

Uncertain Multimodal Intention and Emotion Understanding in the Wild

Qu Yang¹, Qinghongya Shi¹, Tongxin Wang¹ and Mang Ye^{1,2} ✉

¹ School of Computer Science, Wuhan University, Wuhan, China

² Taikang Center for Life and Medical Sciences, Wuhan University, Wuhan, China

{yangqu, shiqinghongya, wangtx, yemang}@whu.edu.cn

<https://github.com/yan9qu/CVPR25-MINE>

Abstract

Understanding intention and emotion from social media poses unique challenges due to the inherent uncertainty in multimodal data, where posts often contain incomplete or missing modalities. While this uncertainty reflects real-world scenarios, it remains underexplored within the computer vision community, particularly in conjunction with the intrinsic relationship between emotion and intention. To address these challenges, we introduce the **Multimodal Intention and Emotion Understanding in the Wild (MINE)** dataset, comprising over 20,000 topic-specific social media posts with natural modality variations across text, image, video, and audio. MINE is distinctively constructed to capture both the uncertain nature of multimodal data and the implicit correlations between intentions and emotions, providing extensive annotations for both aspects. To tackle these scenarios, we propose the **Bridging Emotion-Intention via Implicit Label Reasoning (BEAR)** framework. BEAR consists of two key components: a BEIFormer that leverages emotion-intention correlations, and a Modality Asynchronous Prompt that handles modality uncertainty. Experiments show that BEAR outperforms existing methods in processing uncertain multimodal data while effectively mining emotion-intention relationships for social media content understanding.

1. Introduction

Intention and emotion are fundamental in shaping human social behaviors and decision-making processes [25, 32, 38]. Understanding their interplay is crucial for the advancement of artificial general intelligence. While prior studies [6, 42, 43] in the vision community have independently explored intention or emotion recognition in multimodal contexts (as illustrated in Fig. 1), employing datasets like MIntRec [47] and IEMOCAP [2], these datasets fail

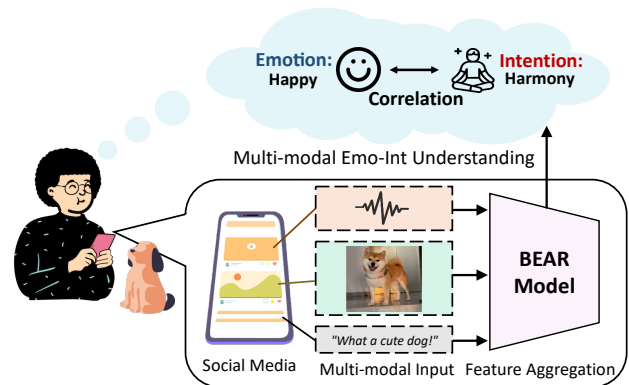


Figure 1. Overview of multi-modal emotion and intention understanding from social media platforms.

to capture the inherent link between intention and emotion and overlook the challenges of uncertain modalities and thematic diversity in real-world scenarios. Addressing these gaps is vital for a comprehensive understanding of intentions and emotions in natural environments. In this paper, we introduce the **Multimodal Intention and Emotion Understanding in the Wild (MINE)** dataset. MINE provides detailed annotations of both intention and emotion, establishing a platform to explore their correlation. It presents the challenges of uncertain modalities, thus mirroring the intricacies of real-life social interactions, as shown in Fig. 2.

The design of MINE is guided by several key principles: 1) MINE addresses a critical gap in existing datasets by enabling comprehensive insights into the interactions between intention and emotion within social contexts. 2) In contrast to data typically curated under controlled conditions (such as scripted dialogues or TV series clips [2, 30, 35, 44, 46, 47]), which exhibit significant disparities with real-world situations, MINE leverages authentic user-generated content from social platforms, thus mirroring real-world dynamics of intention and emotion expression. 3) Unlike related datasets that typically assume fixed modalities [2, 47], MINE incorporates the challenge of uncertain modalities in social media, enhancing the realism and applicability of

✉ Corresponding author.

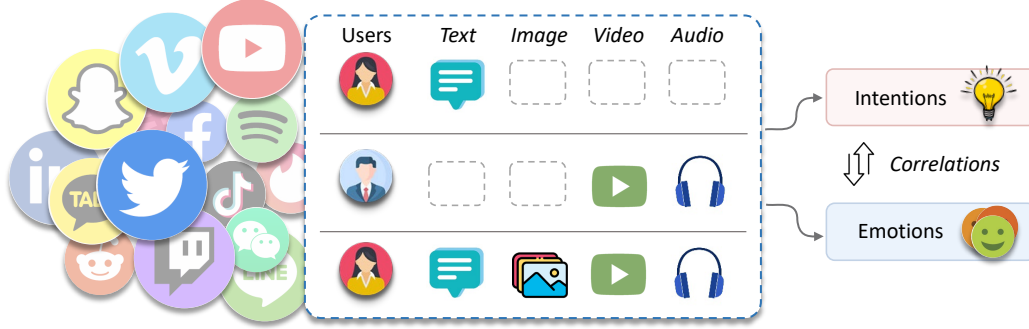


Figure 2. Overview of the MINE dataset, illustrating the multimodal data collection from social media, categorized by topics and including text, image, video, and audio content, each reflecting various intentions and emotions. The modality combination is uncertain.

intention and emotion understanding tasks. At the algorithmic level, this manifests as the missing modality problem, where modality heterogeneity is primarily reflected in the absence of certain modal inputs. Previous research on missing modalities [3, 50] often relies on artificial sampling methods, creating biases and noise due to incomplete information (*e.g.*, an image showing “news about something” with the text “That’s so ironic!”, where removing text leads to misinterpretation). MINE, however, preserves native modal information, eliminating potential errors and offering a reliable benchmark for **uncertain modality** research. 4) By utilizing topical tags (such as #CAMPUS) from social media, MINE provides objective context for analyzing intentions and emotions, enabling a more nuanced understanding of these elements within specific thematic frames.

To harness the intricate relationship between emotions and intentions within the MINE dataset, we propose **Bridging Emotion-Intention via Implicit Label Reasoning** (BEAR) framework, comprising two key components: BEIFormer and Modality Asynchronous Prompt (MAP). The BEIFormer introduces an innovative contrastive task, utilizing [MASK] tokens to conceal intention and emotion labels within training prompts. This design enables the model to reconstruct textual data that is in alignment with ground truth labels, leveraging multimodal inputs and the process of unmasking text. This implicit reconstruction and alignment process in BEAR enhances the understanding by delving into the relationship between emotion and intention. Furthermore, the MAP is specifically engineered to tackle the challenge of uncertain modalities. It provides customized cues tailored for diverse uncertain scenarios, thereby substantially improving understanding capacity in contexts where modality data may be incomplete.

In summary, our contribution is threefold:

(1) We introduce MINE, the first large-scale dataset that captures the inherent relationship between intention and emotion. Comprising over 20,000 topic-specific posts with natural modality distribution, MINE uniquely provides accurate annotations for intention and emotion, enabling systematic study of their correlations in realistic scenarios.

(2) We propose Bridging Emotion-Intention via Implicit Label Reasoning, a novel framework that bridges emotion and intention understanding. It incorporates two key components: (i) BEIFormer that implicitly mines emotion-intention relationships through contrastive learning, and (ii) Modality Asynchronous Prompt that handles uncertain modalities without compromising existing features.

(3) We establish comprehensive benchmarks on MINE by evaluating multiple baselines. Through extensive experiments, we demonstrate that BEAR consistently outperforms existing methods by significant margins (*e.g.*, +2.65% Macro F1 for intention recognition and +3.54% accuracy for emotion recognition), validating the importance of joint emotion-intention modeling.

2. Related Work

2.1. Intention and Emotion Understanding

Intention recognition is to understand the goals behind human actions. While early datasets like ATIS [9, 37] were limited to specific domains, recent datasets like CLINC150[15], Intentionomy [11], and MIntRec [47, 48] have expanded to diverse sources including social media and TV series.

For emotion recognition, datasets have evolved from controlled environments (IEMOCAP [2]) to more naturalistic settings with CMU-MOSEI [46], PanoSent [21], and MELD [26], encompassing multimodal expressions across text, speech, and visual cues. However, as shown in Tab. 1, existing datasets typically focus on either intention or emotion in isolation. While recent works such as EMO-TyDA [31], MC-EIU [17], and EmoInt-MD [35] attempt to combine both aspects, they are constrained by controlled settings and fixed modalities, unlike real-world social media posts which often contain varying combinations of modalities. Our MINE dataset addresses these limitations by capturing authentic emotion-intention relationships while preserving the natural complexity of missing modalities.

Type	Dataset	Modality				Missing	Class Nums		Data Nums	Source
							Emo.	Int.		
Intention	CLINC150 [15]	✓	✗	✗	✗	✗	-	150	23,700	Acted
	HWU64 [18]	✓	✗	✗	✗	✗	-	64	25,716	Acted
	Intentionomy [11]	✓	✓	✗	✗	✗	-	28	14,455	In the wild
	MDID [14]	✓	✓	✗	✗	✗	-	8	1,299	In the wild
	MSED [10]	✓	✓	✗	✗	✗	-	6	9,190	In the wild
	MIntRec [47]	✓	✗	✓	✓	✗	-	20	2,224	TV
	MIntRec2.0 [48]	✓	✗	✓	✓	✗	-	30	15,040	TV
Emotion	IEMOCAP [2]	✓	✗	✓	✓	✗	6	-	7,433	Acted
	CMU-MOSEI [46]	✓	✗	✓	✓	✗	6	-	23,453	In the wild
	GoEmotions [4]	✓	✗	✗	✗	✗	28	-	58,000	In the wild
	EmotionRoI [24]	✓	✗	✗	✗	✗	6	-	1,980	In the wild
	MELD [26]	✓	✗	✓	✓	✗	8	-	13,708	TV
All	EMOTyDA [31]	✓	✗	✓	✓	✗	10	12	19,365	Acted
	MINE	✓	✓	✓	✓	✓	11	20	20,168	In the wild

Table 1. Comparison of existing datasets for intention (Int), emotion (Emo), and combined (All) understanding. MINE uniquely addresses uncertain modalities while combining both emotion and intention analysis in real-world scenarios.

2.2. Uncertain Modality in the Wild

Uncertain modality (or missing modality) presents a fundamental challenge in multimodal understanding. Contemporary studies [3, 16, 22, 23, 50] have investigated this challenge using datasets like IEMOCAP [2] and CMU-MOSEI [46]. Researchers have proposed various approaches, ranging from automated fusion strategy identification [22] and meta-sampling integration [3] to unified frameworks for modality inference [29, 50] and Bayesian meta-learning applications [23].

However, these methods typically rely on random sampling to simulate missing modalities, which can introduce bias and alter the original content meaning (refer to Appendix G). In contrast, MINE captures naturally occurring uncertain modalities in the wild, providing a more authentic foundation for multimodal research.

3. MINE Dataset

In this section, we outline the creation of the MINE dataset, including steps such as data preparation, establishing categories for intentions and emotions, annotating multimodal content with these categories, and analyzing the data. Detailed metrics of the MINE dataset are provided in Tab. 2.

3.1. Topic Definition and Data Preparation

To explore how topics influence the understanding of intentions and emotions in real-world scenarios, we obtained permission from the consent of data subjects and collected high-quality, multimodal data from social media platforms—venues teeming with genuine user expressions. This approach significantly diverges from the artificial confines of laboratory settings or the limited diversity found

in sources like movie clips, frequently utilized in prior research. Our strategy involved organizing a wide range of multimodal data via clusters based on topic-specific hashtags, ensuring a rich and varied data source.

The MINE dataset comprehensively covers ten topics that are integral to daily life. Following the methodology of Kapoor *et al.* [12], we selected these topics from the top 100 most frequently discussed subjects on social media platforms. MINE explores these topics in depth, paying particular attention to intersections identified through related keywords. For instance, the keyword ‘Education’ encompasses associated themes such as ‘Social Media Impact on Teachers and Students’ and ‘Use of Social Media in Academic Learning’. To ensure thematic diversity, we selected ten distinct topics for our study and developed a comprehensive keyword framework for each one, with detailed documentation available in Appendix A. For data collection, we employed a hashtag-based method (*e.g.*, #Education) rather than general topic searches to capture a broader variety of content within each topic.

The data collection spans Twitter posts from 2019 to 2022, with all content obtained with explicit user consent. To ensure data quality, we implemented multiple preprocessing steps, including the removal of nonsensical texts (those consisting solely of links or random characters) and the filtering of harmful content (related to pornography and drugs). While visual content plays a crucial role in conveying emotional and intentional nuances, we maintained a deliberate balance with textual content to ensure comprehensive multimodal analysis. The final dataset composition reflects this approach, with visual content constituting approximately 70% of the data, creating a diverse yet imbal-

Item	Number
Number of intentions	20
Number of emotions	11
Number of text	20044
Number of images	12428
Number of videos	3895
Number of audios	1615
Average length of text utterances (words)	20.0
Total number of words	489069
Average duration of videos (seconds)	24.4
Total duration of videos (hours)	21.9

(a) Data Statistics

Text	Image	Video	Audio	Number
✓				3913
	✓			99
		✓		4
✓	✓			12261
✓		✓		2248
	✓	✓		3
		✓	✓	18
✓	✓	✓		26
✓		✓	✓	1557
✓	✓	✓	✓	39

(b) Distribution of Modality Combinations

Table 2. Data and Modality Statistics of the MINE dataset.

anced multimodal distribution, as detailed in Tab. 2.

3.2. Taxonomy Definition and Data Annotation

Defining labels for intentions and emotions presents a unique challenge due to their subjective nature. While previous work [47] treated intentions as mutually exclusive categories (‘Expressing emotions’ or ‘Achieving goals’), we argue that social interactions naturally blend both elements, aligning with philosophical theories [32].

MINE incorporates 20 intention and 11 emotion labels per data point, drawing from literature on both dialogue and visual intention analysis [11, 47] and emotion research [2, 34]. We frame intention recognition as a multi-label task to capture the nuanced nature of social media interactions, while structuring emotion recognition as a single-label task to identify dominant emotions, consistent with established practices [2]. To address inherent subjectivity, our annotation employs a crowd-sourcing platform with dual annotations. Detailed category definitions and annotation guidelines are provided in Appendices A and H.

The annotation workflow involved two independent groups of annotators (15 and 13 members) under domain expert supervision. Through iterative training and validation, we reduced the initial annotation discrepancy rate from over 25% to below 3%. The final dataset comprises 16,000 training samples, 2,000 validation samples, and 2,000 test samples, with 20% undergoing additional expert review. Privacy protection measures and detailed annotation protocols are documented in Appendices C and H.

Analysis of the annotated data reveals strong correlations between intentions and emotions. For instance, aggressive intentions (“Criticize”, “Disagree”) frequently co-occur with negative emotions. We also observe that certain intentions can complement emotional states - for example, “Comfort” intentions often appear in posts with “Sad” emotions, suggesting an attempt to provide emotional support. Additionally, while topic-intention correlations exist (e.g., “Teach” with education-related content), they are more one-to-one and less pronounced than emotion-intention relationships. Detailed analyses are provided in Appendix B.

3.3. Feature Pre-extraction

We pre-extract features from each modality (text, image, video, and audio) in the dataset to facilitate multimodal fusion. This approach offers three key advantages: it reduces signal complexity and computational costs for researchers, enables consistent and fair comparisons across different multimodal fusion techniques, and minimizes privacy risks in our data. For feature extraction, we employ state-of-the-art pre-trained models: BERT [13] for text, ViT-S [5] for images, Video Swin Transformer [20] for video, and wav2vec 2.0 [1] for audio. All extracted features are aligned to a dimension of 768 to ensure consistent multimodal fusion. Detailed feature extraction procedures are documented in Appendix C.

4. Proposed Method

4.1. Bridging the Relation of Emotion and Intention

As previously highlighted, the interconnection of intentions and emotions is important in both real-life situations and neural network analyses. Existing methodologies that focus solely on either intention or emotion often overlook the valuable insights from their interaction, limiting comprehensive understanding in both domains. To bridge this gap, we introduce the Bridging Emotion-Intention Former (BEIFormer), as depicted in Fig. 3. This model unifies these two critical domains. Specifically, we first establish the baseline for multimodal emotion and intention understanding by pre-extracting features from multimodal data to produce text, visual, and audio modality features X_L , X_V , and X_A , respectively (as detailed in Sec. 3.3). We then employ Multi-modal Transformer [36] to aggregate features from any two modalities, extracting joint information, demonstrated here using visual and audio modalities:

$$\begin{aligned}
Y_V &= \text{softmax} \left(\frac{Q_V K_A^\top}{\sqrt{d_k}} \right) V_A \\
&= \text{softmax} \left(\frac{X_V W_{Q_V} W_{K_A}^\top X_A^\top}{\sqrt{d_k}} \right) X_A W_{V_A},
\end{aligned} \tag{1}$$

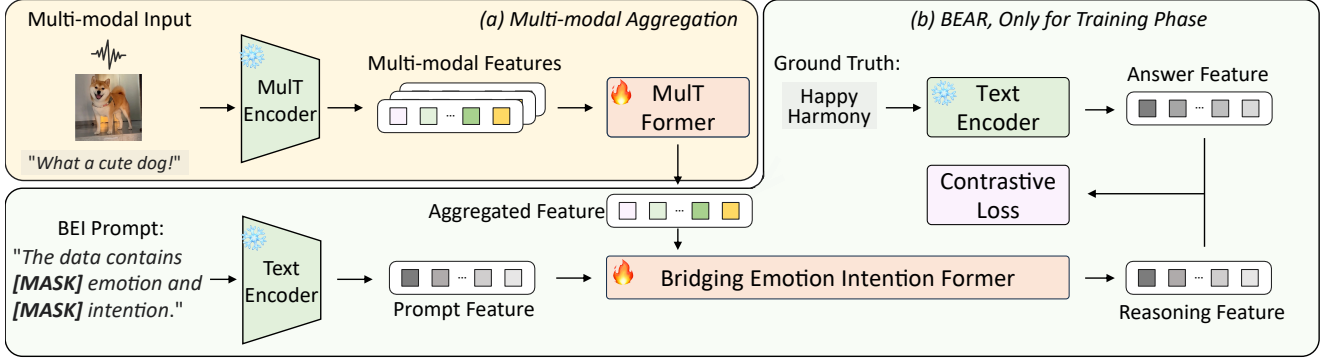


Figure 3. Architecture of the **B**ridging **E**motion-**I**ntention via **I**mplicit **L**abel **R**easoning (BEAR) framework. The framework consists of two main components: (1) a Multi-modal Former that aggregates multimodal features into a unified representation, and (2) a Bridging Emotion Intention Former that fuses aggregated and prompt features. During training, ground truth emotion and intention labels are encoded to match the answer feature, enabling effective integration of emotion-intention signals and enhanced reasoning capabilities.

where $Q_V = X_V W_{Q_V}$, $K_A = X_A W_{K_A}$, and $V_A = X_A W_{V_A}$ represent Query, Key, and Value component from the Audio modality, respectively. The term W denotes the weights. Here, Y_V maintains the same length as Q_V (i.e., T_V), but exists in the feature space of V_A . The softmax function in Eq. (1), scaled by $\sqrt{d_k}$, computes a score matrix $\text{softmax}(\cdot) \in \mathbb{R}^{T_V \times T_A}$, with its (i, j) -th entry quantifying the attention from the i -th time step in modality V to the j -th time step in modality A . Consequently, each i -th time step of Y_V represents a weighted summary of V_A , weighted by the i -th row in $\text{softmax}(\cdot)$.

After applying two-pair cross-modal adaptation to three modalities, we obtain the modality aggregated representation f_{ma} . However, this approach primarily reveals intention or emotion-related features within the multimodal data, driven by a singular training objective. It overlooks the intrinsic correlations between intentions and emotions that could serve as additional supervisory information. To leverage this relational information, we devise an implicit BEIFormer that integrates all-level intermediate tokens from both the Multi-modal Former, establishing implicit Emotion-Intention associations to enhance the understanding of these subjective dimensions. Specifically, we define the BEI Prompt T_{prom} as a textual modality prior (as shown in Fig. 3), using [MASK] tokens to erase Emotion-Intention related priors. This process generates the hint feature f_{prom} via the BEI Former (a transformer-based sub-network), serving as the Query input for the BEIFormer. The BEIFormer then utilizes the modality aggregated feature f_{ma} as Key and Value inputs to generate the final interaction output f_{beif} .

Subsequently, we transform the emotion and intention labels into textual information T_{gt} and use text encoder to align these labels within the same representation space as the hint feature, producing the ground truth feature f_{gt} . The ground truth feature and the BEIFormer interaction output f_{beif} are then used for contrastive learning during training:

$$\text{NCE}(x_i, y_i) = -\log \frac{\exp(x_i^\top y_i / \tau)}{\sum_{j=1}^B \exp(x_i^\top y_j / \tau)}, \quad (2)$$

where x and y are elements from f_{beif} and f_{gt} , respectively, B denotes the batch size, and the temperature hyper-parameter τ is empirically set to 0.05. The full training loss is defined as:

$$\mathcal{L}_{\text{total}} = \begin{cases} \mathcal{L}_{\text{emo}} + \mathcal{L}_{\text{NCE}}, & C_t = \text{emo}, \\ \mathcal{L}_{\text{int}} + \mathcal{L}_{\text{NCE}}, & C_t = \text{int}, \end{cases} \quad (3)$$

where C_t determines the current task. In our model, $\mathcal{L}_{\text{emo/int}}$ represents the conventional loss for emotion or intention classification. For emotion recognition, we employ the standard multi-class cross-entropy loss:

$$\mathcal{L}_{\text{emo}} = \frac{1}{N} \sum_i \mathcal{L}_i = -\frac{1}{N} \sum_i \sum_{c=1}^M y_{ic} \log(p_{ic}), \quad (4)$$

where N is the number of samples, M is the number of classes, y_{ic} indicates the ground truth label (1 if the i -th sample belongs to class c , and 0 otherwise), and p_{ic} is the predicted probability of the i -th sample belonging to class c . This loss function also applies to multi-label intention understanding, ensuring a consistent approach across both emotion and intention classification tasks. Through implicit engagement with semantic correlations between intention and emotion during training, the BEIFormer enhances the network's capacity to discern these subjective dimensions, enabling more nuanced understanding.

4.2. Modality Asynchronous Prompt

In addition to the interplay between Emotion-Intention, the challenge of uncertain modalities presents a significant hurdle. Existing methods fail to provide a reasonable solution. For instance, MMIN [50] focuses on learning inter-modal correlations during multimodal data training to infer missing modal information at deployment. However, this strat-

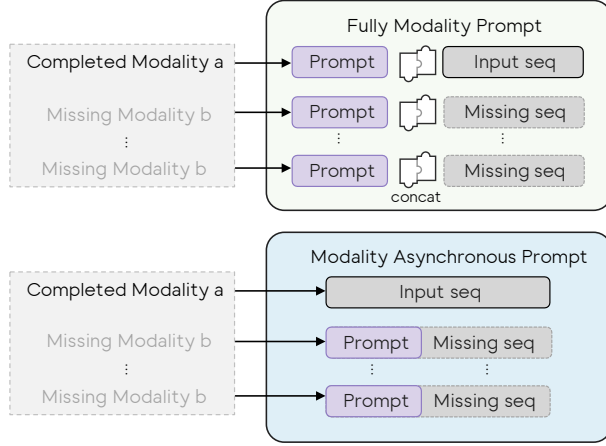


Figure 4. Comparison of Modal Asynchronous Prompt strategy with MPMM [16]. MAP preserves original representations in existing modalities while only prompting missing modalities, whereas the full modality prompt can corrupt existing features.

egy heavily relies on extensive multimodal priors, a requirement rarely met in social media contexts. While methods like MPMM [3, 16] take an alternative approach by using prompts to integrate missing information into neural networks, their strategy of indiscriminately concatenating all modalities, as shown in Fig. 4, risks compromising feature integrity. This approach can lead to data distortion, reducing model effectiveness in handling missing modalities.

In response, we introduce the Modality Asynchronous Prompt (MAP), which dynamically provides asynchronous cues tailored to various types of missing modality. For each missing type, we develop a unique set of prompts that supply contextually relevant cues, enriching the missing modalities while preserving features of existing modalities. For example, when text and visual inputs are missing, our approach focuses exclusively on enhancing the missing modalities, ensuring that audio features remain undisturbed.

Furthermore, we integrate these prompts with existing features through element-wise addition (sums prompt features with missing modality features) rather than concatenation, minimizing parameter count and computational overhead by avoiding the additional Multi Layer Perceptron required in concatenation approaches.

5. Benchmark and Experiments

5.1. Benchmark Methods

For our benchmark, we select established baseline methods in multimodal intention recognition [8, 36, 41] and missing modality challenges [16, 50], chosen for their demonstrated efficacy and influence. They are all trained on MINE.

MuT. The Multimodal Transformer (MuT) [36] employs the Transformer [39] for learning multimodal language representations through directional cross-modal attention. For

consistency, we combine image and video features for visual representation across all multimodal approaches.

MISA. Modality-Invariant and -Specific Representations (MISA) [8] differentiates modality features into invariant and specific subspaces for enhanced fusion and prediction.

MFSA. This approach [41] leverages a transformer-based model with adversarial learning to create modality-specific and -agnostic representations for sentiment recognition.

MMIN. The Missing Modality Imagination Network (MMIN) [50] predicts absent modality representations through joint multimodal representation learning.

MPMM [16] introduces a prompt-based learning framework, enhancing robustness in visual recognition tasks under various missing-modality conditions.

All missing modality experiments use MuT [36] as the baseline framework for consistency, with detailed configurations that can be found in Appendix F.

5.2. Experimental Setup

Evaluation Metrics. Following [11, 33], we utilize three multi-label metrics for intention understanding: Macro F1, Micro F1, and Samples F1, chosen for comprehensive performance assessment. Detailed metric definitions appear in Appendix E. For emotion recognition, following [47], we use accuracy, F1 score, precision, and recall to evaluate model performance comprehensively.

5.3. Main Experiments

Table 3 presents the performance comparison of various methods on intention and emotion understanding tasks. Our approach consistently outperforms existing techniques across both validation and test splits, demonstrating significant advances in handling missing modalities and mining the Emotion-Intention relationship in social media analysis.

On the validation set, our method achieves a Macro F1 score of 36.02% for intention understanding, surpassing the MuT baseline by 2.65%. For emotion recognition, it reaches 64.82% accuracy, exceeding MuT by 3.54%. These improvements persist in the test split, with our method achieving 35.79% Macro F1 and 64.53% accuracy, outperforming the baseline by 2.68% and 1.81%.

These results highlight the effectiveness of our strategy, particularly our Bridging Emotion-Intention via Implicit Label Reasoning (BEAR) approach. BEAR significantly enhances the accuracy of intention and emotion recognition, achieving substantial gains over the metrics compared to conventional training methods that treat these tasks separately. This underscores our method’s capacity to offer a deeper and more accurate analysis of intention and emotion, validating its superiority in multimodal data analysis within the context of social media.

Split	Method	Intention			Emotion			
		Macro F1	Micro F1	Samples F1	Acc.	F1 Score	Precision	Recall
Validation	Random	15.12	18.03	17.69	4.91	1.85	4.82	2.25
	MISA [8]	23.89	53.32	53.63	63.93	18.60	20.22	17.55
	MFSA [41]	31.55	53.88	53.52	64.15	19.56	20.07	17.91
	MuT [36]	34.37	52.47	53.18	61.28	19.29	20.88	18.79
	TCL.MAP [51]	34.25	53.14	53.71	63.15	19.14	20.01	18.79
	MMIN [50]	27.91	48.85	49.50	59.30	17.45	20.67	16.97
	MPMM [16]	32.88	52.07	53.00	63.15	17.65	21.10	17.58
	Ours	36.02	54.06	53.88	64.82	19.89	21.31	19.85
Testing	Random	15.18	17.96	17.61	6.25	2.06	5.17	2.00
	MISA [8]	23.74	52.27	52.54	63.28	15.55	23.98	16.52
	MFSA [41]	32.40	53.75	53.01	64.05	17.69	26.58	17.91
	MuT [36]	34.11	52.05	52.77	62.72	20.38	23.83	19.58
	TCL.MAP [51]	34.73	51.64	52.41	63.31	20.14	25.27	19.32
	MMIN [50]	27.15	48.14	49.02	57.71	15.47	17.60	15.17
	MPMM [16]	32.81	52.13	52.74	62.82	17.78	25.27	17.94
	Ours	35.25	52.90	53.35	64.67	20.55	29.64	19.95

Table 3. Performance comparison of multimodal fusion and missing modality methods on intention and emotion understanding tasks. Results show consistent improvements of our approach across all metrics on both validation and test sets.

5.4. Ablation Study

Component Analysis. Our study delves into the advantages of combining the Modality Asynchronous Prompt (MAP) with the Bridging Emotion-Intention Former (BEIF), offering a detailed examination in Tab. 4. Our results reveal that while each component individually contributes to performance improvement—MAP elevating the Macro F1 score to 33.36% and BEIF to 35.25% for intention understanding—it’s their combined application that significantly amplifies effectiveness, as evidenced by the compelling outcomes achieved (35.79% Macro F1, 53.92% Micro F1, and 54.88% Samples F1). The ablation study on MAP particularly highlights the effectiveness of our method, especially in addressing the challenge of missing modalities. The asynchronous prompting of missing information, crucially, does not compromise the feature space of the existing modalities. This approach, combined with our preference for element-wise addition over concatenation, ensures the dimensional consistency and integrity of the data representation remain intact. These findings indicate that our holistic approach is adept at offering a nuanced and thorough understanding of complex multimodal data.

Emotion-Intention Correlation Mining. To validate the effectiveness of various strategies for mining intention and emotion correlations, we conduct ablation studies, as shown in Tab. 5. The experiments explore various methods, including a generic multi-task learning paradigm that directly combines emotion and intention losses. This paradigm, while straightforward, shows limited improvement (Macro F1: +1.21%) due to the absence of a specialized Emotion-Intention mechanism during training. It can even harm the ability to understand intention in certain scenarios, with decreases in Micro F1 (-1.06%) and Samples F1 (-0.30%) compared to the baseline. While the Complete Information method (Comp. Info.), which employs one aspect as a prior during both training and inference, exhibits notable gains (+9.06% Macro F1, +9.80% Micro F1), its practical application is limited by available prior knowledge in real-world scenarios. The limitations of these alternative approaches further emphasize the importance of BEAR, which achieves consistent improvements (+2.44% Macro F1) without requiring additional inference-time information.

Feature Extractor Impact. Beyond model architecture, we investigate how different feature extractors influence in-

Component			Intention			Emotion			
MAP	EA	BEIF	Macro F1	Micro F1	Samples F1	Acc.	F1 Score	Precision	Recall
			32.81	52.13	52.74	62.82	17.78	25.27	17.94
✓			33.36	52.74	53.43	62.72	18.95	28.43	19.30
✓	✓		34.37	52.47	53.18	61.28	19.29	25.88	18.79
		✓	35.25	52.90	53.35	64.67	20.02	29.19	19.35
✓	✓	✓	35.79	53.92	54.88	64.53	20.55	29.64	20.35

Table 4. Ablation study showing the contribution of each component. Integration of both Modality Asynchronous Prompt (MAP) and Bridging Emo-Int Former (BEIF) yields optimal performance across all metrics. EA means element-wise addition.

Method	Intention		
	Macro F1	Micro F1	Samples F1
Baseline	32.81	52.13	52.74
w/ Multi-task	34.02	51.07	52.44
	(+1.21)	(-1.06)	(-0.30)
w/ EmoInt-Trans [35]	31.59	52.06	51.32
	(-1.22)	(-0.06)	(-1.42)
w/ BEIFormer	35.25	52.90	53.35
	(+2.44)	(+0.77)	(+0.61)
w/ Comp. Info.†	41.87	61.93	62.09
	(+9.06)	(+9.80)	(+8.54)

Table 5. Advanced relation mining strategies are important in enhancing emotion-intention analysis. †represents the utilization of supplementary information during the inference process.

tention and emotion understanding performance. We evaluate various state-of-the-art extractors for visual [20, 27], audio [7], and textual [49] modalities. As shown in Tab. 6, more advanced feature extractors consistently improve performance. CLIP-ViT-S [27] provides the most substantial gains for visual features (+1.81% Macro F1), while SemBERT [49] significantly enhances text representation (+1.42% Micro F1). Notably, while AST shows modest improvements in the metrics (+0.76% Macro F1), its impact is more significant than these numbers suggest, given that audio data comprises only a small portion of the test set.

Generalization Analysis on CMU-MOSEI. To evaluate the generalizability of our MAP approach, we conduct experiments on the CMU-MOSEI dataset across various baseline architectures, as detailed in Tab. 7. MAP demonstrates consistent improvements across all tested configurations. These consistent gains across different architectures and modality combinations validate MAP’s robustness in handling missing modalities.

Data Imbalance. The modality imbalance in MINE is intentional, reflecting the natural, in-the-wild modality distribution of social media content. To evaluate its impact, we conduct experiments with modality resampling. As shown in Tab. 8, resampling (the rare modalities by 2x and 3x), especially feature-level contribution resampling [40], can improve the understanding performance.

6. Conclusion and Discussion

We introduce MINE, a novel dataset capturing the inherent relationship between intention and emotion in natural social media settings. Unlike existing datasets that treat these aspects in isolation or rely on controlled environments, MINE preserves the authentic challenges of real-world social media data—particularly uncertain modalities and diverse thematic contexts. By providing detailed annotations for both intention and emotion across multiple modalities, MINE establishes a comprehensive benchmark for understanding human social behavior in naturalistic settings.

Method	Intention		
	Macro F1	Micro F1	Samples F1
Baseline	32.81	52.13	52.74
w/ AST [7]	33.57	52.95	53.34
	(+0.76)	(+0.82)	(+0.60)
w/ CLIP-ViT-S [27]	34.62	53.70	54.02
	(+1.81)	(+1.57)	(+1.28)
w/ SwinT-S [20]	33.14	52.67	52.98
	(+0.33)	(+0.54)	(+0.24)
w/ SemBERT [49]	33.89	53.55	54.02
	(+1.08)	(+1.42)	(+1.28)

Table 6. Impact of various feature extractors.

Model	Utilized modality		
	T, V	T, A	T
MuT [36]	37.15	38.28	41.91
w/ MAP	38.93 ^{+1.78}	39.14 ^{+0.86}	43.35 ^{+1.44}
MFN [45]	58.48	60.54	59.74
w/ MAP	61.33 ^{+2.85}	61.95 ^{+1.41}	63.72 ^{+3.98}
BERT_MAG [28]	62.83	61.39	62.83
w/ MAP	64.51 ^{+1.68}	62.55 ^{+1.16}	64.42 ^{+1.59}
LMF [19]	42.38	57.15	55.94
w/ MAP	44.25 ^{+1.87}	58.67 ^{+1.52}	57.12 ^{+1.18}

Table 7. Evaluation MAP across various architectures and modality combinations on CMU-MOSEI. T: Text, V: Visual, A: Audio.

Methods	ours	2*rare	3*rare	Wei et al [40]
Average	51.45	51.79	51.08	53.20

Table 8. Results of various methods for resisting modal imbalance.

To effectively process MINE’s complex multimodal data, we propose BEAR (Bridging Emotion-Intention via Implicit Label Reasoning). It addresses two critical challenges: the intrinsic relationship between emotion and intention, and the uncertainty of modalities. The BEAR employs an alignment task with masked emotion and intention categories, enabling implicit mining of their relationships. Meanwhile, the Modality Asynchronous Prompt handles missing modalities through targeted prompting that preserves the integrity of available features. Experimental results demonstrate BEAR’s effectiveness, outperforming existing methods across various metrics. In summary, MINE opens new avenues for understanding human social behavior through multimodal content.

Acknowledgements. This work is partially supported by the National Key Research and Development Program of China (2024YFC3308400), National Natural Science Foundation of China under Grant (62176188, 62361166629), Major Project of Science and Technology Innovation of Hubei Province (2024BCA003) and Key Project of the Health Commission of Hubei Province (WJ2023Z003).

References

- [1] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *NIPS*, 2020. 4
- [2] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 2008. 1, 2, 3, 4
- [3] Haozhe Chi, Minghua Yang, Junhao Zhu, Guan hong Wang, and Gaoang Wang. Missing modality meets meta sampling (m3s): An efficient universal approach for multimodal sentiment analysis with missing modality. *arXiv*, 2022. 2, 3, 6
- [4] Dorottya Demszy, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. Goemotions: A dataset of fine-grained emotions. *arXiv*, 2020. 3
- [5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv*, 2020. 4
- [6] Yiyang Fang, Wenke Huang, Guancheng Wan, Kehua Su, and Mang Ye. Emoe: Modality-specific enhanced dynamic emotion experts. In *CVPR*, 2025. 1
- [7] Yuan Gong, Yu-An Chung, and James Glass. Ast: Audio spectrogram transformer. *arXiv*, 2021. 8
- [8] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *ACM MM*, 2020. 6, 7
- [9] Charles T Hemphill, John J Godfrey, and George R Doddington. The atis spoken language systems pilot corpus. In *Speech and Natural Language: Proceedings of a Workshop*, 1990. 2
- [10] Ao Jia, Yu He, Yazhou Zhang, Sagar Uprety, Dawei Song, and Christina Lioma. Beyond emotion: A multi-modal dataset for human desire understanding. In *NACCL*, 2022. 3
- [11] Menglin Jia, Zuxuan Wu, Austin Reiter, Claire Cardie, Serge Belongie, and Ser-Nam Lim. Intentonomy: a dataset and study towards human intent understanding. In *CVPR*, 2021. 2, 3, 4, 6
- [12] Kawaljeet Kaur Kapoor, Kuttimani Tamilmani, Nripendra P Rana, Pushp Patil, Yogesh K Dwivedi, and Sridhar Nerur. Advances in social media research: Past, present and future. *Information Systems Frontiers*, 2018. 3
- [13] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NACCL*, 2019. 4
- [14] Julia Kruk, Jonah Lubin, Karan Sikka, Xiao Lin, Dan Jurafsky, and Ajay Divakaran. Integrating text and image: Determining multimodal document intent in Instagram posts. In *EMNLP-IJCNLP*, 2019. 3
- [15] Stefan Larson, Anish Mahendran, Joseph J Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K Kummerfeld, Kevin Leach, Michael A Laurenzano, Lingjia Tang, et al. An evaluation dataset for intent classification and out-of-scope prediction. In *EMNLP-IJCNLP*, 2019. 2, 3
- [16] Yi-Lun Lee, Yi-Hsuan Tsai, Wei-Chen Chiu, and Chen-Yu Lee. Multimodal prompting with missing modalities for visual recognition. In *CVPR*, 2023. 3, 6, 7
- [17] Rui Liu, Haolin Zuo, Zheng Lian, Xiaofen Xing, Björn W Schuller, and Haizhou Li. Emotion and intent joint understanding in multimodal conversation: A benchmarking dataset. *arXiv*, 2024. 2
- [18] Xingkun Liu, Arash Eshghi, Pawel Swietojanski, and Verena Rieser. Benchmarking natural language understanding services for building conversational agents. *arXiv*, 2019. 3
- [19] Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Efficient low-rank multimodal fusion with modality-specific factors. *arXiv*, 2018. 8
- [20] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *CVPR*, 2022. 4, 8
- [21] Meng Luo, Hao Fei, Bobo Li, Shengqiong Wu, Qian Liu, Soujanya Poria, Erik Cambria, Mong-Li Lee, and Wynne Hsu. Panosent: A panoptic sextuple extraction benchmark for multimodal conversational aspect-based sentiment analysis. In *ACM MM*, 2024. 2
- [22] Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. Smil: Multimodal learning with severely missing modality. In *AAAI*, 2021. 3
- [23] Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. Are multimodal transformers robust to missing modality? In *CVPR*, 2022. 3
- [24] Kuan-Chuan Peng, Amir Sadovnik, Andrew Gallagher, and Tsuhan Chen. Where do emotions come from? predicting the emotion stimuli map. In *IEEE ICIP*, 2016. 3
- [25] Wei Peng, Yue Hu, Yuqiang Xie, Luxi Xing, and Yajing Sun. Cogintac: Modeling the relationships between intention, emotion and action in interactive process from cognitive perspective. In *2022 IEEE Congress on Evolutionary Computation*, 2022. 1
- [26] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *ACL*, 2019. 2, 3
- [27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Aspell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 8
- [28] Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, AmirAli Bagher Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. Integrating multimodal information in large pretrained transformers. In *ACL*, 2020. 8
- [29] Md Kaykobad Reza, Ashley Prater-Bennette, and M Salman Asif. Robust multimodal learning with missing modalities via parameter-efficient adaptation. *IEEE TPAMI*, 2024. 3
- [30] Tulika Saha, Aditya Patra, Sriparna Saha, and Pushpak Bhat-tacharyya. Towards emotion-aided multi-modal dialogue act classification. In *ACL*, 2020. 1

- [31] Tulika Saha, Aditya Patra, Sriparna Saha, and Pushpak Bhattacharyya. Towards emotion-aided multi-modal dialogue act classification. In *ACL*, 2020. [2](#), [3](#)
- [32] Tobias Schröder, Terrence C Stewart, and Paul Thagard. Intention, emotion, and action: A neural theory based on semantic pointers. *Cognitive science*, 2014. [1](#), [4](#)
- [33] QingHongYa Shi, Mang Ye, Ziyi Zhang, and Bo Du. Learnable hierarchical label embedding and grouping for visual intention understanding. *IEEE Transactions on Affective Computing*, 2023. [6](#)
- [34] Mohammad Faridul Haque Siddiqui, Parashar Dhakal, Xiaoli Yang, and Ahmad Y Javaid. A survey on databases for multimodal emotion recognition and an introduction to the viri (visible and infrared image) database. *Multimodal Technologies and Interaction*, 2022. [4](#)
- [35] Gopendra Vikram Singh, Mauajama Firdaus, Asif Ekbal, and Pushpak Bhattacharyya. Emoint-trans: A multimodal transformer for identifying emotions and intents in social conversations. *IEEE/ACM TASLP*, 2022. [1](#), [2](#), [8](#)
- [36] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *ACL*, 2019. [4](#), [6](#), [7](#), [8](#)
- [37] Gokhan Tur, Dilek Hakkani-Tür, and Larry Heck. What is left to be understood in atis? In *IEEE Spoken Language Technology Workshop*, 2010. [2](#)
- [38] Chai M Tyng, Hafeez U Amin, Mohamad NM Saad, and Amir S Malik. The influences of emotion on learning and memory. *Frontiers in psychology*, 2017. [1](#)
- [39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017. [6](#)
- [40] Yake Wei, Ruoxuan Feng, Zihe Wang, and Di Hu. Enhancing multimodal cooperation via sample-level modality valuation. In *CVPR*, 2024. [8](#)
- [41] Dingkan Yang, Haopeng Kuang, Shuai Huang, and Lihua Zhang. Learning modality-specific and-agnostic representations for asynchronous multimodal language sequences. In *ACM MM*, 2022. [6](#), [7](#)
- [42] Qu Yang, Mang Ye, and Bo Du. Emollm: Multimodal emotional understanding meets large language models. *arXiv*, 2024. [1](#)
- [43] Qu Yang, Mang Ye, and Dacheng Tao. Synergy of sight and semantics: visual intention understanding with clip. In *ECCV*, 2024. [1](#)
- [44] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv*, 2016. [1](#)
- [45] Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Memory fusion network for multi-view sequential learning. In *AAAI*, 2018. [8](#)
- [46] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *ACL*, 2018. [1](#), [2](#), [3](#)
- [47] Hanlei Zhang, Hua Xu, Xin Wang, Qianrui Zhou, Shaojie Zhao, and Jiayan Teng. Mintrec: A new dataset for multimodal intent recognition. In *ACM MM*, 2022. [1](#), [2](#), [3](#), [4](#), [6](#)
- [48] Hanlei Zhang, Xin Wang, Hua Xu, Qianrui Zhou, Kai Gao, Jianhua Su, Wenrui Li, Yanting Chen, et al. Mintrec2. 0: A large-scale benchmark dataset for multimodal intent recognition and out-of-scope detection in conversations. In *ICLR*, 2024. [2](#), [3](#)
- [49] Zhuosheng Zhang, Yuwei Wu, Hai Zhao, Zuchao Li, Shuailiang Zhang, Xi Zhou, and Xiang Zhou. Semantics-aware bert for language understanding. In *AAAI*, 2020. [8](#)
- [50] Jinming Zhao, Ruichen Li, and Qin Jin. Missing modality imagination network for emotion recognition with uncertain missing modalities. In *ACL*, 2021. [2](#), [3](#), [5](#), [6](#), [7](#)
- [51] Qianrui Zhou, Hua Xu, Hao Li, Hanlei Zhang, Xiaohan Zhang, Yifan Wang, and Kai Gao. Token-level contrastive learning with modality-aware prompting for multimodal intent recognition. In *AAAI*, 2024. [7](#)