0.0	21.4	28.6	**71.4**	35.7	21.4	7.1	0.0	0.0	7.1	35.7	42.9	58.8	64.3	9.1
Apologize, Thank	Removed														

Table 7: MIntRec intent-wise results post debiasing (Intent Accuracy in %); Models grouped on the basis of approach; Due to space constraints, extended tables for M-1 and M-2.0 are in Tables 28, 29, 32 and 33

cline, especially in the *Joke*, *Comfort*, and *Flaunt* (Table 7) intents where visual cues like facial expressions or tonal cues play a crucial role. 3) Small multimodal models, which were competitive before, now perform similarly to BERT, highlighting the limits of BERT-based multimodal approaches and needs for better fusion mechanisms to effectively balance the contributions of different modalities. 4) MLLMs like VChatGPT achieve better results in categories such as *Acknowledge* (Table 7) despite the absence of fine-tuning. These findings suggest that advanced techniques improve the recognition of specific intents. More detailed and intent-specific results are in Appendix Section 9.7 and Table 30 and 31.

e) Analyzing the Impact of Dataset Size vs. Textual Bias on Performance: We analyze whether the performance drop observed in different classes of models is due to the samples being textually unbiased or a reduction in sample size. We use a random subset of the original dataset. We then train several models on this subset to compare their performance with the debiased dataset. The train, test, and dev splits of the random sample and the debiased sample are the same. Table 8 indicates that the observed performance drop across various model categories is primarily due to their inability to accurately classify textually unbiased samples, rather than a reduction in dataset size.

5.4 Experimental Settings

All our experiments are conducted on a single A100 GPU with 80 GB RAM and 2*960 GB SSD. The training of small multimodal models takes ~1 hour. Text-LLMs require ~3-5 hours for training and 1-2 hours for inference. MLLMs need 1-4 hours for inference. Small multimodal models are trained using a cross-entropy loss with a learning rate of 5e-6 over 100 epochs. The Text-LLMs are fine-tuned using prompts and 4-bit quantization, supported by QLoRA (Dettmers et al., 2023), with a learning rate of 2e-4. AdamW optimizer, enabled with early termination, is used in all cases. We use the NLTK, spaCy, and NumPy packages for further analysis and ablations. Further details are in Appendix section 9.4.

6 Human Annotation Verification

To assess automated annotation accuracy, we randomly sample 15-20% of M-1 (444 samples) and M-2.0 (1,489 samples) for manual verification. Three Annotators are selected to classify samples into five modality categories: Text, Video, Audio+Text, Video+Text, or Audio+Video+Text, excluding audio-only samples. Initial labeling is done by two annotators and any annotation discrepancy is checked and resolved by the third annotator after discussing with others. Annotators are chosen after discussions and conditions of fulfilling many criteria like domain knowledge expertise along with a good working proficiency in English. Inter-annotator agreement 85.7%.

Results indicate strong textual bias, with over 80% of samples in both datasets primarily relying on text. In M-1, the distribution was: T (82.46%), T+V (5.69%), T+V+A (5.01%), T+A (4.56%), and V (2.28%). In M-2.0, it was: T (80.63%), T+V+A (7.60%), T+V (6.50%), T+A (4.31%), and V (0.96%). Findings suggest that humans rely more on visual cues than audio for interpretation. A possible approach to address bias is training a classifier to predict the required modality subset for each sample, ensuring decisions are based only on the most relevant modalities. Detailed guidelines are in Appendix section 9.8.

7 Ablation Analysis

Modality Contributions: Table 5 highlights the impact of individual modalities. Audio column exhibits greater variation than the video column,
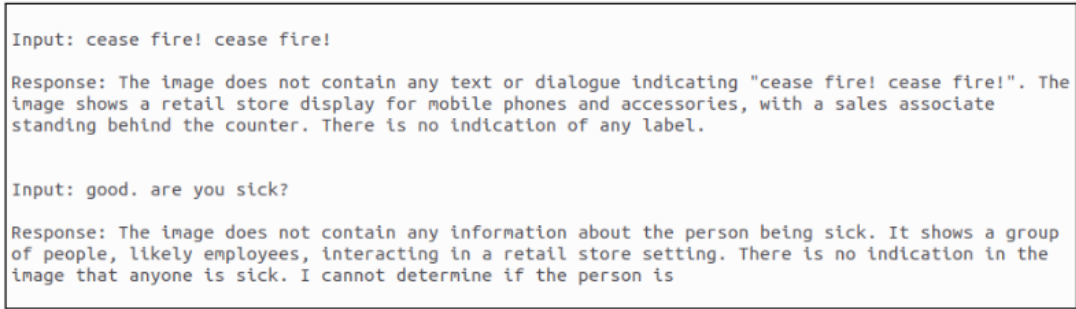
```
Input: cease fire! cease fire!

Response: The image does not contain any text or dialogue indicating "cease fire! cease fire!". The
image shows a retail store display for mobile phones and accessories, with a sales associate
standing behind the counter. There is no indication of any label.


Input: good. are you sick?

Response: The image does not contain any information about the person being sick. It shows a group
of people, likely employees, interacting in a retail store setting. There is no indication in the
image that anyone is sick. I cannot determine if the person is
```

Figure 2: Multimodal model confusion caused due to image frames provided

Model	M-1				M-2.0			
	Random		Debiased		Random		Debiased	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1
BERT [A]	67.6	56.4	16.7	23.6	57.7	52.7	18.5	22.6
M-7B [B]	72.2	73.2	32.4	38.7	59.3	57.7	33.6	34.7
SDIF [C]	59.4	59.0	22.2	25.6	59.2	52.1	19.4	23.1
GPT-4T [D]	73.1	75.3	44.9	40.2	50.7	49.6	26.0	26.1
GPT-4V [D]	63.0	66.7	52.9	54.1	50.7	49.5	30.5	31.0

Table 8: Performance Comparison: Random Subset vs. Debiased Data for MIntRec Datasets

which predominantly classifies samples into broad categories (e.g., positive, negative, neutral). This suggests that tonal variations in speech provide richer information than visual features. Over 80% of intents like *Apologize*, *Thank*, and *Praise* show strong textual bias, whereas intents such as *Joke*, *Taunt*, and *Asking for opinions* require video for accurate classification (Appendix Fig. 7 and 8). For further details refer to Appendix 9.5.3.

Effectiveness of Multimodal Models: Multimodal models like MulT, SDIF perform better by using text to guide attention over audio and video on textually biased original dataset. Large MLLMs (e.g., VLLaMA, VLLaVA, VChatGPT) struggle due to limited fine-tuning and high computational requirements. Smaller multimodal models achieve comparable or better performance, making them more practical for real-world applications. Further details on fusion techniques in Appendix 9.9.

Generalizability Analysis: Generalizability is a key concern for intent detection since user intents evolve and datasets differ in label space. Traditional multimodal and BERT-based models, with fixed output sizes, often need retraining to adapt. Large language models, however, show stronger transfer across datasets. For instance, Mistral-7B fine-tuned on MIntRec2.0 reached 72.58% accuracy and 72.85% F1 on MIntRec-1, while a model trained on MIntRec-1 achieved 47.47% accuracy and 40.42% F1 on MIntRec2.0. The drop mainly comes from label mismatches, but overall performance remains competitive with major baselines. These findings highlight the promise of LLMs for

handling evolving intents with less retraining. Full analysis is provided in Appendix 9.5.2.

Analysis of Fair Comparisons: Due to resource limits, textual LLMs are fine-tuned with QLoRA, while multimodal LLMs are tested in few-shot settings. This raises concerns about fairness, since textual models may appear to perform better largely because they benefit from fine-tuning. We argue that the performance gap is not only due to differences in model size or training strategies, but is also strongly influenced by text-biased samples in the dataset, which naturally favor textual LLMs over multimodal ones. To study this, we run an ablation with LLaMA-3-8B Instruct and Mistral-7B Instruct, checking if fine-tuning explains the advantage of textual LLMs. We use the same few-shot prompts and system messages for fairness for all models.

Dataset	Model	M-1 (Acc)	M-1 (F1)	M-2.0 (Acc)	M-2.0 (F1)
Original	LLaMA-3-8B-Instruct	44.49	38.34	31.19	27.40
	Mistral-7B-Instruct	64.27	59.80	39.80	34.27
Debiased	LLaMA-3-8B-Instruct	26.85	15.93	25.64	5.73
	Mistral-7B-Instruct	26.85	19.78	28.65	5.00

Table 9: Performance of Instruction-tuned models on MIntRec-1 and 2.0 datasets (Original vs Debiased)

Table: 9 shows that few-shot textual LLMs perform worse than their fine-tuned counterparts, confirming that fine-tuning provides task-specific benefits. However, the observed gap between textual and multimodal LLMs cannot be attributed solely to fine-tuning. While fine-tuning alters the absolute accuracy levels, it does not fundamentally change the relative behavior; textual LLMs consistently outperform open-source multimodal models even without fine-tuning on the original dataset. This suggests that the benchmarks themselves are predominantly text-heavy, with only a limited contribution from other modalities. Furthermore, our debiasing step reduces the dataset's over-reliance on textual content, causing both fine-tuned and non-fine-tuned textual models to drop in performance, thereby narrowing the gap. This underscores that

Input	Textual LLM	Multimodal Model	Type of Issue
Ugh, calm down	Comfort	Taunt	Missing facial expression cues
It's a pipe organ	Introduce	Explain	Object not visible in frames
Great job	Praise	Taunt	Missing tonal cues
Anyway, Zephra asked me to set up a call	Arrange	Arrange	Confusion in similar labels (Inform)

Table 10: Erroneous cases in textual and multimodal models, highlighting intent recognition limitations

the original dataset's strong text bias is a substantial factor, but not the only reason for the performance differences observed.

8 Discussion and Error Analysis

We closely examine the mistakes made by the evaluated models and find the following key insights:

a) Multimodal LLMs: These models often struggle to understand speaker intent in situations involving emotions or abstract meanings. While they use both visual and audio cues, they get confused when the visual information does not match the spoken words. For example, if someone says *"It is a pipe organ"* (correct label: *Introduce*), the model may fail when the pipe organ is not visible in the video. In such cases, it may rely on unrelated signals like facial expressions or the background, leading to errors, as shown in Table 10. Image or audio quality can also play a major role in performance. When a video is converted into a small set of frames (sampled uniformly or randomly), the model may become confused because it loses the temporal information linking the frames. This issue is shown in Figure 2. In addition, these models require considerable computational resources for fine-tuning and have difficulty combining multiple modalities when those signals are weak or noisy.

b) Textual LLMs: These models work well for intent classes that are clear from text alone but struggle when body language or tone is needed. For example, *"Ugh, calm down"* is often mislabeled as *Comfort* instead of *Taunt*, or *"Great job"* is labeled as *Praise* instead of *Taunt*. These errors show the limitations of relying solely on text for tasks that require a deeper understanding of tone or context.

c) Small Multimodal Models: These models with BERT backend act like a noisier version of BERT and only give small improvements, if any. When textually biased samples are removed, their performance drops significantly. They also suffer from similar issues as larger multimodal models, misalignment of modalities, and difficulty handling abstract or emotion-driven samples.

Correlated Labels: All models face challenges when intent classes are highly similar. For example, *"Anyway, Zephra asked me to set up a call"* is incorrectly labeled as *Arrange* instead of *Inform*. Though both labels might seem valid, a closer look shows the true intent is to inform. Both textual and multimodal models miss fine distinctions like these. More details are provided in Appendix 9.10.

Our analysis shows that the modality needed for correct classification varies across samples. Some samples require only text, others need audio or visual cues. Any fixed rule at the dataset or label level fails to capture this variation. To handle the sample-specific nature of modality use, future work should explore adaptive models that can route and process inputs based on their modality needs. Mixture-of-experts or routing-based frameworks can help focus compute on the relevant parts of the input. This also requires datasets that are better balanced across modalities and robust fusion methods that can work well even when only certain modalities are informative. Refer to Appendix Section 9.11 for more discussion.

9 Conclusion

In this work, we study multimodal intent detection using benchmark datasets like MIntRec. We observe that Mistral-7B, a text-only model, achieves the best performance due to the strong text bias in the datasets. While around 50% and 70% of the samples can be accurately classified using only textual information, over 90% of the samples require textual input, either alone or in combination with other modalities, for correct classification, which limits the usefulness of these benchmarks for evaluating multimodal models. To address this, we propose a debiasing strategy that removes text-dominant samples (around 70% and 55% samples in M-1 and M-2.0, respectively). After debiasing, model performances drop significantly. The drop exceeds 60% in smaller multimodal models, while textual models also see a drop of more than 50%, showing that current models depend heavily on text features. Our comprehensive analysis shows the need for unbiased datasets, more effective fusion frameworks, and adaptive approaches that can select relevant modalities for each sample.

Limitations

Our work has the following limitations:

(I) While our experiments focus on the MIntRec datasets, we have not conducted experiments on domain-specific data due to the lack of publicly available datasets, restricting the scope of our findings. Future research can explore intent detection in specific domains such as education and healthcare.

(II) The experimental datasets used in this study are limited to English, and extending this work to a multilingual setting, particularly for low-resource languages, would provide a broader understanding of multimodal intent detection and address language-related biases.

(III) Fine-tuning MLLMs is challenging due to their large size, modality alignment issues, and modality-specific noise. These challenges lead to variations in training approaches; however, since our objective is to gain deeper insights into model behavior rather than achieve state-of-the-art performance, the majority of our observations should remain unaffected.

(IV) Due to hardware limitations, we were unable to fine-tune the open-source MLLMs. While fine-tuning could have improved the performance of the best model, it is unlikely to change key findings such as the impact of textual bias or the relative performance of different models.
Addressing these limitations in future studies would provide a more comprehensive evaluation of the models and methods explored in this work.

Ethical Concerns

We use publicly available datasets and approaches, ensuring compliance with all ethical guidelines. The datasets do not contain personally sensitive information, avoiding ethical concerns.

References

Urmil Bharti, Deepali Bajaj, Hunar Batra, Shreya Lalit, Shweta Lalit, and Aayushi Gangwani. 2020. Medbot: Conversational artificial intelligence powered chatbot for delivering tele-health after covid-19. In *2020 5th international conference on communication and electronics systems (ICCES)*, pages 870–875. IEEE.

Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. 2020. Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 38–45, Online. Association for Computational Linguistics.

Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, Maël Primet, and Joseph Dureau. 2018. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *Preprint*, arXiv:1805.10190.

Yusheng Dai, Hang Chen, Jun Du, Ruoyu Wang, Shihao Chen, Haotian Wang, and Chin-Hui Lee. 2024. A Study of Dropout-Induced Modality Bias on Robustness to Missing Video Frames for Audio-Visual Speech Recognition . In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27435–27445, Los Alamitos, CA, USA. IEEE Computer Society.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems*, volume 36, pages 10088–10115. Curran Associates, Inc.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Liqing Gao, Fan Lyu, Peng Shi, Lei Zhu, Junfu Pu, Liang Wan, and Wei Feng. 2024. Overcoming modality bias in question-driven sign language video translation. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(11):11724–11738.

Souradip Guha, Ankan Mullick, Jatin Agrawal, Swetarekha Ram, Samir Ghui, Seung-Cheol Lee, Satadeep Bhattacharjee, and Pawan Goyal. 2021. Matscie: An automated tool for the generation of databases of methods and parameters used in the computational materials science literature. *Computational Materials Science (Comput. Mater. Sci.)*, 192:110325.

Yangyang Guo, Liqiang Nie, Harry Cheng, Zhiyong Cheng, Mohan Kankanhalli, and Alberto Del Bimbo. 2023. On modality bias recognition and reduction. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(3):1–22.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM international conference on multimedia*, pages 1122–1131.

Shijue Huang, Libo Qin, Bingbing Wang, Geng Tu, and Ruifeng Xu. 2024. Sdif-da: A shallow-to-deep interaction framework with data augmentation for multimodal intent detection. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10206–10210.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego

de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

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.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2024. Video-LLaVA: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5971–5984, Miami, Florida, USA. Association for Computational Linguistics.

Xingkun Liu, Arash Eshghi, Pawel Swietojanski, and Verena Rieser. 2021. *Benchmarking Natural Language Understanding Services for Building Conversational Agents*, pages 165–183. Springer Singapore, Singapore.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2024. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*.

Xin Mu, Kai Ming Ting, and Zhi-Hua Zhou. 2017a. Classification under streaming emerging new classes: A solution using completely-random trees. *IEEE Transactions on Knowledge and Data Engineering*, 29(8):1605–1618.

Xin Mu, Feida Zhu, Juan Du, Ee-Peng Lim, and Zhi-Hua Zhou. 2017b. Streaming classification with emerging new class by class matrix sketching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Ankan Mullick. 2022. Multilingual intent detection and key information extraction for domain specific applications.

Ankan Mullick. 2023a. Exploring multilingual intent dynamics and applications. *IJCAI Doctoral Consortium*.

Ankan Mullick. 2023b. Novel intent detection and active learning based classification (student abstract). *arXiv e-prints*, pages arXiv–2304.

Ankan Mullick, Animesh Bera, and Tapas Nayak. 2021. Rte: A tool for annotating relation triplets from text. *arXiv preprint arXiv:2108.08184*.

Ankan Mullick, Sombit Bose, Abhilash Nandy, Gajula Chaitanya, and Pawan Goyal. 2024a. A pointer network-based approach for joint extraction and detection of multi-label multi-class intents. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15664–15680.

Ankan Mullick, Akash Ghosh, G Sai Chaitanya, Samir Ghui, Tapas Nayak, Seung-Cheol Lee, Satadeep Bhattacharjee, and Pawan Goyal. 2024b. Matscire: Leveraging pointer networks to automate entity and relation extraction for material science knowledge-base construction. *Computational Materials Science*, 233:112659.

Ankan Mullick, Surjodoy Ghosh D, Shivam Maheswari, Srotaswini Sahoo, Suman Kalyan Maity, and Pawan Goyal. 2018a. Identifying opinion and fact subcategories from the social web. In *Proceedings of the 2018 ACM International Conference on Supporting Group Work*, pages 145–149.

Ankan Mullick, Pawan Goyal, and Niloy Ganguly. 2016. A graphical framework to detect and categorize diverse opinions from online news. In *Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES)*, pages 40–49.

Ankan Mullick, Pawan Goyal, Niloy Ganguly, and Manish Gupta. 2017a. Extracting social lists from twitter. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, pages 391–394.

Ankan Mullick, Pawan Goyal, Niloy Ganguly, and Manish Gupta. 2018b. Harnessing twitter for answering opinion list queries. *IEEE Transactions on Computational Social Systems*, 5(4):1083–1095.

Ankan Mullick, Mukur Gupta, and Pawan Goyal. 2024c. Intent detection and entity extraction from biomedical literature. *LREC-COLING 2024*, page 271.

Ankan Mullick, Shivam Maheshwari, Pawan Goyal, and Niloy Ganguly. 2017b. A generic opinion-fact classifier with application in understanding opinionatedness in various news section. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 827–828.

Ankan Mullick, Ishani Mondal, Sourjyadip Ray, R Raghav, G Chaitanya, and Pawan Goyal. 2023. Intent identification and entity extraction for healthcare queries in indic languages. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1825–1836.

Ankan Mullick, Abhilash Nandy, Manav Kapadnis, Sohan Patnaik, R Raghav, and Roshni Kar. 2022a. An evaluation framework for legal document summarization. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4747–4753.

Ankan Mullick, Abhilash Nandy, Manav Nitin Kapadnis, Sohan Patnaik, and R Raghav. 2022b. Fine-grained intent classification in the legal domain. *arXiv preprint arXiv:2205.03509*.

Ankan Mullick, Shubhraneel Pal, Tapas Nayak, Seung-Cheol Lee, Satadeep Bhattacharjee, and Pawan Goyal. 2022c. Using sentence-level classification helps entity extraction from material science literature. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4540–4545.

Ankan Mullick, Sourav Pal, Projjal Chanda, Arijit Panigrahy, Anurag Bharadwaj, Siddhant Singh, and Tanmoy Dam. 2019. D-fj: Deep neural network based factuality judgment. *Technology*, 50:173.

Ankan Mullick, Sukannya Purkayastha, Pawan Goyal, and Niloy Ganguly. 2022d. A framework to generate high-quality datapoints for multiple novel intent detection. *arXiv preprint arXiv:2205.02005*.

Parth Patwa, Mohit Bhardwaj, Vineeth Guptha, Gitanjali Kumari, Shivam Sharma, Srinivas Pykl, Amitava Das, Asif Ekbal, Md Shad Akhtar, and Tanmoy Chakraborty. 2021. Overview of constraint 2021 shared tasks: Detecting english covid-19 fake news and hindi hostile posts. In *International Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situation*, pages 42–53. Springer.

Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, AmirAli Bagher Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. 2020. Integrating multimodal information in large pretrained transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2359–2369, Online. Association for Computational Linguistics.

Gopendra Vikram Singh, Mauajama Firdaus, Asif Ekbal, and Pushpak Bhattacharyya. 2023. Emoint-trans: A multimodal transformer for identifying emotions and intents in social conversations. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:290–300.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti 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 conference. Association for computational linguistics. Meeting*, volume 2019, page 6558. NIH Public Access.

Gokhan Tur, Dilek Hakkani-Tür, and Larry Heck. 2010. What is left to be understood in atis? In *2010 IEEE Spoken Language Technology Workshop*, pages 19–24. IEEE.

Ali Vosoughi, Shijian Deng, Songyang Zhang, Yapeng Tian, Chenliang Xu, and Jiebo Luo. 2024. Cross modality bias in visual question answering: A causal view with possible worlds vqa. *IEEE Transactions on Multimedia*, 26:8609–8624.

Min Wang, Ke Fu, Fan Min, and Xiuyi Jia. 2020. Active learning through label error statistical methods. *Knowledge-Based Systems*, 189:105140.

Congying Xia, Wenpeng Yin, Yihao Feng, and S Yu Philip. 2021. Incremental few-shot text classification with multi-round new classes: Formulation, dataset and system. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1351–1360.

Congying Xia, Chenwei Zhang, Xiaohui Yan, Yi Chang, and S Yu Philip. 2018. Zero-shot user intent detection via capsule neural networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3090–3099.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 543–553.

Hanlei Zhang, Xin Wang, Hua Xu, Qianrui Zhou, Kai Gao, Jianhua Su, jinyue Zhao, Wenrui Li, and Yanting Chen. 2024. MIntrec2.0: A large-scale benchmark dataset for multimodal intent recognition and out-of-scope detection in conversations. In *The Twelfth International Conference on Learning Representations*.

Hanlei Zhang, Hua Xu, Xin Wang, Qianrui Zhou, Shaojie Zhao, and Jiayan Teng. 2022. Mintrec: A new dataset for multimodal intent recognition. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, page 1688–1697, New York, NY, USA. Association for Computing Machinery.

Appendix

9.1 Motivation for Multimodal Intent Detection

Multimodal intent detection is chosen as the focus of this study because it is interpretable, involves dynamic selection of input types, and provides a compact but challenging setup to evaluate fusion strategies. Unlike tasks such as Visual Question Answering (VQA), where the main input modality is usually fixed (e.g., vision), intent detection requires the model to decide which combination of audio, video, and text is most useful for each example. This makes the task more complex and better aligned with real-world scenarios, where input relevance can vary significantly.

The task has a wide range of practical applications, including robotics, conversational agents, and task-driven systems. Robots interacting with humans need to understand intent through gestures, tone, and spoken content. Chatbots and virtual assistants also benefit from multimodal understanding to respond more effectively. Although the study focuses on intent detection, the proposed tools and findings are general and can be applied to other multimodal tasks such as visual dialogue, video-based agents, and embodied interaction.

This work takes a systematic approach to understanding multimodal intent detection by analyzing a variety of models and datasets. The analysis reveals strong textual biases in existing benchmarks, 70% in the M-1 dataset and 50% in M-2.0 (see Fig. 11), indicating that models often rely heavily on text while ignoring other modalities. In the second phase of the study, a detailed modality-wise analysis shows that certain intents like *Apologize* and *Thank* are easy to detect from text alone, while others, such as *Complain*, *Flaunt*, and *Inform*, benefit significantly from audio or video input (Figs. 8, 9). These findings highlight the need for more balanced and less biased datasets to evaluate multimodal models fairly. This concept of intent is different from the notion of entity-phrases (Mullick et al., 2021; Guha et al., 2021; Mullick et al., 2022c, 2024b) and opinion (Mullick et al., 2016, 2017a,b, 2018a,b, 2019).

To address this issue, the study introduces a method to reduce textual bias in datasets and evaluates model performance under these more controlled conditions. In addition, it proposes tools to measure how different modalities contribute to predictions and how models adapt when a modal-ity is missing or misleading. These tools provide deeper insights into model behavior beyond simple accuracy. While the experiments focus on intent detection, the approach is general and can support the development of robust systems for any task involving multiple input modalities. The dynamic nature of modality relevance in intent detection makes it an ideal testbed for studying adaptive multimodal learning in real-world scenarios.

9.2 Dataset

MIntRec-1 **[M-1]** (Zhang et al., 2022) contains 2,224 annotated samples and is introduced as the first benchmark dataset for multimodal intent classification with 20 intent labels across text, audio, and video modalities.

The MDID dataset (Kruk et al., 2019) provides binary intent labels using Instagram posts with image context. EmoInt-Trans (Singh et al., 2023) presents the EmoInt dataset with annotations for 32 emotions and 15 empathetic intents. However, the dataset is currently not publicly accessible due to technical issues.

MIntRec2.0 **[M-2.0]** (Zhang et al., 2024) is the largest available dataset for multimodal intent detection, labeled under a 30-class intent taxonomy, with inputs from text, audio, and video sources.

In our study, we use both MIntRec-1 and MIntRec2.0 under the fine-grained intent classification setting, which consists of 20 and 30 intent labels, respectively.

The authors of the MIntRec datasets have evaluated various baseline models, including BERT, MAG-BERT, MISA, MulT, and SDIF_DA. They claim that extensive experiments on their datasets show that using multimodal data leads to better performance compared to only using text. However, based on our observations, we find that both MIntRec datasets contain a significant number of samples that are heavily biased towards the textual modality.

9.3 Approach

9.3.1 Textual LLM Approach

We chose LLaMA and Mistral models because they cover a range of practical sizes, their weights are openly available, and they have strong community support. By working with models in the 7 billion to 13 billion parameter range, we avoid the engineering overhead of very large models while still observing how capacity affects performance. This

```
Each of the input sequences either indicate an emotion expressed by the speaker or an intent
expressed by speaker. From the list of given 20 labels, identify which label best describes the
emotion or intent of the speaker in the input sequence. List of labels:
advise
agree
apologise
arrange
ask_help
care
comfort
complain
criticize
flaunt
greet
inform
introduce
joke
leave
oppose
praise
prevent
taunt
thank


Input: jonah, you set up the chairs. cheyenne, can you make sure that the sound system's working?
### Response: arrange


### End
```

Figure 3: Textual LLM prompt for finetuning

```
Each of the input sequences either indicate an emotion expressed by the speaker or an intent
expressed by speaker. From the list of given 20 labels, identify which label best describes the
emotion or intent of the speaker in the input sequence. List of labels:
advise
agree
apologise
arrange
ask_help
care
comfort
complain
criticize
flaunt
greet
inform
introduce
joke
leave
oppose
praise
prevent
taunt
thank

Input: text
Response: label


Input: but i could really use the help.
Response: Ask for help


Input: 'cause he eats a lot of sugar, so he'd be the sweetest.
Response: Joke


Only generate the label from the list of above 20 labels.
```

Figure 4: Inferencing prompt (for models used without training, few-shot)

24041

size span lets us study small-to-mid models that can run on single GPUs without distributed setups.

According to Hugging Face download statistics, the LLaMA2-7B [4] model has significantly more (last month) downloads compared to the relatively newer Qwen2-7B [5] and Qwen2.5-7B [6] models. The relatively recent Qwen2.5-7B shows growing interest, but it remains well behind LLaMA2-7B. These figures show that any insights we draw from LLaMA models will be broadly relevant to the research community and practitioners already using these checkpoints.

Our experiments include four models: LLaMA2-7B, LLaMA2-13B, LLaMA3-8B, and Mistral-7B, alongside Qwen2.5-7B for comparison as a newer 7 billion-parameter alternative. Comparing LLaMA3-8B to Mistral-7B holds the size constant to highlight differences in architecture and training data. Comparing LLaMA2-7B to LLaMA2-13B isolates the effect of roughly doubling parameters within the same family. Including Qwen2.5-7B adds perspective on a newer entrant at the 7 billion-parameter scale.

Our goal is not to declare one model universally **best**, but to understand how design choices and size affect behavior along two key dimensions. First, we measure intent-specific performance: how accurately each model follows different task types, from direct factual queries to multi-step instructions. Second, we test bias mitigation by probing for unwanted associations and stereotyping, then comparing how those biases change with model size and training source. By focusing on relative performance within each model class and size, we ensure our observations reflect architectural and training differences rather than raw capacity alone.

9.3.2 Multimodal Non-LLM Approach

The following multimodal baselines are utilized for our study:

MISA (Hazarika et al., 2020): The fundamental idea behind MISA (Multimodal Invariant and Specific Representations) lies in learning two distinct subspaces for each modality: a modality-invariant subspace and a modality-specific subspace. In essence, MISA takes multiple modalities into account by learning representations that capture both shared features and modality-specific

---

	Input	Output
GPT-4o	$5.00 / 1M tokens	$15.00 / 1M tokens
Claude-3-Haiku	$0.25 / 1M tokens	$1.25 / 1M tokens

Table 11: Close Source API Cost

nuances, thus enabling a more comprehensive understanding of the underlying affective content in multimodal data.

MulT (Tsai et al., 2019): The Multimodal Transformer (MulT) model has a crossmodal attention module, which allows the model to attend to crossmodal interactions across entire utterances. This module dynamically adapts features from one modality to another, facilitating the modeling of long-range crossmodal contingencies without the need for manual word alignment preprocessing.

MAG-BERT (Rahman et al., 2020): MAG-BERT handles multiple modalities by leveraging its Multimodal Adaptation Gate (MAG) component, which enables BERT to process multimodal input data effectively. This framework allows BERT to adapt seamlessly to the integration of nonverbal information, such as visual and acoustic cues, alongside linguistic input. During the fine-tuning process, MAG-BERT utilizes attention mechanisms conditioned on nonverbal behaviors to map informative features from different modalities to a vector representation.

SDIF-DA (Huang et al., 2024): The Shallow-to-Deep Interaction Framework with Data Augmentation (SDIF-DA) utilizes a hierarchical structure composed of shallow and deep interaction modules. Initially, the shallow interaction module preliminarily aligns video and audio features with text features. Subsequently, a KQV attention mechanism is employed to obtain bimodal representations of video and audio aligned with text. These aligned features are then concatenated with the original text features to create a trimodal representation, facilitating shallow integration across modalities. SDIF-DA incorporates a ChatGPT-based data augmentation approach to fine-tune the text encoder, aiming to enhance the model's performance.

9.3.3 Multimodal LLM Approach

With the introduction of advanced multimodal large language models, we also evaluate their capabilities in constrained generation tasks such as intent detection. Additionally, we assess their performance beyond traditional applications like object detection and question answering.

**Video-ChatGPT (Maaz et al., 2024)[VChat-

Model	Num Param	Memory
MAG	112.5M	429.20 MB
MISA	140M	534.15 MB
MulT	136M	519.39 MB
BERT	109.5M	417.70 MB
M-7B	7B	15 GB
L2-7B	7B	13 GB
L3-8B	8B	16 GB
L2-13B	13B	26 GB
VLLaMA	15B	30 GB
VLLaVA	11B	22 GB
VChatGPT	11B	22 GB

Table 12: Model Parameter and Memory Requirement Details

	M-1	M-2.0
M-7B	140	180
L2-7B	120	150
L3-8B	150	200
L2-13B	180	230

Table 13: Finetuning-LLM Timings in min after quantization

GPT]: Video-ChatGPT is a novel multimodal model that combines a video-adapted visual encoder with a large language model (LLM). The main idea is to enable detailed video understanding and generation of meaningful conversations about videos. It leverages a pretrained visual encoder like CLIP to extract spatiotemporal video features, which are then integrated with an LLM like Vicuna.

Video-LLaMA (Zhang et al., 2023)[VL-LaMA]: Video-LLaMA enables large language models to simultaneously understand both visual and auditory content in videos. Its unique idea is to have separate vision and audio branches that process the video frames and audio signals in parallel. The vision branch uses techniques like a Video Q-former to extract spatiotemporal visual features, while the audio branch leverages pre-trained audio models like ImageBind and an Audio Q-former to obtain audio representations. These multimodal representations are then aligned to the language model's embedding space through cross-modal pre-training on video/audio captioning data.

Video-LLaVA (Lin et al., 2024)[VLLaVA]: Video-LLaVA employs an encoder-decoder architecture where the encoder maps raw image and video inputs into a unified visual feature space, while the decoder is a large language model like Vicuna. Specifically, it utilizes LanguageBind encoders pretrained on vision-language data to extract visual features from images and videos. These vi-

sual encoders are designed to align the representations of different modalities into a common textual feature space. This allows the image and video features to be mapped into a single unified visual representation.

GPT-4o, Claude-3-Haiku: Closed-source chatbots such as GPT and Claude have shown impressive capabilities to understand multiple modalities. Despite their APIs not directly supporting video input, we utilize their functionality by including 6 uniformly sampled frames from the video along with the text utterance in the input. We test them under two settings: text-only **[ClaudeT, GPT-4T]** and text with 6 frames **[ClaudeV, GPT-4V]**.

9.4 Experimental Setup

All small multimodal models, including SDIF, MAG, MulT, and MISA, are trained with a learning rate of 5e-6 for a maximum of 100 epochs with early termination enabled. On average, these models are trained for 20 epochs, and their results are averaged across 5 seed values. Textual LLMs are trained and evaluated using carefully curated prompts, as shown in Figure 3, for fine-tuning. These prompts are refined through experimentation to determine the most effective format for intent classification. The best-performing prompt is then adapted across different models to ensure consistency in evaluation. For inference, similar prompts are used with the label removed from the response for both fine-tuned and multimodal LLMs, ensuring a uniform testing strategy across all models.. The textual LLMs are fine-tuned using the AdamW optimizer with a learning rate of 2e-4, 4-bit quantization, and are modeled as a text-generation task. Table 12 shows that smaller models like MAG, MISA, MulT, and BERT have fewer parameters (around 100–140 million) and require very little memory (less than 1 GB), which makes them easier to deploy on less powerful systems. In contrast, models with billions of parameters need significantly more memory, up to 30 GB, highlighting the trade-off between model size and the computational resources required.

Tables 13 and 15 summarize the time taken to fine-tune and perform inference on various LLMs. As model size increases, the time required for both fine-tuning and inference also grows significantly. This increase in time results from the higher computational complexity and greater number of parameters in larger models. Additionally, the time taken for inference is influenced by model size and

	BERT	M-7B	L2-7B	Q-7B	L3-8B	L2-13B	MulT	MAG	MISA	SDIF	ClaudeT	ClaudeV	GPT-4T	GPT-4V	VChatGPT
Advise	74.8	84.0	88.0	60.0	**96.0**	80.0	74.4	75.5	72.8	80.0	48.0	52.0	76.0	40.0	12.0
Agree	95.8	**100.0**	91.7	91.7	91.7	**100.0**	96.7	95.0	96.7	100.0	**100.0**	100.0	100.0	83.3	0.0
Apologize	96.3	96.3	**100.0**	96.3	92.6	92.6	96.3	97.3	97.4	96.3	**100.0**	100.0	85.2	96.3	92.6
Arrange	66.8	86.4	81.8	**95.5**	63.6	77.3	71.4	67.6	66.8	68.2	22.7	54.5	31.8	18.2	0.0
Ask Help	64.0	80.0	70.0	70.0	70.0	60.0	63.0	66.7	63.0	50.0	80.0	40.0	**100.0**	90.0	15.8
Care	89.5	**94.7**	89.5	94.7	84.2	89.5	94.7	89.5	89.5	84.2	15.8	31.6	15.8	15.8	27.8
Comfort	78.3	**88.9**	88.9	66.7	83.3	83.3	73.3	81.9	78.9	77.8	44.4	72.2	66.7	77.8	24.6
Complain	61.8	**82.5**	75.4	80.7	75.4	71.9	63.5	65.7	58.6	66.7	64.9	59.6	66.7	57.9	43.5
Criticize	47.8	**78.3**	69.6	73.9	65.2	**78.3**	50.0	51.3	52.2	39.1	69.6	73.9	30.4	69.6	0.0
Flaunt	48.0	50.0	70.0	**80.0**	50.0	70.0	46.0	42.0	54.0	40.0	40.0	40.0	60.0	60.0	58.3
Greet	77.5	91.7	83.3	83.3	91.7	75.0	74.2	85.0	71.7	83.3	**100.0**	100.0	83.3	91.7	24.6
Inform	68.6	78.9	73.7	35.1	75.4	**80.7**	70.2	68.2	72.3	70.2	54.4	47.4	64.9	66.7	14.3
Introduce	68.6	**85.7**	66.7	85.7	71.4	76.2	71.0	71.1	68.1	57.1	14.3	23.8	23.8	19.0	60.0
Joke	26.0	70.0	80.0	20.0	70.0	**90.0**	33.0	32.0	40.0	60.0	30.0	30.0	50.0	50.0	11.8
Leave	72.4	76.5	76.5	58.8	**88.2**	88.2	71.2	75.3	72.4	82.4	41.2	47.1	41.2	47.1	10.0
Oppose	33.0	60.0	**80.0**	50.0	60.0	70.0	32.0	37.3	35.0	40.0	50.0	**80.0**	40.0	60.0	48.8
Praise	84.9	**90.7**	86.0	90.7	81.4	83.7	83.3	87.8	85.8	79.1	72.1	72.1	65.1	62.8	40.0
Prevent	80.0	93.3	66.7	80.0	73.3	**100.0**	79.3	84.0	78.0	80.0	40.0	33.3	73.3	46.7	0.0
Taunt	16.7	8.3	16.7	25.0	16.7	25.0	16.7	11.1	26.7	16.7	16.7	8.3	33.3	41.7	**84.0**
Thank	98.0	**100.0**	100.0	88.0	96.0	**100.0**	97.6	99.5	99.6	96.0	**100.0**	84.0	84.0	92.0	60.0

Table 14: Extended MIntRec-1 Results (Accuracy in %; MulT, MISA, MAG and SDIF are small multimodal models; L2-7B, L2-13B, M-7B, Q-7B and L3-8B are textual LLMs; VChatGPT is Multimodal LLM)(Extended version of Main Table 3)

	M-1	M-2.0
M-7B	80	100
L2-7B	65	100
L3-8B	80	100
L2-13B	90	120
VChatGPT	90	180
VLLaVA	100	240
VLLaMA	100	240

Table 15: Inferencing-LLM Timings in min (First four are after quantization, meanwhile the remaining three are without quantization)

the amount of test data; larger datasets result in longer inference times.

Small-scale multimodal models, on the other hand, are more time-efficient. They take considerably less time for training compared to larger LLMs, making them very resource-efficient and practical for scenarios with limited computational resources or where quick turnaround is essential. This efficiency is particularly advantageous in iterative development environments or real-time applications where rapid model updates and low-latency inference are critical.

All no-training experiments on textual LLMs, including GPT-4T and ClaudeT, are performed by appending 4/5 input samples to the prompt (Fig. 4), which is then appended to every input. For GPT-4V and ClaudeV with visual input, due to cost constraints (Table 11), we uniformly sample 6 images from the video. These images are then cropped down to 1024x1024 resolution, each image translating roughly to 170 tokens.

Different models are employed in varying con-figurations not to establish state-of-the-art performance, but rather to derive deeper insights into the behavior of current systems under real-world constraints. Our study prioritizes understanding the limitations of models and datasets for multimodal intent detection over optimizing performance. While it might appear unfair to compare fine-tuned textual models against non-fine-tuned multimodal models, this decision is grounded in practical considerations such as limited computational resources and the feasibility of training large models on a single GPU. Models like BERT, being lightweight, can be fully trained from scratch. Textual LLMs such as GPT-4T and ClaudeT, which have already undergone extensive pretraining, can be adapted to the task using parameter-efficient fine-tuning methods like LoRA. In contrast, large multimodal LLMs like GPT-4V and ClaudeV are significantly more complex due to the inclusion of pretrained encoders for each modality. These models are commonly used in zero-shot or few-shot settings, as full fine-tuning is not only computationally expensive but also conceptually challenging. Specifically, fine-tuning multimodal models requires careful coordination across different modality-specific encoders to avoid misalignment, which can lead to degraded or inconsistent outputs.

Moreover, the scale, architecture, and training history of these models differ vastly. Attempting to normalize these differences by enforcing a single training or inference strategy would be infeasible under our resource constraints and less informative for our practical goal. Our comparisons aim

	BERT	M-7B	L2-7B	Q-7B	L3-8B	L2-13B	MulT	MAG	MISA	SDIF	ClaudeT	ClaudeV	GPT-4T	GPT-4V	VLLaVA
Acknowledge	72.3	55.9	76.5	8.8	**86.8**	66.2	65.9	71.2	69.4	70.9	76.5	75.0	76.5	76.5	29.4
Advise	58.6	67.0	67.0	69.1	**77.3**	46.4	60.2	61.6	53.6	60.2	41.2	41.2	55.9	53.8	8.2
Agree	63.1	75.0	64.0	71.9	62.5	70.3	66.6	68.7	61.9	64.1	51.6	**79.7**	38.1	34.9	0
Apologize	92.9	97.0	94.0	**100.0**	93.8	96.9	94.5	94.5	93.2	93.2	98.5	98.5	81.5	84.6	81.5
Arrange	67.1	**82.0**	65.0	66.7	60.8	45.1	65.1	67.1	62.3	59.6	14.6	14.6	10.0	10.0	5.9
Ask Help	52.4	70.0	49.0	73.0	21.6	32.4	57.8	51.3	63.8	64.9	83.8	**86.5**	72.9	72.9	2.1
Ask Opinions	56.2	50.0	39.6	**79.2**	27.1	35.4	57.9	60.0	54.2	61.7	50.0	50.0	59.6	59.6	34.7
Care	51.1	**61.7**	40.4	21.3	31.9	27.7	55.3	52.8	51.5	50.2	2.2	2.2	8.5	8.5	10.3
Comfort	42.0	**48.9**	46.9	34.7	38.8	44.9	44.5	45.7	46.5	42.9	29.2	29.2	22.9	25.0	0
Complain	42.2	55.1	51.4	**71.0**	54.2	53.3	45.8	51.2	43.7	51.4	52.4	52.4	57.8	57.8	23.4
Confirm	46.7	44.1	38.7	15.1	**75.3**	44.1	49.9	47.9	48.4	47.7	8.9	8.9	18.3	18.3	7.4
Criticize	40.8	38.3	34.0	40.4	34.0	40.4	40.9	37.4	36.6	44.7	**58.7**	**58.7**	36.4	36.4	17.6
Doubt	62.4	**80.0**	64.4	72.6	60.7	53.3	64.0	64.6	67.6	61.8	20.0	20.0	40.1	40.1	2.1
Emphasize	2.3	23.5	5.9	0.0	11.8	17.6	2.3	1.2	5.9	3.5	23.5	23.5	**31.2**	**31.2**	9.1
Explain	57.2	**76.2**	61.1	68.9	63.7	60.6	63.2	59.3	60.8	61.4	12.4	12.9	17.0	17.0	46.9
Flaunt	9.1	36.4	9.1	18.2	27.3	9.1	15.4	8.2	21.8	20	36.4	36.4	**73.7**	**73.7**	7.4
Greet	82.7	86.4	83.3	80.3	84.8	**90.9**	86.4	85.1	81.8	80.6	83.3	83.3	70.3	70.3	13.3
Inform	60.9	64.1	60.8	68.7	**71.4**	53.9	55.0	56.8	58.9	54.3	45.2	46.7	46.9	46.9	0
Introduce	44.0	41.7	33.3	43.3	40.0	23.3	40.7	**49.7**	42.3	45.0	11.8	11.8	20.7	20.7	41.4
Invite	34.1	52.9	52.9	11.8	52.9	76.5	38.8	24.7	29.4	34.1	76.5	76.5	**82.3**	**82.3**	7.6
Joke	4.8	10.3	17.2	6.9	6.9	3.4	4.1	0	15.2	12.4	28.6	28.6	**48.1**	37.0	0
Leave	45.3	50.9	45.3	37.7	49.1	47.2	50.9	49.8	44.9	**55.1**	30.2	30.2	34.6	34.6	2.6
Oppose	73.7	71.8	69.2	**83.8**	71.8	64.9	73.8	74.9	66.5	69.1	12.9	13.8	9.6	9.6	49.6
Plan	56.3	**71.0**	50.0	60.5	47.4	57.9	46.3	55.3	60.5	53.7	33.3	33.3	19.4	19.4	9.7
Praise	75.9	**85.8**	67.3	84.1	77.9	73.4	77.8	78.7	76.5	75.6	59.5	59.5	56.7	56.7	11.8
Prevent	57.4	64.5	41.9	45.2	**74.2**	64.5	66.4	64.5	58.7	56.1	16.1	16.1	36.7	26.7	20.7
Refuse	14.1	58.8	41.2	0.0	29.4	35.3	12.9	10.6	23.5	22.3	47.1	52.9	52.9	52.9	**84.5**
Taunt	14.5	31.0	18.9	34.5	20.7	**37.9**	25.2	4.8	23.8	23.1	20.7	20.7	37.0	37.0	26.3
Thank	96.5	**100.0**	98.3	94.8	**100.0**	98.3	96.5	96.5	96.2	96.5	94.8	94.8	81.0	81.0	67.6
Warn	13.7	42.1	36.8	42.1	42.1	15.8	15.8	10.5	25.3	32.6	47.4	47.4	**72.2**	**72.2**	31.2

Table 16: Extended Results for MIntRec2.0 (Accuracy in %; MulT, MISA, MAG and SDIF are small multimodal models; L2-7B, L2-13B, M-7B, Q-7B and L3-8B are textual LLMs; VLLaVA is multimodal LLM)(Extended version of Main Table 3)

to examine the relative performance trends within and across model classes rather than absolute performance. Even though fine-tuning may improve certain metrics like accuracy or F1 score, our key findings, particularly those concerning bias, modality dominance, and ablation performance, are based on comparative patterns rather than isolated outcomes. For instance, we observe that text often dominates multimodal fusion and balancing modalities can cause significant drops in performance. These behaviors persist regardless of model size or training strategy, indicating that they are systemic issues rather than artifacts of specific setups. Thus, the fairness of comparing models under different settings does not undermine the validity of our conclusions. Instead, it reflects a realistic evaluation framework for developing robust and scalable intent detection systems in constrained environments.

9.5 Phase-1: Performance Analysis Across Models

9.5.1 Intent-Wise Analysis

In the MIntRec-1 dataset (Table 14), textual LLMs excel in intent labels like *Agree*, *Prevent*, and *Thank*. Small-scale multimodal models perform better on labels like *Care* and *Apologize*, while larger multimodal LLMs excel in *Taunt*, *Praise*,

Greet, and *Ask for help*.

In the MIntRec2.0 dataset (Table 16), textual LLMs perform better on labels such as *Agree*, *Thank*, *Acknowledge*, *Greet*, and *Praise*. Small-scale multimodal models outperform on *Oppose*, *Leave*, *Introduce*, and *Ask for help*, while larger multimodal LLMs excel in *Refuse*, *Invite*, *Ask for help*, and *Complain*. The MIntRec2.0 dataset, with its numerous intent labels, displays less consistent performance trends compared to MIntRec-1. We, therefore, report overall trends wherein most of the models of that type perform better.

Tables 17 and 18 summarize the results of the intent detection task performed on MIntRec-1 and MIntRec2.0, respectively.

9.5.2 Generalizability of Models

Generalizability is a key aspect for intent detection models, as user intents continuously evolve and a fixed model may require frequent retraining. Traditional small multimodal models and BERT-based architectures use a fixed output layer size to predict a set number of probabilities (20 for MIntRec-1 and 30 for MIntRec2.0). In contrast, large language models (LLMs) such as GPT-4o and Claude benefit from extensive pretraining, allowing them to perform well on multiple datasets in zero- or few-shot

24045

	VLLaMA	VLLaVA	VChatGPT
Advise	**16.0**	0.0	12.0
Agree	0.0	0.0	0.0
Apologize	66.7	**92.6**	**92.6**
Arrange	**4.6**	**4.6**	0
Ask Help	10.5	10.5	**15.8**
Care	22.2	**61.1**	27.8
Comfort	21.0	14.0	**24.6**
Complain	21.7	30.4	**43.5**
Criticize	0.0	**10.0**	0.0
Flaunt	50.0	**66.7**	58.3
Greet	17.5	7.0	**24.6**
Inform	**14.3**	**14.3**	**14.3**
Introduce	10.0	20.0	**60.0**
Joke	0.0	**11.8**	**11.8**
Leave	0.0	0.0	**10.0**
Oppose	30.2	**67.4**	48.8
Praise	0.0	26.7	**40.0**
Prevent	0.0	**16.7**	0.0
Taunt	68.0	**100.0**	84.0
Thank	90.0	**100.0**	60.0

Table 17: MultiModal LLM Results on MIntRec-1 (Accuracy in %)(Extended version of Main Table 3)

	VLLaMA	VLLaVA	VChatGPT
Accuracy	11.8	**18.8**	18.7
Acknowledge	22.1	**29.4**	13.2
Advise	4.1	**8.2**	**8.2**
Agree	0.0	0.0	0.0
Apologize	70.8	81.5	**95.4**
Arrange	3.9	**5.9**	0.0
Ask Help	0.0	**2.1**	0.0
Ask Opinions	14.3	**34.7**	4.3
Care	7.5	10.3	**42.9**
Comfort	1.1	0.0	**17.8**
Complain	12.77	**23.4**	0
Confirm	6.7	7.4	**23.4**
Criticize	11.8	**17.6**	4.4
Doubt	3.1	2.1	**11.8**
Emphasize	0.0	**9.1**	0.5
Explain	39.4	**46.9**	9.1
flaunt	3.7	7.4	**42.4**
Greet	5.0	**13.3**	12.9
Inform	5.9	0.0	**6.7**
Introduce	3.4	**41.5**	0.0
Invite	3.8	7.5	**62.1**
Joke	0.8	0.0	**24.5**
Leave	5.3	2.6	**12.8**
Oppose	15.0	**49.6**	2.6
Plan	6.4	9.7	**38.9**
Praise	0.0	11.8	**25.8**
Prevent	1.7	**20.7**	11.8
Refuse	**84.5**	**84.5**	0.0
Taunt	5.3	26.3	**82.8**
Thank	43.2	**67.6**	21.0
Warn	8.3	31.2	**67.6**

Table 18: MultiModal LLM Results on MIntRec2.0 (Accuracy in %)(Extended version of Main Table 3)

settings without additional training.

We evaluate the generalizability of a textual LLM by selecting Mistral-7B for its strong performance. Mistral-7B, fine-tuned on the MIntRec2.0 dataset, is tested on the MIntRec-1 dataset by adjusting the class labels in the prompt. The system achieves an Overall Accuracy of 72.58% and an Overall F1 Score of 72.85%. Given that MIntRec2.0 already contains most MIntRec-1 labels, this result was expected despite a slight decline caused by differences between the two datasets. MIntRec2.0 includes labels such as *Oppose*, *Asking for opinions* that are absent in MIntRec-1, leading to some misclassifications.

In another experiment, the model fine-tuned on MIntRec-1 with 20 labels is tested on MIntRec2.0 with 30 labels, resulting in an Overall Accuracy of 47.47% and an Overall F1 Score of 40.42%. This performance remains competitive compared to several closed-source and open-source models in zero- or few-shot settings.

In summary, while parameter-efficient fine-tuning can limit generalizability, the models still demonstrate acceptable performance across different datasets.

9.5.3 Performance with limited data analysis

Table 19 compares model performance under zero-shot, 10-shot, and full-shot settings. Zero-shot performance is weak across all models, while few-shot learning significantly improves results. Tables 20 and 22 display the performance of these models

under zero-shot settings, while Tables 21 and 23 present their performance under 10-shot settings.

	M-1			M-2.0		
	0-shot	10-shot	Full-shot	0-shot	10-shot	Full-shot
BERT	5.0	57.2	70.9	4.3	40.9	57.1
M-7B	4.7	67.4	**82.9**	3.9	**45.4**	**65.2**
L2-7B	5.4	28.8	79.3	2.9	39.1	57.2
L3-8B	4.9	67.2	77.3	0.3	42.4	61.3
L2-13B	4.9	66.5	80.7	**10.1**	41.1	54.8
MulT	6.5	**68.8**	71.5	4.1	36.9	58.4
MAG	4.9	56.5	72.7	3.6	39.2	58.2
MISA	4.8	53.4	71.8	3.3	37.6	57.8
SDIF	**12.8**	52.4	72.8	6.6	38.6	58.2

Table 19: Zero/Few-shot Results

Notably, MulT surpasses M-7B on MIntRec-1 in the 10-shot setting, likely due to its cross-modal attention mechanism and compact architecture, which facilitates learning patterns more efficiently than large language models (LLMs) that require a sufficient number of samples for improved performance. However, M-7B achieves the best overall performance, emphasizing the critical role of adequate training data. Since we use pre-trained variants of the models, rather than chat-finetuned variants, the zero-shot experiments fail to accurately predict several intents. However, as the amount of training data increases, the performance

Model	M-7B	L2-7B	L3-8B	L2-13B
Accuracy	4.7	**5.4**	4.9	4.9
Advise	8.0	**12.0**	0.0	0.0
Agree	**8.3**	0.0	0.0	0.0
Apologize	**3.7**	0.0	0.0	0.0
Arrange	9.1	90.9	**100.0**	**100.0**
Ask Help	**10.0**	0.0	0.0	0.0
Care	0.0	0.0	0.0	0.0
Comfort	**5.6**	0.0	0.0	0.0
Complain	**3.5**	0.0	0.0	0.0
Criticize	0.0	0.0	0.0	0.0
Flaunt	0.0	0.0	0.0	0.0
Greet	**16.7**	0.0	0.0	0.0
Inform	**3.5**	1.7	0.0	0.0
Introduce	**9.5**	0.0	0.0	0.0
Joke	0.0	0.0	0.0	0.0
Leave	0.0	0.0	0.0	0.0
Oppose	**10.0**	0.0	0.0	0.0
Praise	**4.6**	0.0	0.0	0.0
Prevent	0.0	0.0	0.0	0.0
Taunt	**8.3**	0.0	0.0	0.0
Thank	**4.0**	0.0	0.0	0.0

Table 20: Zero Shot results MIntRec-1 (Extended version of Table 19)

Epoch	M-7B	L2-7B	L3-8B	L2-13B
Accuracy	67.2	28.7	65.4	66.5
Advise	**88.0**	80.0	88.0	68.0
Agree	**100.0**	33.3	100.0	**100.0**
Apologize	**88.9**	85.2	88.9	81.5
Arrange	**68.2**	13.6	54.5	45.4
Ask Help	70.0	50.0	70.0	**80.0**
Care	52.6	15.8	52.6	78.9
Comfort	**77.8**	33.3	66.7	72.2
Complain	42.1	7.0	42.1	**68.4**
Criticize	**69.6**	21.7	60.9	65.2
Flaunt	**90.0**	0.0	90.0	40.0
Greet	**91.7**	33.3	91.7	75.0
Inform	**52.6**	10.5	54.4	19.3
Introduce	**85.7**	19.0	85.7	**85.7**
Joke	60.0	10.0	60.0	**70.0**
Leave	52.9	52.9	52.9	**82.3**
Oppose	70.0	40.0	70.0	**80.0**
Praise	65.1	18.6	65.1	**79.1**
Prevent	66.7	26.7	53.3	**73.3**
Taunt	25.0	0.0	25.0	**33.3**
Thank	96.0	60.0	96.0	**100.0**

Table 21: 10-shot Results on MIntRec-1 (Extended version of Table 19)

of the LLMs improves significantly, highlighting their sensitivity to the quantity of training data. Notably, even under 10-shot settings, the LLMs perform remarkably well, nearly achieving the performance levels seen with full fine-tuning. These trends are consistent across both the MIntRec-1 and MIntRec2.0 datasets.

9.6 Phase-2: Analysis of Modality Importance

9.6.1 Wordcloud analysis

All textual LLMs, such as LLaMA-2-7B, LLaMA-2-13B, LLaMA-3-8B, and Mistral-7B, perform consistently well on labels like *Agree*, *Apologize*, and *Thank*. To understand this, we analyze word clouds for these labels (Fig. 5).

For instance, the names 'Judy,' 'Amy,' and 'Cheyenne' appear in the dataset, with 'Amy' occurring 14 times in the test set across various intents: 'Inform' (4), 'Greet' (3), 'Apologize' (2), 'Arrange' (1), 'Thank' (1), 'Care' (1), 'Comfort' (1), and 'Complain' (1). The worst-performing LLM model on MIntRec-1, LLaMA-3-8B, correctly classifies 10 out of these 14 instances. Word clouds highlight phrases such as 'sorry' and 'apologize,' which contribute to accurate intent classification. This confirms that the models categorize intents based on content rather than specific character names, ensuring the dataset does not introduce character-based bias.

	M-7B	L2-7B	L3-8B	L2-13B
Accuracy	3.9	2.9	0.3	**10.1**
Acknowledge	1.5	4.4	**76.5**	1.5
Advise	3.1	1.0	0.0	**12.4**
Agree	0.0	**3.1**	1.6	1.6
Apologize	4.6	1.5	3.1	**6.1**
Arrange	**3.9**	**3.9**	0.0	0.0
Ask Help	0.0	0.0	0.0	0.0
Ask Opinions	0.0	0.0	0.0	0.0
Care	**2.1**	0.0	**2.1**	0.0
Comfort	6.1	0.0	0.0	**8.2**
Complain	0.9	0.0	0.9	**5.6**
Confirm	**3.2**	0.0	0.0	2.1
Criticize	0.0	0.0	0.0	**2.1**
Doubt	2.2	0.7	0.0	**5.2**
Emphasize	5.9	0.0	0.0	**11.8**
Explain	1.0	0.0	0.0	**4.7**
Flaunt	**4.5**	0.0	0.0	**4.5**
Greet	**7.6**	1.5	4.5	1.5
Inform	4.6	1.4	0.5	**38.7**
Introduce	40.0	**66.7**	3.3	38.3
Invite	0.0	0.0	0.0	**5.9**
Joke	**6.9**	0.0	0.0	3.4
Leave	3.8	0.0	1.9	**26.4**
Oppose	**4.3**	1.7	1.7	**3.4**
Plan	**2.6**	0.0	0.0	**2.6**
Praise	**4.4**	0.0	0.9	**4.4**
Prevent	0.0	3.2	0.0	**6.4**
Refuse	**5.9**	0.0	0.0	0.0
Taunt	1.7	1.7	0.0	**18.9**
Thank	0.0	0.0	0.0	0.0
Warn	5.3	5.3	0.0	**42.1**

Table 22: Zero-shot Results on MIntRec2.0 (Extended version of Table 19)

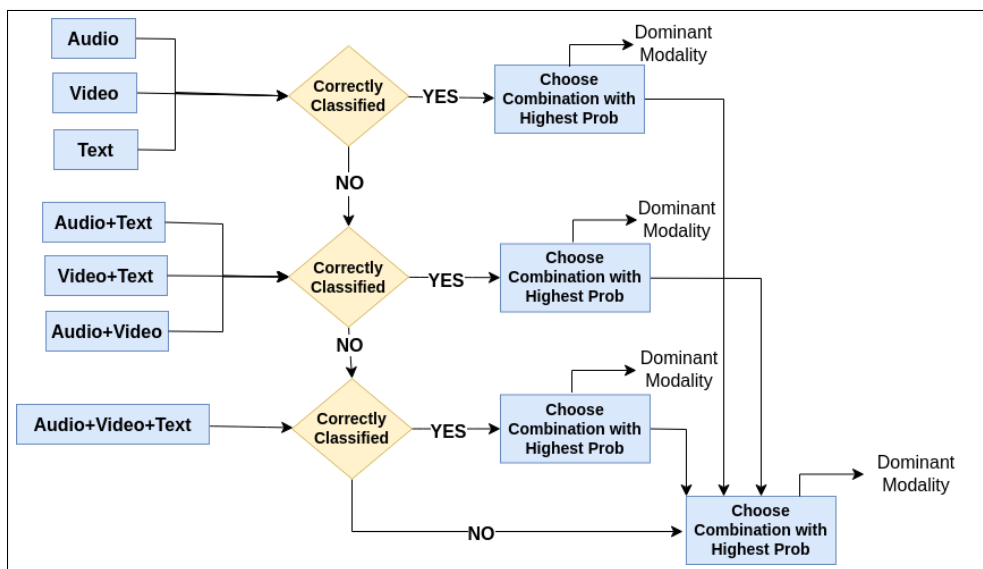Figure 5: WordCloud of Agree, Apologize, and Thank labels
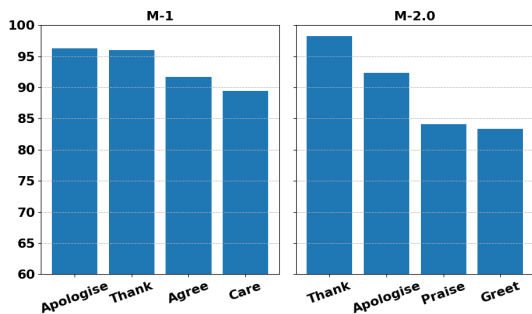

Figure 6: Modality Annotation Algorithm

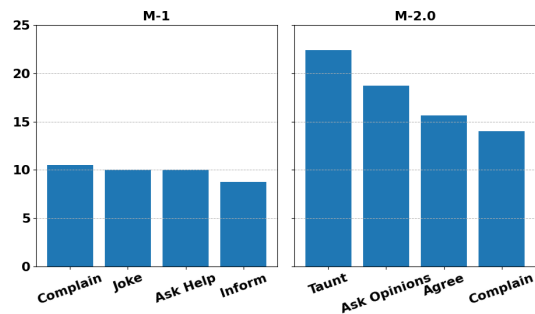Figure 7: Textually Biased Labels

Figure 8: Visually Biased Labels

	M-7B	L2-7B	L3-8B	L2-13B
Accuracy	**45.4**	39.1	29.4	14.1
Acknowledge	89.7	58.8	**92.6**	16.2
Advise	**56.7**	49.5	23.7	9.3
Agree	23.4	**43.7**	6.2	18.7
Apologize	**83.1**	67.7	15.4	41.5
Arrange	**45.1**	43.1	33.3	11.7
Ask Help	0.0	**21.6**	13.5	0.0
Ask Opinions	0.0	10.4	**12.5**	4.2
Care	40.4	**44.7**	17.0	14.9
Comfort	42.9	**46.9**	22.4	2.0
Complain	**33.6**	29.9	14.0	14.0
Confirm	**41.9**	23.7	8.6	11.8
Criticize	**61.7**	55.3	10.6	12.8
Doubt	**51.1**	23.7	20.7	2.9
Emphasize	29.4	**58.8**	0.0	5.9
Explain	**52.8**	32.1	38.9	6.2
Flaunt	**45.4**	**45.4**	27.3	9.1
Greet	66.7	77.3	**84.8**	33.3
Inform	**33.6**	31.3	12.9	5.1
Introduce	**35.0**	23.3	23.3	10.0
Invite	47.1	**52.9**	23.5	29.4
Joke	13.8	**27.6**	13.8	10.3
Leave	39.6	49.1	**58.5**	28.3
Oppose	15.4	**17.1**	14.5	12.8
Plan	**81.6**	39.5	47.4	15.8
Praise	**56.6**	49.6	32.7	9.7
Prevent	48.4	**64.5**	**64.5**	25.8
Refuse	**58.8**	41.2	47.1	17.6
Taunt	**18.9**	5.2	13.8	17.2
Thank	93.1	94.8	**98.3**	68.9
Warn	57.9	52.6	**63.2**	26.3

Table 23: 10-Shot Results on MIntRec2.0 (Extended version of Table 19)

	T+V+A	T	V	A	T+V	T+A	V+A
Acc	**72.8**	69.9	17.8	28.5	69.7	70.8	25.2
Advise	**80.0**	60.0	0.0	0.0	50.0	50.0	0.0
Agree	**100.0**	63.2	61.4	40.4	59.7	61.4	49.1
Apologize	**96.3**	81.4	11.6	48.8	79.1	86.1	41.9
Arrange	68.2	**96.3**	7.4	63.0	**96.3**	**96.3**	55.6
Ask Help	50.0	**100.0**	0.0	64.0	**100.0**	**100.0**	56.0
Care	**84.2**	39.1	0.0	13.0	43.5	52.2	4.4
Comfort	**77.8**	25.0	0.0	0.0	25.0	16.7	8.3
Complain	66.7	91.7	0.0	8.3	91.7	**100.0**	8.3
Criticize	39.1	30.0	0.0	0.0	30.0	**50.0**	0.0
Flaunt	**60.0**	**60.0**	0.0	0.0	30.0	30.0	0.0
Greet	**83.3**	40.0	0.0	10.0	30.0	30.0	10.0
Inform	70.2	72.2	0.0	16.7	**77.8**	**77.8**	5.6
Introduce	57.1	89.5	0.0	5.3	**94.7**	89.5	5.3
Joke	60.0	61.4	61.4	40.4	**73.7**	68.4	28.1
Leave	**82.4**	76.0	0.0	16.0	76.0	72.0	8.0
Oppose	40.0	**72.7**	0.0	9.1	68.2	**72.7**	9.1
Praise	**79.1**	66.7	0.0	4.8	61.9	76.2	9.5
Prevent	80.0	**82.4**	0.0	11.8	70.6	64.7	5.9
Taunt	16.7	**73.3**	0.0	20.0	**73.3**	66.7	20.0
Thank	**96.0**	66.7	16.7	50.0	75.0	75.0	41.7

Table 24: Ablation analysis for MIntRec-1. SDIF is the best-performing small multimodal model (Extended version of Table 5)

	T+V+A	T	V	A	T+V	T+A	V+A
Acc	**58.4**	56.7	12.4	23.3	56.4	56.4	24.6
Acknowledge	65.9	69.1	0.0	61.8	**70.6**	**70.6**	48.5
Advise	**60.2**	50.5	2.1	1.0	53.6	53.6	15.5
Agree	**66.6**	65.6	0.0	28.1	57.8	57.8	34.4
Apologize	**94.5**	90.8	0.0	53.9	92.3	92.3	58.5
Arrange	**65.1**	64.7	2.0	0.0	60.8	60.8	7.8
Ask Help	57.8	62.2	0.0	0.0	**64.9**	**64.9**	2.7
Ask Opinions	**57.9**	50.0	0.0	4.2	47.9	47.9	10.4
Care	55.3	**57.5**	0.0	2.1	46.8	46.8	6.4
Comfort	44.5	**49.0**	2.0	0.0	44.9	44.9	0.0
Complain	45.8	43.9	7.5	0.0	38.3	38.3	11.2
Confirm	**49.9**	46.2	0.0	15.1	45.2	45.2	4.3
Criticize	40.9	38.3	0.0	0.0	**48.9**	**48.9**	4.3
Doubt	64.0	62.2	49.6	26.7	**68.2**	**68.2**	33.3
Emphasize	2.3	**11.8**	0.0	0.0	5.9	5.9	0.0
Explain	**63.2**	58.0	16.6	31.6	62.2	62.2	40.4
Flaunt	15.4	13.6	0.0	0.0	**22.7**	**22.7**	0.0
Greet	**86.4**	80.3	12.1	31.8	80.3	80.3	42.4
Inform	55.0	55.3	58.1	**62.2**	56.2	56.2	37.8
Introduce	**40.7**	38.3	0.0	1.7	38.3	38.3	10.0
Invite	**38.8**	29.4	0.0	0.0	17.7	17.7	0.0
Joke	4.1	6.9	0.0	0.0	**13.8**	**13.8**	0.0
Leave	50.9	56.6	0.0	11.3	**56.6**	**56.6**	24.5
Oppose	73.8	67.5	3.4	52.1	70.1	70.1	41.0
Plan	46.3	**57.9**	0.0	0.0	44.7	44.7	10.6
Praise	**77.8**	72.6	3.5	9.7	67.3	67.3	18.6
Prevent	**66.4**	61.3	0.0	0.0	64.5	64.5	3.2
Refuse	12.9	**29.4**	0.0	0.0	29.4	29.4	0.0
Taunt	25.2	**25.9**	0.0	0.0	13.8	13.8	8.6
Thank	96.5	**96.6**	0.0	48.3	96.6	96.6	53.5
Warn	15.8	**26.3**	0.0	0.0	21.1	21.1	0.0

Table 25: Ablation analysis for MIntRec2.0. MulT is the best-performing small multimodal model on MIntRec2.0
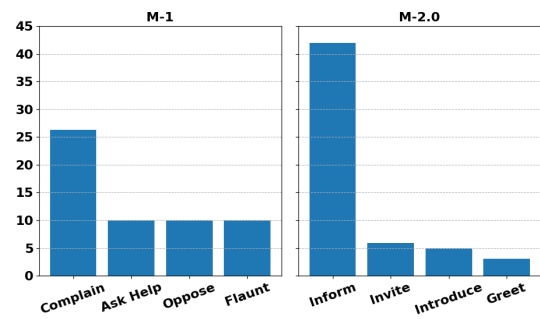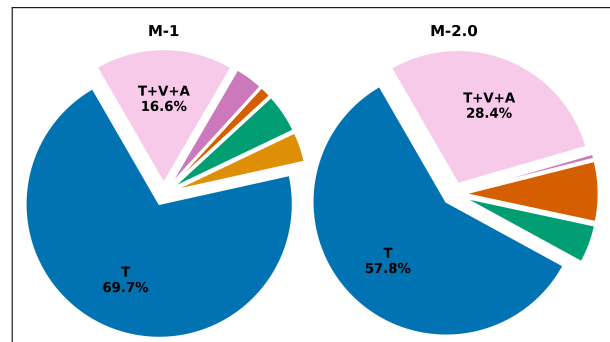

Figure 9: Labels that benefit from audio cues

Figure 10: Distribution of Modality Importance
■ **Text,** ■ **Audio,** ■ **Video,** ■ **Text+Video,** ■ **Text+Audio,** ■ **Audio+Video,** ■ **Text+Video+Audio**

9.6.2 Modality Ablation Analysis

The importance of individual modalities for different intent categories is summarized in Tables 24 and 25. Figures 7 illustrate that over 80% of the samples for intents such as *Apologize*, *Thank*, and *Praise* exhibit a strong textual bias. On the other hand, only a small fraction of samples for intents like *Joke*, *Taunt*, and *Asking for opinions* require the video modality for correct classification (Figure 8. Similarly *Inform*, *Complain* might benefit from the presence of audio cues (Fig. 9).

9.6.3 Modality Combination Annotation

Fig. 6 shows the algorithm used to label MIntRec-1 and MIntRec2.0 with the smallest subset of modalities needed to classify each instance correctly.

9.7 Phase-3: Debiasing MIntRec Datasets

9.7.1 Debiasing Analysis

Tables 26 and 27 present the number of samples before and after debiasing across various intent levels. Tables 28 and 29 summarize the performance of various models following the debiasing process. Fig. 12 shows reduced performance for both small multimodal models and textual LLMs on the debiased datasets. The results before vs. after the debiasing are summarized in Tables 30 and 31.

Modality Ablation on Debiased Dataset We also perform a modality ablation analysis on the debiased dataset (Tables 34 and 35), following the methodology described in Section 9.6.2. The results reveal that the performance gap between text-inclusive combinations and combinations without text is reduced, indicating that the textual bias has been mitigated. Notably, several intent categories, such as *Care*, *Arrange*, and *Agree*, which previously benefited from text-based combinations, now exhibit diminished performance.

During our analysis, we discover that words such as *Right*, *Yeah*, *Okay*, and *Definitely* are strong indicators of agreement, while words like *Thank*, *Appreciate*, and *Thanking* clearly signify gratitude. These words are highly representative of their respective intent categories. Their presence in the text provides unambiguous signals that assist textual LLMs in accurately classifying these labels. This is because the LLMs leverage the direct association of these words with specific intents, reducing the need for additional contextual information from other modalities such as visual or auditory cues. Consequently, even in the absence of multimodal data, textual LLMs efficiently and accurately

Intent	Before	After	% Reduction
Thank	124	2	98.39
Apologize	136	4	97.06
Greet	60	6	90.00
Agree	59	7	88.14
Praise	213	26	87.79
Care	95	13	86.32
Comfort	88	18	79.55
Advise	122	26	78.69
Complain	286	66	76.92
Prevent	73	17	76.71
Leave	85	24	71.76
Inform	284	90	68.31
Arrange	110	35	68.18
Introduce	105	34	67.62
Oppose	51	18	64.71
Criticize	117	47	59.83
Ask Help	51	21	58.82
Flaunt	52	22	57.69
Joke	51	37	27.45
Taunt	62	50	19.35
Total	2224	563	74.69

Table 26: MIntRec-1 Debiasing Statistics

Intent	Before	After	% Reduction
Thank	281	14	95.02
Apologize	271	23	91.51
Greet	313	47	84.98
Praise	472	109	76.91
Oppose	508	183	63.98
Advise	375	145	61.33
Agree	343	140	59.18
Explain	765	315	58.82
Doubt	687	298	56.62
Comfort	233	104	55.36
Arrange	259	116	55.21
Confirm	471	218	53.72
Inform	926	433	53.24
Prevent	128	60	53.13
Plan	188	91	51.60
Introduce	310	161	48.06
Care	222	117	47.30
Complain	512	272	46.88
Acknowledge	307	170	44.63
Ask Help	155	87	43.87
Leave	235	134	42.98
Criticize	198	132	33.33
Invite	109	74	32.11
Ask Opinions	251	174	30.68
Warn	99	70	29.29
Refuse	94	73	22.34
Taunt	274	228	16.79
Flaunt	93	80	13.98
Emphasize	94	86	8.51
Joke	131	122	6.87
Total	9304	4276	54.04

Table 27: MIntRec2.0 Debiasing Statistics
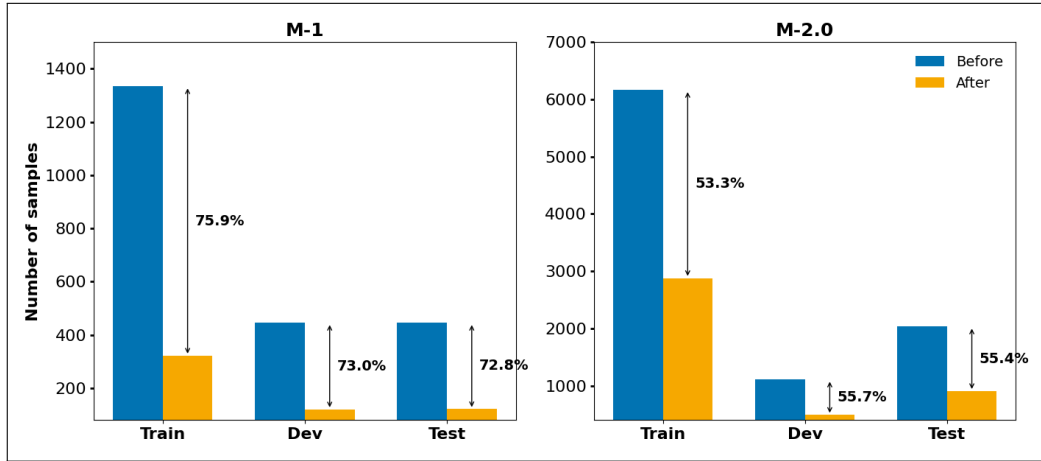
Figure 11: Before vs After Statistics of MIntRec Debiasing

	BERT	M-7B	L-7B	Q-7B	L-8B	L-13B	MulT	MAG	MISA	SDIF	ClaudeT	ClaudeV	GPT-4T	GPT-4V	VChatGPT
Advise	16.7	16.7	33.3	**66.7**	50.0	33.3	33.3	33.3	0.0	33.3	33.3	16.7	14.3	0.0	0.0
Arrange	0.0	0.0	14.3	**42.9**	28.6	14.3	28.6	0.0	0.0	14.3	14.3	28.6	20.0	14.3	0.0
Ask Help	0.0	66.7	66.7	66.7	33.3	33.3	33.3	33.3	0.0	0.0	**100.0**	33.3	25.0	33.3	0.0
Comfort	0.0	0.0	0.0	50.0	25.0	0.0	25.0	0.0	0.0	0.0	0.0	0.0	28.6	**42.9**	0.0
Complain	30.0	**75.0**	20.0	40.0	20.0	60.0	40.0	15.0	20.0	35.0	45.0	45.0	33.3	50.0	50.0
Criticize	30.0	40.0	20.0	40.0	30.0	**70.0**	30.0	40.0	40.0	20.0	60.0	50.0	33.3	50.0	22.2
Flaunt	0.0	0.0	**50.0**	16.7	33.3	33.3	16.7	0.0	0.0	16.7	**50.0**	33.3	33.3	**50.0**	40.0
Inform	31.3	68.8	18.8	18.8	12.5	43.8	50.0	43.8	**81.3**	37.5	31.3	50.0	43.8	55.0	0.0
Introduce	14.3	0.0	57.1	42.9	42.9	28.6	28.6	14.3	0.0	28.6	0.0	0.0	50.0	**60.0**	0.0
Joke	33.3	33.3	83.3	66.7	50.0	50.0	33.3	16.7	16.7	33.3	50.0	16.7	50.0	**62.5**	16.7
Leave	0.0	0.0	25.0	25.0	0.0	50.0	0.0	0.0	0.0	0.0	0.0	0.0	60.0	**66.7**	0.0
Oppose	0.0	0.0	50.0	0.0	50.0	16.7	0.0	0.0	0.0	0.0	33.3	33.3	**66.7**	**66.7**	40.0
Prevent	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	66.7	0.0	**80.0**	66.7	0.0
Taunt	0.0	0.0	30.0	30.0	0.0	0.0	0.0	20.0	20.0	10.0	10.0	10.0	**100.0**	**100.0**	14.3

Table 28: Extended Debiased MIntRec-1 Results (Accuracy in %; MulT, MISA, MAG and SDIF are small multimodal models; L2-7B, L2-13B, M-7B, Q-7B and L3-8B are textual LLMs; VChatGPT is a multimodal LLM)(Extended version of Table 7)

	BERT	M-7B	L-7B	Q-7B	L-8B	L-13B	MulT	MAG	MISA	SDIF	ClaudeT	ClaudeV	GPT-4T	GPT-4V	VChatGPT
Acknowledge	41.5	78.1	56.1	78.1	41.5	61.0	63.4	41.5	63.4	48.8	73.2	73.2	0.0	0.0	**81.6**
Advise	13.3	6.7	42.2	**55.6**	26.7	33.3	15.6	15.6	26.7	4.4	26.7	26.7	0.0	3.9	0.0
Agree	26.9	26.9	23.1	26.9	23.1	23.1	11.5	38.5	0.0	15.4	34.6	**65.4**	2.6	4.8	0.0
Arrange	5.9	5.9	5.9	11.8	0.0	**23.5**	5.9	17.7	17.7	5.9	6.3	6.3	5.9	5.1	0.0
Ask Help	9.1	27.3	36.4	36.4	36.4	9.1	9.1	9.1	27.3	0.0	54.6	**63.6**	6.1	8.3	0.0
Ask Opinions	28.0	36.0	28.0	**72.0**	60.0	28.0	16.0	40.0	28.0	40.0	36.0	36.0	7.7	12.0	0.0
Care	11.5	7.7	23.1	7.7	26.9	26.9	19.2	11.5	19.2	23.1	4.0	4.0	9.4	12.2	0.0
Comfort	4.0	**36.0**	24.0	4.0	**36.0**	**36.0**	8.0	0.0	20.0	8.0	4.2	4.2	11.1	22.2	9.1
Complain	11.9	32.2	**57.6**	35.6	44.1	42.4	8.5	17.0	18.6	22.0	39.0	39.0	11.5	22.2	34.6
Confirm	28.6	2.0	22.5	14.3	38.8	24.5	16.3	**44.9**	34.7	24.5	8.7	8.7	12.0	23.1	3.2
Criticize	20.6	2.9	41.2	17.7	50.0	44.1	29.4	14.7	14.7	23.5	**57.6**	**57.6**	16.5	23.5	23.3
Doubt	17.7	**64.7**	49.0	21.6	31.4	51.0	54.9	23.5	19.6	19.6	21.6	21.6	20.0	25.0	55.6
Emphasize	6.7	13.3	6.7	6.7	20.0	**33.3**	6.7	6.7	0.0	0.0	20.0	20.0	23.7	26.7	7.1
Explain	24.7	1.2	34.1	**45.9**	35.3	30.6	32.9	22.4	37.7	18.8	8.6	8.6	23.8	29.4	1.5
Flaunt	0.0	38.9	27.8	**44.4**	5.6	33.3	11.1	0.0	5.6	11.1	27.8	27.8	25.5	30.6	0.0
Inform	31.3	**84.4**	29.2	27.1	37.5	34.4	35.4	19.8	17.7	34.4	29.4	30.4	33.3	31.3	0.0
Introduce	11.1	13.9	25.0	19.4	22.2	27.8	22.2	22.2	8.3	11.1	5.7	5.7	**36.4**	34.8	6.9
Invite	16.7	**75.0**	25.0	41.7	41.7	33.3	25.0	16.7	33.3	25.0	66.7	66.7	37.0	36.6	0.0
Joke	12.0	28.0	12.0	24.0	4.0	8.0	4.0	4.0	12.0	8.0	29.2	29.2	40.0	**44.0**	0.0
Leave	22.2	25.9	14.8	0.0	14.8	11.1	14.8	29.6	14.8	18.5	0.0	0.0	44.0	**46.2**	0.0
Oppose	20.5	23.1	12.8	5.1	23.1	12.8	15.4	28.2	25.6	15.4	5.1	5.1	46.3	**50.0**	2.9
Plan	14.3	47.6	9.5	19.1	38.1	19.1	4.8	19.1	9.5	9.5	25.0	25.0	48.0	**55.6**	7.1
Praise	4.2	45.8	41.7	20.8	41.7	50.0	4.2	4.2	12.5	8.3	17.4	17.4	50.0	**56.0**	17.4
Prevent	18.2	27.3	0.0	18.2	18.2	9.1	36.4	0.0	9.1	0.0	0.0	0.0	53.9	**59.3**	0.0
Refuse	0.0	21.4	28.6	**71.4**	35.7	21.4	7.1	0.0	0.0	7.1	35.7	42.9	58.8	64.3	9.1
Taunt	8.7	43.5	8.7	54.4	23.9	23.9	17.4	28.3	15.2	17.4	17.4	17.4	66.7	**72.7**	7.7
Warn	0.0	30.8	38.5	23.1	46.2	30.8	0.0	0.0	0.0	7.7	30.8	30.8	71.4	**91.7**	0.0

Table 29: Extended Debiased MIntRec2.0 Results (Accuracy in %; MulT, MISA, MAG and SDIF are small multimodal models; L2-7B, L2-13B, M-7B, Q-7B and L3-8B are textual LLMs; VChatGPT is a multimodal LLM)(Extended version of Table 7)

Model	Apologize		Thank		Flaunt		Advise		Prevent		Complain		Taunt	
	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After
BERT	96.3		98.0		48.0	0.0	74.8	16.7	80.0	0.0	61.8	30.0	14.5	0.0
M-7B	96.3		**100.0**		50.0	0.0	84.0	16.7	93.3	0.0	**82.5**	**75.0**	31.0	0.0
L2-7B	**100.0**		**100.0**		**70.0**	**50.0**	88.0	33.3	66.7	0.0	75.4	20.0	18.9	30.0
L3-8B	92.6		96.0		50.0	33.3	**96.0**	**50.0**	73.3	0.0	75.4	20.0	20.7	0.0
L2-13B	92.6		**100.0**		**70.0**	33.3	80.0	33.3	**100.0**	0.0	71.9	60.0	**37.9**	0.0
MulT	96.3		97.6		46.0	16.7	74.4	33.3	79.3	0.0	63.5	40.0	25.2	0.0
MISA	97.3	Removed	99.5	Removed	42.0	0.0	75.5	0.0	84.0	0.0	65.7	20.0	4.8	20.0
MAG	97.4		99.6		54.0	0.0	72.8	33.3	78.0	0.0	58.6	15.0	23.8	20.0
SDIF	96.3		96.0		60.0	16.7	80.0	33.3	80.0	0.0	66.7	35.0	23.1	10.0
ClaudeT	**100.0**		**100.0**		40.0	**50.0**	48.0	33.3	40.0	66.7	64.9	45.0	20.7	10.0
ClaudeV	**100.0**		84.0		40.0	33.3	52.0	16.7	33.3	0.0	59.6	45.0	20.7	10.0
GPT-4T	85.2		84.0		60.0	33.3	76.0	14.3	73.3	**80.0**	66.7	33.3	37.0	**100.0**
GPT-4V	96.3		92.0		60.0	**50.0**	40.0	0.0	46.7	66.7	57.9	50.0	37.0	**100.0**
VChatGPT	92.6		60.0		58.3	40.0	12.0	0.0	0.0	0.0	43.5	50.0	26.3	14.3

Table 30: Performance of models before and after debiasing on MIntRec-1. Pre-debiasing results for all intent labels are reported in 14. Extended debiasing results are reported in Table 28

Model	Apologize		Thank		Joke		Ask Help		Agree		Warn		Invite	
	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After
BERT	92.9		96.5		4.8	12.0	52.4	9.1	63.1	26.9	13.7	0.0	34.1	16.7
M-7B	97.0		**100.0**		10.3	28.0	70.0	27.3	75.0	26.9	42.1	30.8	52.9	75.0
L2-7B	94.0		98.3		17.2	12.0	49.0	36.4	64.0	23.1	36.8	38.5	52.9	25.0
L3-8B	93.8		**100.0**		6.9	4.0	21.6	36.4	62.5	23.1	15.8	46.2	52.9	41.7
L2-13B	96.9		98.3		3.4	8.0	32.4	9.1	70.3	23.1	15.8	30.8	76.5	33.3
MulT	94.5		96.5		4.1	4.0	57.8	9.1	66.6	11.5	15.8	0.0	38.8	25.0
MISA	94.5	Removed	96.5	Removed	0.0	4.0	51.3	9.1	68.7	38.5	10.5	0.0	24.7	16.7
MAG	93.2		96.2		15.2	12.0	63.8	27.3	61.9	0.0	25.3	0.0	29.4	33.3
SDIF	93.2		96.5		12.4	8.0	64.9	0.0	64.1	15.4	32.6	7.7	34.1	25.0
ClaudeT	**98.5**		94.8		28.6	29.2	83.8	54.6	51.6	34.6	47.4	30.8	76.5	**66.7**
ClaudeV	**98.5**		94.8		28.6	29.2	**86.5**	63.6	**79.7**	65.4	47.4	30.8	76.5	**66.7**
GPT-4T	81.5		81.0		**48.1**	40.0	72.9	6.1	38.1	2.6	**72.2**	71.4	**82.3**	37.0
GPT-4V	84.6		81.0		37.0	**44.0**	72.9	8.3	34.9	4.8	**72.2**	**91.7**	**82.3**	36.6
VChatGPT	81.5		67.6		0.0	0.0	2.1	0.0	0.0	0.0	31.2	0.0	7.6	0.0

Table 31: Performance of models before and after debiasing on MIntRec2.0. Pre-debiasing results for all intent labels are reported in 16. Extended debiasing results are reported in Table 29
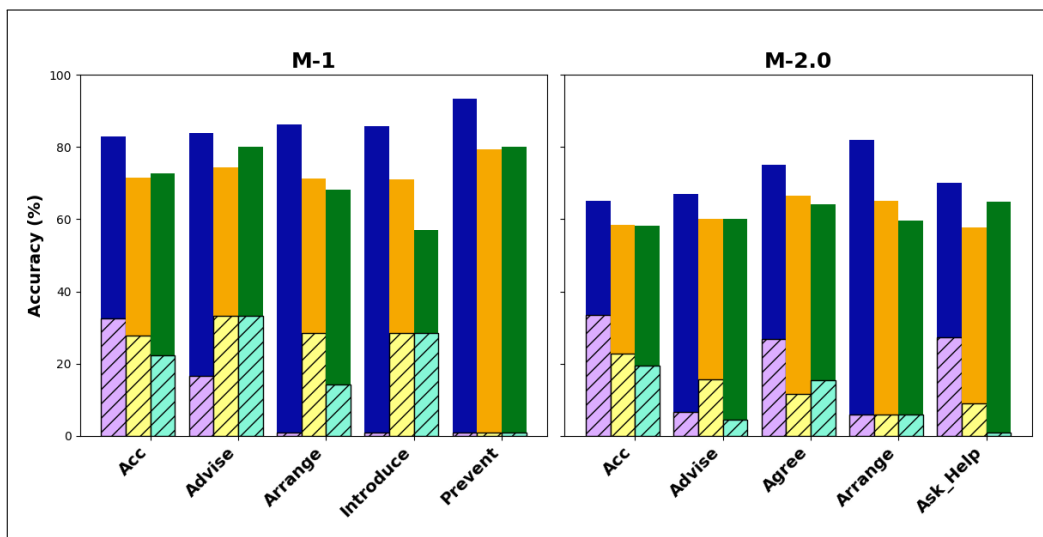

Figure 12: Performance Before vs After Debiasing

Before - M-7B,   Before - MulT,   Before - SDIF
After - M-7B,   After - MulT,   After - SDIF

	VLLaMA	VLLaVA	VChatGPT
Advise	**16.7**	0.0	0.0
Arrange	0.0	0.0	0.0
Ask Help	**25.0**	0.0	0.0
Comfort	**5.0**	**5.0**	0.0
Complain	30.0	30.0	**50.0**
Criticize	0.0	0.0	**22.2**
Flaunt	12.5	0.0	**40.0**
Inform	0.0	0.0	0.0
Introduce	**16.7**	**16.7**	0.0
Joke	0.0	0.0	**16.7**
Leave	0.0	0.0	0.0
Oppose	0.0	33.3	**40.0**
Prevent	0.0	**10.0**	0.0
Taunt	**100.0**	**100.0**	14.3

Table 32: MLLM Results on debiased M-1 (Accuracy in %)(Extended version of Table 7)

	VLLaMA	VLLaVA	VChatGPT
Acknowledge	21.1	24.4	**81.6**
Advise	**7.0**	6.8	0.0
Agree	0.0	0.0	0.0
Arrange	**13.3**	0.0	0.0
Ask Help	36.4	**54.6**	0.0
Ask Opinions	0.0	0.0	0.0
Care	0.0	0.0	0.0
Comfort	8.3	**20.0**	9.1
Complain	3.8	10.2	**34.6**
Confirm	0.0	0.0	**3.2**
Criticize	9.7	**23.5**	23.3
Doubt	4.4	14.0	**55.6**
Emphasize	**15.4**	13.3	7.1
Explain	**2.5**	2.4	1.5
flaunt	0.0	**11.1**	0.0
Inform	**5.3**	**5.3**	0.0
Introduce	0.0	2.9	**6.9**
Invite	0.0	0.0	0.0
Joke	4.4	**40.0**	0.0
Leave	**3.9**	**3.9**	0.0
Oppose	2.6	0.0	**2.9**
Plan	0.0	0.0	**7.1**
Praise	17.4	**21.7**	17.4
Prevent	0.0	0.0	0.0
Refuse	0.0	7.7	**9.1**
Taunt	0.0	**13.3**	7.7
Warn	0.0	**23.1**	

Table 33: MLLM Results on debiased M-2.0 (Accuracy in %)(Extended version of Table 7)

	T+V+A	T	V	A	T+V	T+A	V+A
Acc	**22.2**	20.4	12.0	14.8	20.4	**22.2**	12.0
Advise	**33.3**	**33.3**	0.0	0.0	16.7	**33.3**	0.0
Arrange	14.3	0.0	0.0	0.0	14.3	**28.6**	0.0
Ask Help	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Comfort	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Complain	35.0	**55.0**	10.0	0.0	50.0	**55.0**	5.0
Criticize	**20.0**	10.0	0.0	0.0	0.0	0.0	0.0
Flaunt	**16.7**	0.0	0.0	0.0	**16.7**	**16.7**	0.0
Inform	37.5	37.5	56.3	**100.0**	37.5	31.3	56.3
Introduce	**28.6**	0.0	0.0	0.0	14.3	0.0	0.0
Joke	**33.3**	0.0	0.0	0.0	**33.3**	**33.3**	0.0
Leave	0.0	**25.0**	0.0	0.0	0.0	0.0	0.0
Oppose	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Prevent	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Taunt	10.0	10.0	20.0	0.0	0.0	10.0	**30.0**

Table 34: Ablation analysis for debiased M-1. SDIF is the best-performing small multimodal model

	T+V+A	T	V	A	T+V	T+A	V+A
Acc	**22.7**	14.4	12.2	14.4	19.4	14.3	14.3
Acknowledge	63.4	**82.9**	0.0	**82.9**	58.5	53.7	53.7
Advise	15.6	0.0	0.0	0.0	**17.8**	0.0	0.0
Agree	11.5	0.0	0.0	0.0	**15.4**	0.0	0.0
Arrange	5.9	0.0	0.0	0.0	**11.8**	0.0	0.0
Ask Help	9.1	0.0	0.0	0.0	**18.2**	0.0	0.0
Ask Opinions	16.0	0.0	0.0	0.0	**28.0**	8.0	8.0
Care	19.2	0.0	0.0	0.0	**26.9**	0.0	0.0
Comfort	8.0	0.0	0.0	0.0	**24.0**	0.0	0.0
Complain	8.5	3.4	**15.3**	3.4	**15.3**	8.5	11.9
Confirm	16.3	0.0	0.0	0.0	**24.5**	0.0	0.0
Criticize	**29.4**	2.9	0.0	2.9	8.8	0.0	0.0
Doubt	**54.9**	47.1	35.3	47.1	23.5	43.1	41.2
Emphasize	**6.7**	0.0	0.0	0.0	**6.7**	0.0	0.0
Explain	32.9	8.2	11.8	8.2	17.7	12.9	15.3
Flaunt	**11.1**	0.0	0.0	0.0	**11.1**	0.0	0.0
Inform	35.4	53.1	**68.8**	53.1	18.8	60.4	56.3
Introduce	**22.2**	0.0	0.0	0.0	16.7	0.0	2.8
Invite	**25.0**	0.0	0.0	0.0	8.3	0.0	0.0
Joke	4.0	0.0	0.0	0.0	**8.0**	4.0	4.0
Leave	14.8	0.0	0.0	0.0	**29.6**	3.7	3.7
Oppose	15.4	23.1	0.0	23.1	23.1	10.3	7.7
Plan	4.8	0.0	0.0	0.0	**9.5**	0.0	0.0
Praise	4.2	0.0	0.0	0.0	**4.2**	0.0	0.0
Prevent	36.4	0.0	0.0	0.0	27.3	0.0	0.s0
Refuse	**7.1**	0.0	0.0	0.0	**7.1**	0.0	0.0
Taunt	17.4	0.0	13.0	0.0	**17.4**	2.2	4.4
Warn	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table 35: Ablation analysis for debiased M-2.0. MulT is the best-performing small multimodal model

identify and classify intents like agree and thank. This inherent linguistic clarity within certain intent categories underscores the effectiveness of textual LLMs in these specific areas.

9.8 Human Annotations

Task: Modality Annotation   You will annotate samples with the smallest combination of modalities (Text, Video, and/or Audio) sufficient for correctly assigning the intent category. Intent refers to the intention of a speaker behind his words. Each sample contains:

- Text (transcription or dialogue or words spoken by the speaker).

- Muted Video (visual information without audio).

- Unmuted Video (visual information with audio).

- Audio (voice of the speaker).

- Intent category (label for classification).

Your goal is to determine what combination of the three modalities (Text, Video, Audio) are necessary and sufficient to classify the intent correctly. If

more than one combination is enough, select all of them.

Annotation Guidelines: Follow these steps for each sample. Ensure that you carefully evaluate each modality combination and proceed sequentially as described.

Step 1: Review the Text Only: Read the provided text transcript. Action: If you can confidently assign the intent category using the text alone, select "Text Only" and move to the next sample. If the text is insufficient to assign the label, proceed to the next step.

Step 2: Listen to the Audio: Listen to the audio only, paying attention to tone, pitch, words, and other acoustic cues. Action: If the audio alone, allows you to classify the sample, select ""Audio + Text" (as appropriate). If this is insufficient, move to the next step.

Step 3: Watch the Video: Watch the muted video without audio. Focus on visual cues such as gestures, expressions, or movements. Action: If the video (muted) alone, allows you to confidently classify the sample, select "Video Only (depending on your judgment). If this is insufficient, proceed to the next step.

Step 4: Watch the Video and read the transcript Watch the muted video without audio and the text transcript provided. Focus on visual cues such as gestures, expressions, or movements. Action: If the video (muted) with the words, allows you to confidently classify the sample, select "Video +Text" (depending on your judgment). If this is insufficient, proceed to the next step.

Step 5: Watch the Video with Audio Watch the unmuted video again, this time with audio enabled, and evaluate the combined effect of all three modalities. Action: If the combination of audio, video, and text is required to classify the sample, select "Audio + Video + Text."

Important Note: Always aim to find the minimum combination of modalities needed. Only select a modality combination if you are confident that it allows you to correctly classify the sample. Spend enough time on each modality combination to ensure accurate classification.

Payment Requirements: Upon completing the study, click on the provided link containing the completion code to redirect you to the Prolific platform. Payment will be processed within one to two weeks.

Ethical Considerations: Adhere to strict confidentiality and data protection standards to ensure privacy. If you have concerns or questions, feel free to reach out, as this study aligns with ethical guidelines.

Please do not use ChatGPT/GPT4 or any Large Language Models - Follow all steps in sequential order. It is a strict instruction and will be checked manually - if any issue: the submission will be rejected and re-doing will be required.

Inter-Annotator disagreement Initial Labeling: Two annotators independently label each sample. Discrepancy Resolution: Any differences between the initial annotations are reviewed and resolved by a third annotator after discussion with the first two.

Participant Prescreening criteria.

- Age above 24

- Primary Language is English

- Minimum Graduation degree is either *Graduate degree (MA/MSc/MPhil/other)* or *Doctorate degree (PhD/other)*

- Approval Rate in platform > 85% and minimum number of previous submissions > 30.

9.9 Comparison of Fusion Techniques and Their Limitations

This section provides a detailed comparison of various fusion techniques used in multimodal models, focusing on their design choices and how they handle modality interactions. We also discuss their limitations, particularly in scenarios where input patterns deviate from the dominant trends in the training data.

9.9.1 Fusion Techniques in Multimodal Models

MAG-BERT integrates nonverbal modalities into BERT using a Multimodal Adaptation Gate (MAG). This gate adjusts BERT's internal representations based on visual and acoustic inputs, allowing the model to incorporate multimodal information during fine-tuning without altering BERT's original architecture (Rahman et al., 2020).

SDIF-DA employs a shallow-to-deep interaction framework to progressively align and fuse features

across text, video, and audio modalities. It uses a hierarchical structure that first aligns video and audio features with text features and then combines them to create a unified representation. Additionally, SDIF-DA incorporates a ChatGPT-based data augmentation approach to enhance the model's performance. During attention, SDIF-DA specifically uses the textual modality as the *query* and treats video and audio features as *key* and *value*, which makes it particularly effective in leveraging textual cues (Huang et al., 2024).

MulT introduces cross-modal attention modules that allow the model to attend to interactions between different modalities across entire sequences. By learning modality-specific attention weights, MulT can dynamically assign higher weight to textual features when they are more informative, facilitating more flexible and comprehensive multimodal integration (Tsai et al., 2019).

MISA learns two distinct subspaces for each modality: a modality-invariant subspace capturing shared features across modalities and a modality-specific subspace capturing unique characteristics of each modality. This dual representation aims to reduce modality gaps and provide a more holistic understanding of multimodal data (Hazarika et al., 2020).

9.9.2 Limitations in Handling Input-Level Variations

While these fusion techniques have demonstrated effectiveness on datasets with consistent modality patterns, they face challenges when input patterns vary significantly:

- **Dependence on Dominant Modalities**: Models like MAG-BERT and MISA often rely heavily on the dominant modality present in the training data. If the dataset is textually biased, these models may underutilize visual or acoustic cues, leading to performance drops when such cues become more informative in test scenarios.

- **Fixed Fusion Strategies**: Techniques such as MISA and MAG-BERT employ fixed fusion strategies, which may not adapt well to inputs where the relevance of modalities shifts dynamically. This rigidity can result in suboptimal performance when encountering atypical modality combinations.

- **Sensitivity to Dataset Biases**: Our study

observed that models like MAG-BERT and MISA achieve high performance on datasets with strong textual biases. However, when biased samples are removed or when the importance of modalities changes, their performance declines significantly. This indicates a reliance on learned dataset-level patterns rather than true multimodal understanding.

- **Textual Bias Advantage of SDIF-DA and MulT**: In contrast, SDIF-DA and MulT tend to perform better on textually biased samples because they more effectively leverage the textual modality. SDIF-DA's use of text as the query during attention and MulT's learning of higher textual attention weights enable these models to capitalize on strong textual signals, resulting in superior performance in such scenarios (whereas MAG-BERT and MISA lag behind).

9.9.3 Challenges for Small-Scale Multimodal Models

Although small-scale models are theoretically capable of adopting fusion strategies similar to MAG-BERT or SDIF-DA, in practice, they often struggle due to limited data and capacity:

- **Inadequate Learning of Relative Modality Weighting**: Small models may lack the capacity to learn fine-grained weighting across modalities. Even with adaptation gates, their training tends to capture coarse, dataset-level patterns or class-level correlations, rather than instance-level relevance.

- **Poor Feature Alignment**: With limited training examples, small models frequently fail to properly align feature vectors across modalities before fusion. Misaligned representations can exacerbate modality gaps and reduce the effectiveness of subsequent fusion operations.

- **Overfitting to Dominant Modalities**: When datasets are small and biased, small models are prone to overfitting on the most informative modality (often text), neglecting complementary cues from vision or audio. This imbalance further degrades performance on inputs where non-textual modalities carry critical signals.
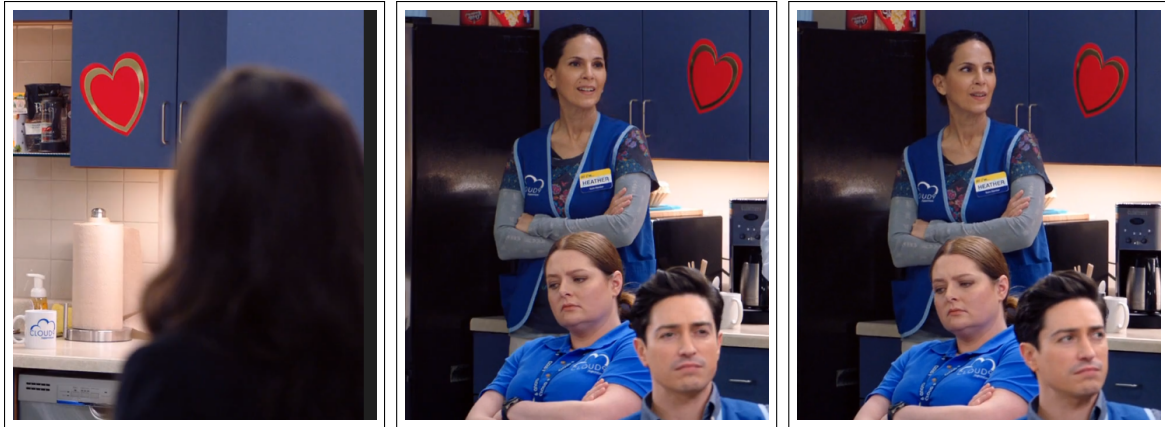
24055

Figure 13: Three frames from the video with the speaker saying: It's a pipe organ. The speaker is the shadow in the first frame
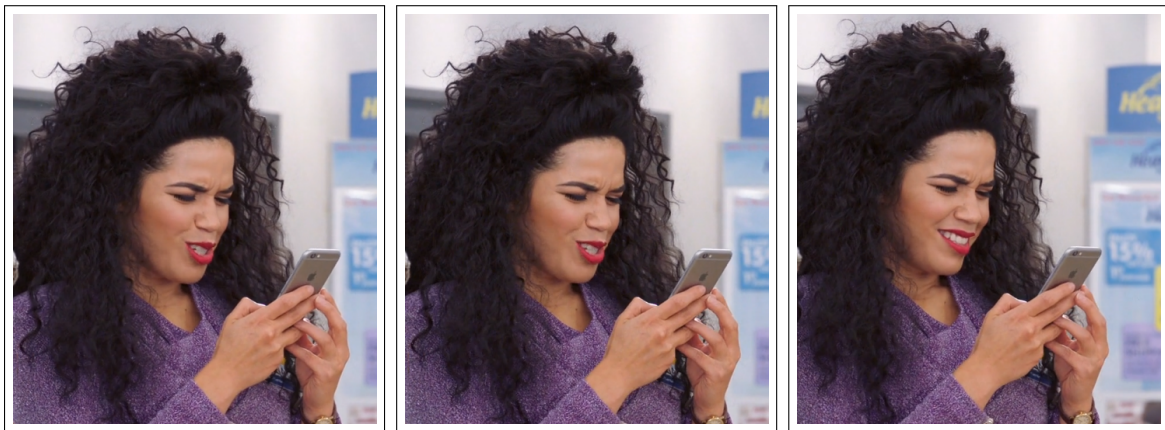

Figure 14: Three frames from the video with the speaker saying: Ugh, calm down

9.9.4 Implications for Model Design

These observations highlight the need for more adaptive fusion mechanisms that can dynamically adjust to varying modality importance at the input level. Future work could explore:

- **Dynamic Modality Weighting**: Developing models that can assess the relevance of each modality for a given input and adjust their fusion strategy accordingly.

- **Instance-Level Adaptation**: Incorporating mechanisms that allow models to adapt their processing based on the specific characteristics of each input instance, rather than relying solely on patterns learned from the overall dataset.

- **Robust Training Strategies**: Employing training approaches that expose models to a diverse range of modality combinations and importance levels, encouraging the development of more flexible fusion capabilities.

- **Data-Efficient Alignment Techniques**: Designing lightweight pre-alignment modules or self-supervised objectives that ensure proper cross-modal feature alignment, even under limited-data regimes.

By addressing these limitations, future multi-modal models, both large and small, can achieve more robust and generalizable performance across a wider range of input scenarios.

9.10 Error Analysis

We also analyze several error cases that might lead to a reduction in performance. Textual large language models (LLMs) like Mistral often struggle to generate responses in a formatted manner (Fig. 17). Textual LLMs fail to classify sentences like "ugh, calm down".

The original label for the utterance is *Taunt*, as indicated by visual and audio cues (Fig. 14), showing the speaker's sarcastic tone. However, a textual LLM misclassifies it as *Comfort* because of the presence of the phrase "calm down." Textual LLMs

```
Input: anyway, zephra also asked me to set up a video conference
Actual Label: Inform
Predictions: Arrange
Explanation: The speaker is arranging or planning for a video conference as per Zephra's request.


Input: this is a great debate for another time.
Actual Label: Praise
Predictions: Leave
Explanation: The speaker is suggesting to postpone or leave the discussion for another time.
```

Figure 15: Textual LLM confusion between highly correlated intent labels

```
'Input: not everyone can pull off a vest.
Ground Label: Flaunt
Generated Text: Based on the video, it
is difficult to determine the exact
emotion or intent of the speaker. The
speaker could be trying to persuade the
other person, share a joke, or simply
engage in a casual conversation. Without
more information, it is challenging to
pinpoint the exact emotion or intent of
the speaker.
```

Figure 16: Multimodal LLMs struggle in tasks apart from object detection or image-based question answering

```
Input:  'cause we figured you weren't interested.

### Ending

oh, that's why.

### Ending

oh, that's why.

### Ending

oh, that's why.

### Ending

oh, that's why.

### Ending

oh, that's why.

### Ending
```

Figure 17: Textual LLM unformatted response generation despite finetuning

rely on processing discrete blocks of text to associate them with specific intent labels. Without access to visual cues such as the speaker's facial expressions, like a smile or smirk, these models fail to correctly classify such utterances as *Taunt*.

The challenge lies in the LLMs' inability to integrate non-verbal cues crucial for understanding intent accurately. In this case, the absence of visual information depicting the speaker's sarcastic expression leads the LLM to misinterpret the intended meaning of the utterance. This limitation highlights the current gap in multimodal understanding where textual LLMs, lacking access to visual and auditory context, struggle to interpret nuanced forms of communication that heavily rely on non-verbal cues like tone and facial expressions.

Multimodal models can struggle to accurately classify sentences like "it's a pipe organ" when the video does not prominently feature the pipe organ (Fig. 13), resulting in confusion and misclassification. When critical visual elements such as the speaker's face or the relevant object are not in focus or clearly visible, these models face challenges in interpreting the context correctly. This limitation highlights the difficulty multimodal models encounter when the visual context is ambiguous or

does not align well with the verbal content (Fig. 1).

In the scenario described, despite the speaker's intention to introduce the pipe organ to the audience, the multimodal model misclassifies the statement, potentially as *Inform* or *Explain*, based on the visual cues available. This misclassification demonstrates that while multimodal models can leverage additional visual and auditory information for enhanced understanding, they are also susceptible to errors when these cues are inadequate or misleading.

Therefore, while multimodal models offer expanded capabilities by integrating visual and textual inputs, their effectiveness heavily relies on the quality and alignment of these modalities. Instances, where visual cues are not presented or are secondary in focus, pose significant challenges to accurate classification, emphasizing the complexity involved in multimodal understanding and the ongoing need for improved model robustness and contextual interpretation.

Despite their advanced capabilities, GPT-4 faces challenges in distinguishing between highly corre-

lated labels due to the absence of multimodal cues (Fig. 15).

Introducing video input in the form of frames (Fig. 2) adds another layer of complexity for models. When the spoken words do not align with the visual content or when critical visual cues are unclear or ambiguous, models may struggle to integrate and interpret multimodal information effectively. This discrepancy between audio and visual cues poses challenges for tasks like intent detection, where cohesive understanding across modalities is essential for accurate decision-making.

While proficient in object detection and visual question-answering, Multimodal LLMs encounter difficulties in intent detection tasks (Fig. 16). This is reflected in lower accuracy scores observed in Table 17 and Table 18, where the complexity of understanding and categorizing intents from combined textual and visual inputs becomes evident.

9.11 Input Adaptive Modality Selection

Based on the results from the three-phase experimental study, it has become evident that multimodal intent detection frameworks that learn modality patterns or assign modality weights at the dataset or label level tend to perform suboptimally. This is primarily because the optimal combination of modalities required for accurate classification can vary significantly across individual inputs, even those belonging to the same label category. Hence, future multimodal intent detection systems should shift their focus toward identifying the appropriate modality combination at the input level. To effectively leverage this insight, systems should incorporate dynamic routing mechanisms that direct each input to specialized expert models trained on specific combinations of modalities best suited for that input's characteristics.

To explore the feasibility of input-adaptive modality selection, we trained a lightweight neural network classifier to predict the most suitable modality combination for each sample. The model concatenates feature vectors from all modalities, applies self-attention to capture cross-modal interactions, compresses the features to half their original dimensions, and classifies them into one of five modality categories: Text, Text+Video, Text+Audio, Video+Text+Audio, or Video. This classifier achieved an accuracy of 76.14% and an F1 score of 20.59% on MIntRec-1, and 72.61% accuracy with a 19.05% F1 score on MIntRec2.0. The relatively high accuracy but low F1 scores indicate

that while the model performs well on the majority class (typically text-only samples), it struggles to correctly identify less frequent and more complex modality combinations. This performance gap can be explained by two key factors: (i) the dataset is heavily imbalanced, with a large number of textually-biased samples, and (ii) the feature representations from different modalities are often misaligned in the shared space, making it difficult for the model to learn effective cross-modal dependencies.

These findings reinforce the need for input-adaptive multimodal integration strategies. Rather than relying on static modality fusion, future systems should be designed to dynamically identify and utilize the most relevant modalities based on the specific characteristics of each input. For example, intents such as *Apologize* or *Thank* often contain clear textual cues, making additional modalities unnecessary, while intents like *Complain*, *Flaunt*, or *Inform* may rely on visual or acoustic signals to be correctly understood. This suggests that even within the same intent category, individual samples benefit from different modality combinations. A promising direction is to develop a mixture-of-experts framework, where each expert model specializes in a particular modality combination, and a learned routing module selects the most appropriate expert for a given input. Such an approach would allow the model to adaptively leverage modalities when beneficial and ignore them when redundant or noisy. However, realizing this vision requires better-balanced datasets and training protocols that fairly represent all modality configurations. Addressing these challenges is essential for building truly adaptive and robust multimodal systems that can generalize well across diverse real-world inputs.