Dynamic Emotion and Personality Profiling for Multimodal Deception Detection

Li Zheng[1], **Yanyi Luo**[1], **Hao Fei**[2], **Yuzhe Ding**[1], **Yujie Huang**[1], **Fei Li**[1]*, **Chong Teng**[1], **Donghong Ji**[1]*

[1]Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education, School of Cyber Science and Engineering, Wuhan University, Wuhan, China

[2]National University of Singapore, Singapore, Singapore

{zhengli,lne.luoyanyi,yuzheding,huang-yj}@whu.edu.cn

{lifei_csnlp,tengchong,dhji}@whu.edu.cn, haofei7419@gmail.com

Abstract

Deception detection is of great significance for ensuring information security and conducting public opinion analysis, with personality factors and emotion cues playing a critical role. However, existing methods lack sample-level dynamic annotations for emotions and personality. In this paper, we propose an innovative multi-model multi-prompt annotation scheme and a strict label quality evaluation standard, and establish a multimodal joint detection dataset DDEP for deception, emotion, and personality. Meanwhile, we propose Rel-DDEP, an adaptive reliability-weighted fusion framework. Our framework quantifies uncertainty by mapping modal features to a high-dimensional Gaussian distribution space. It then performs reliability-weighted fusion and incorporates an alignment module and a sorting constraint module to achieve joint detection of deception, emotion, and personality. Experimental results on the MDPE and DDEP datasets show that our Rel-DDEP significantly outperforms the existing state-of-the-art baseline models in three tasks. The F1 score of the deception detection increases by 2.53%, that of the emotion detection increases by 2.66%, and that of the personality detection increases by 9.30%. The experiments fully verify the necessity of annotating dynamic emotion and personality labels for each sample and the effectiveness of reliability-weighted fusion.

1 Introduction

Deception detection aims to accurately identify deceptive behavior in individuals, which is critical for information security (Gutierrez et al., 2015; Han et al., 2018) and public opinion analysis (Alowibdi et al., 2014; Luke et al., 2023; Zheng et al., 2025a). Prior works have yielded notable progress in multimodal deception detection. Pérez-Rosas et al. (2015) introduce a multimodal deception dataset using real-world courtroom trial videos. Gupta et al. (2019) propose the Bag-of-Lies dataset for daily deceptive behavior detection. Vance et al. (2022) construct the multimodal deception database DDPM. Despite these advances, existing studies are confined to the single task of multimodal deception detection.

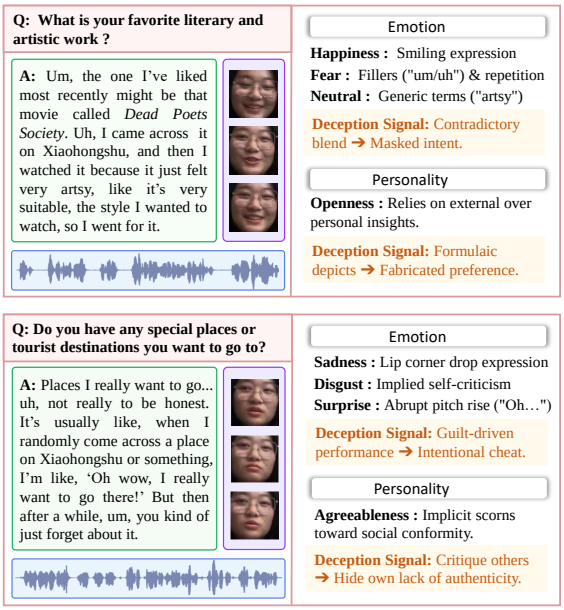


Figure 1: Examples of deceptive detection with different emotion and personality information.

Numerous studies (Gaspar and Schweitzer, 2013; Levitan et al., 2015; Zheng et al., 2025c) show that personality factors and emotion cues play important roles in deception detection. Personality factors affect an individual's behavior and ways of thinking, thus playing a role in deception detection. Emotion cues, whether they are genuine emotions or feigned emotions, can serve as key bases for detecting deception. Moreover, emotion and personality influence each other. Cai et al. (2024) propose a multimodal deception dataset with personality and emotion features. However, their dataset only provides subject-level (per-participant) annotations, failing to capture sample-level dynamics, the fact that an individual's emotions and personality can

*Corresponding author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 3930–3940

vary significantly across different situations.

As shown in Figure 1, the two examples correspond to the same subject. Subject-level annotation would assign identical emotion and personality features to both cases, yet their actual emotional and personality traits differ substantially. Furthermore, these discrepancies in emotion and personality provide critical cues for deception detection. The left example exhibits a blend of feigned happiness and fear of exposure, which is a hallmark of deceptive behavior. While the right example combines sadness and disgust, reflecting the subject's internal disapproval of their perfunctory actions and facilitating deception identification as a result. Thus, sample-specific annotation is imperative.

Considering the complexity of emotions in real-world scenarios where emotions aren't presented in a single form, we adopt multi-label annotation for emotions (See the detailed analysis in Section 2.). Since personality is relatively stable within a certain period, we set single-label annotation for personality. Notably, manual annotation is costly, a longstanding major challenge in data annotation (Cai et al., 2021; Hang et al., 2024; Zheng et al., 2024a). Recent advances have shown that large language models (LLMs) possess strong text understanding and generation capabilities, and are increasingly used to assist human annotation tasks (Ding et al., 2022; Wang et al., 2024). However, the application of LLMs in multimodal annotation is still in the exploratory stage, with no unified standards for verifying annotation quality.

In view of this, we propose *an innovative multi-model multi-prompt annotation scheme and a rigorous label quality evaluation standard*, and obtain the *Dynamic Deception-Emotion-Personality joint detection multimodal dataset* ***DDEP***. Specifically, we first employ multiple distinct LLMs for initial annotation and design diverse prompts for each model. These prompts guide LLMs to deeply understand data from varied perspectives, mitigating single perspective biases and reducing misjudgments. Subsequently, we design a voting mechanism and construct a quality scoring system incorporating consistency and uncertainty scores to comprehensively assess label quality. For data failing to meet the quality threshold, we leverage multimodal LLMs for advanced re-annotation followed by a second quality evaluation. Finally, data still not meeting requirements are re-annotated by professional human annotators.

We further propose ***Rel-DDEP***, *an adaptive reliability-weighted fusion framework for the joint multimodal detection of deception, emotion, and personality.* Specifically, we project each modality's features into a high-dimensional Gaussian distribution space to quantify uncertainty, then perform reliability-weighted fusion to assign rational weights to modalities (prioritizing highly reliable ones). Concurrently, we introduce an alignment module (to match uncertainty estimates with actual prediction errors) and a sorting constraint module (to ensure uncertainty estimates reflect modality importance order in joint detection). Finally, we derive the final predictions for emotion, personality, and deception.

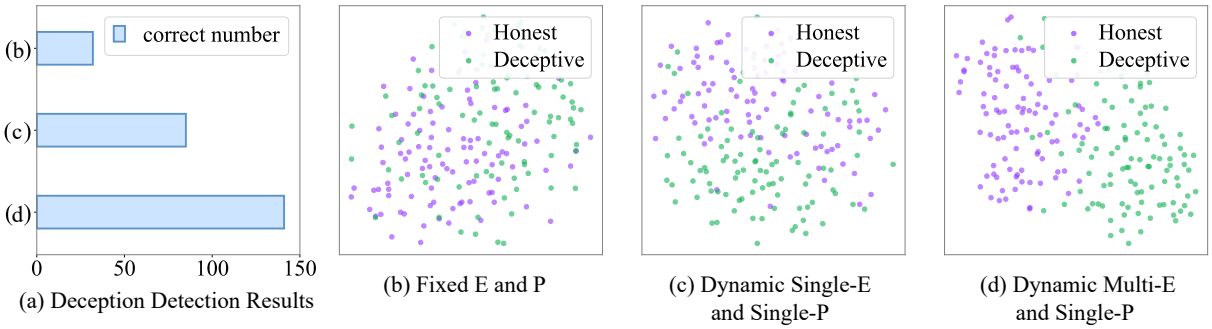
To verify the effectiveness of our model, we conduct experiments on the widely used multimodal deception dataset MDPE (Cai et al., 2024) and our DDEP dataset. The results show that our model significantly outperforms all state-of-the-art (SoTA) baselines on the three tasks of the two datasets. In the deception, emotion and personality detection task, the F1 score is increased by 2.53%, 2.66% and 9.30% respectively. Extensive experiments verify the necessity of annotating dynamic emotion and personality labels for each data and the effectiveness of reliability-weighted fusion.

Our main contributions are summarized as follows:

- We propose a multi-model multi-prompt annotation scheme and a rigorous label quality evaluation standard, and establish a multimodal dataset DDEP for the joint detection of deception, emotion, and personality.
- We propose an adaptive reliability-weighted fusion framework to fully leverage the advantages of modalities with high reliability.
- Our extensive experimentation on the MDPE and DDEP datasets demonstrate that our Rel-DDEP achieves SoTA performance.

2 Observation and Key Intuition

Previous deception detection study (Cai et al., 2024) incorporates emotion and personality information at the subject-level. However, in reality, the emotions and personalities of the same person vary significantly in different situations. To understand the impact of this variability on deception detection, we conduct a series of exploratory analyses.



(a) Deception Detection Results (b) Fixed E and P (c) Dynamic Single-E and Single-P (d) Dynamic Multi-E and Single-P

Figure 2: Visualization of different emotion (E) and personality (P) information.

We randomly select 20 subjects and choose 10 data instances with different emotion and personality characteristics for each subject. The deception detection results are shown in Figure 2 (a). When relying solely on the fixed subject-level emotion and personality information (b) for deception detection, only 32 instances are correctly detected. This low success rate indicates that such fixed information is insufficient for effective deception detection. The visualization experiment in Figure 2 (b) shows that deceptive and honest samples are intertwined in the visual space, with blurry boundaries that make them difficult to distinguish. This visual evidence suggests that fixed subject-level annotations fail to capture the nuances that are crucial for accurate deception detection.

These findings inspire us to consider an alternative: *annotating each data instance with single emotion and personality labels.* When applying this method to the same set of 200 samples, the number of correctly detected instances samples to 85, leading to an improvement in deception detection performance. The visualization results in Figure 2 (c) also indicate that the distinction between deceptive and honest samples is enhanced. This highlights the importance of annotating dynamic emotion and personality labels for individual data instances. However, there is still some overlap between deceptive and honest samples.

By observing the data, we find that the emotions in each sample are complex and diverse. For instance, the examples shown in Figure 1 contain multiple emotions. This indicates that a more comprehensive annotation method is required. Therefore, *we explore using multiple emotion labels for each sample while maintaining single-label annotations due to the relative stability of personality.* Implementing this new strategy on the set of 200 samples brings about significant improvements, with 141 samples correctly detected. Figure 2 (d) shows that deceptive and honest samples form distinct and compact clusters with concentrated feature distributions and clear boundaries. This result shows that dynamic multi-label emotion and single-label personality annotations can effectively capture sample-specific information, yielding more discriminative features for deception detection. We formalize these findings into two theorems.

Theorem 1. *Information Gain Improvement Theorem.* Let X_{fixed} be the feature set when using fixed labels, and X_{new} be the feature set after re-annotating the emotion and personality labels of each sample. Let Y be the category of deception detection results. Then, the information gain $IG(Y|X_{new}) > IG(Y|X_{fixed})$.

Theorem 2. *Situational Feature Difference Capture Theorem.* Let the sample s be in different situations c_i and c_j. With fixed labels, the feature difference degree $D_{fixed}(X_{s,c_i}, X_{s,c_j}) = 0$ (fixed labels ignore situational differences), and the degree after re-annotation is $D_{new}(X_{s,c_i}, X_{s,c_j})$. Then $D_{new}(X_{s,c_i}, X_{s,c_j}) > 0$.

3 Multi-Model and Multi-Prompt Data Annotation

The MDPE dataset (Cai et al., 2024) offers rich data for multimodal deception detection, yet it suffers from limitations in emotion and personality annotation: it only provides subject-level emotion and personality labels, lacking sample-specific dynamic annotations. To better characterize data features per Theorems 1 and 2, we propose an innovative multi-model multi-prompt annotation scheme to assign emotion and personality labels to each sample, along with a set of label quality evaluation criteria. The annotation process is shown in Figure 3.

3.1 Low-level Annotation

To comprehensively annotate the data, we select multiple LLMs of different types (e.g., GPT4o,

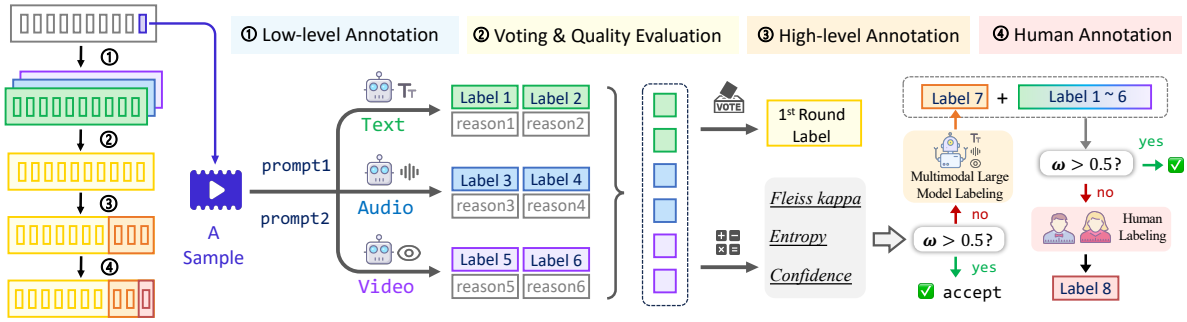


Figure 3: The flowchart of our data annotation. Label denote both emotion and personality.

Llama3, VideoLlama3, and Qwen2 Audio) for low-level annotation and design multiple prompting methods for each model. Through diverse prompts, we guide the models to understand the data from different perspectives, thereby generating more comprehensive annotation results. In the emotion annotation, one prompt guides the model to judge emotions from the overall atmosphere, as follows:

Prompt1: `Please determine the emotions of the characters based on the overall atmosphere presented.`

Another prompt focuses on the emotions reflected by specific behaviors or expressions:

Prompt2: `Please observe the facial expressions and body movements of the characters and determine their emotions.`

3.2 Voting Mechanism and Annotation Quality Score Evaluation

After generating annotations via the multi-model multi-prompt strategy, we derive the initial annotation round using a voting mechanism. For multi-label emotion annotation, we select labels with over half the votes as initial results. For personality labels, we choose the top-voted label if it garners more than half the votes. If the number of votes for each label does not meet the requirement of more than half of the votes, it is recorded as an object that requires high-level re-annotation. Our voting mechanism aggregates outputs from multiple models, effectively mitigating single-model errors and substantially boosting annotation reliability.

To ensure that the annotation quality meets a high standard, we construct a comprehensive quality score evaluation system, which is mainly composed of a consistency score and an uncertainty score. The consistency score is measured by the kappa coefficient (Cohen, 1960):

$$k = \frac{p_o - p_e}{1 - p_e} \tag{1}$$

where p_o represents the observed consistency, p_e represents the expected consistency. The value range of the Kappa coefficient is between -1 and 1.

The uncertainty score consists of two key metrics: entropy and self-evaluation confidence. Entropy is used to measure the degree of uncertainty of a model's data classification.

$$u_i = -\sum_{j=1}^{n} p_{ij} \log(p_{ij}) \tag{2}$$

where p_{ij} denotes the probability that the i-th sample belongs to the j-th class, and n is the total number of classes. When the model is confident in a sample's classification, the probability p_{ij} of the target class approaches 1 while those of other classes approach 0, yielding a low entropy value. Self-evaluation confidence refers to the model's confidence in its annotation outputs. Specifically, we require each LLM to assign a confidence score s_c to its annotations with corresponding explanations. By jointly considering the consistency score and uncertainty score, we derive the final quality score:

$$S_q = \alpha_1 k + \alpha_2 u_i + \alpha_3 s_c \tag{3}$$

We screen the annotation results according to the set quality threshold. Annotations with a quality score lower than the threshold are considered to have low reliability and are marked as objects that need to be re-annotated at a high-level.

3.3 High-level Annotation

For the data marked as requiring high-level re-annotation, we use the multimodal large model as a high-level annotator to review the low-level annotation results of these unimodal LLMs and

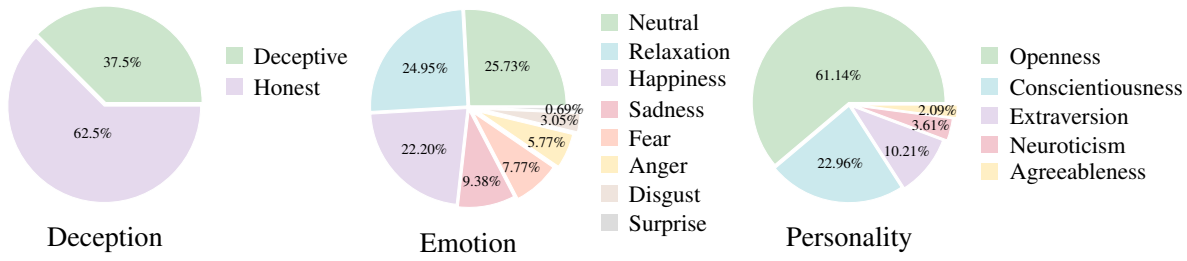


Figure 4: The data distribution of our DDEP dataset.

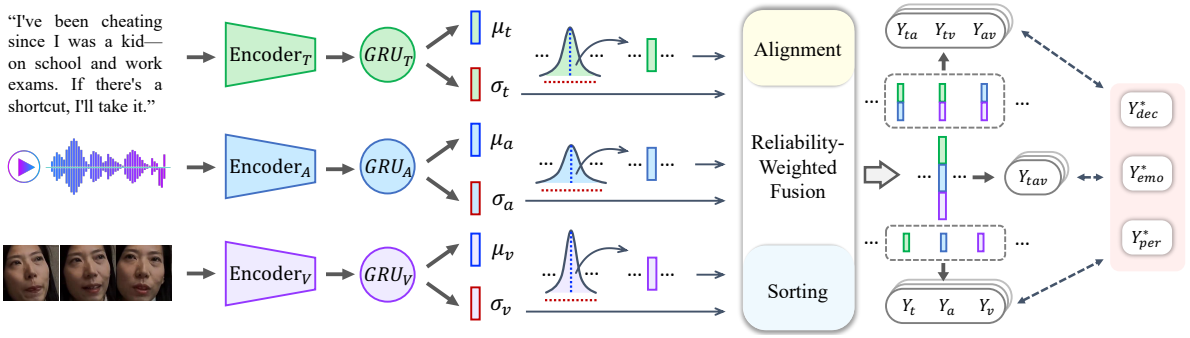


Figure 5: The overview of our framework.

provide new annotation results along with explanations. This model conducts a more comprehensive and in-depth understanding and annotation of the data from the perspective of multimodal fusion. Subsequently, we evaluate the quality of the annotation results to obtain the final quality scores. If the quality score is lower than the threshold, the data is marked as requiring re-annotation and then handed over to human annotators for processing.

3.4 Human Annotation

Five natural language processing and emotion analysis experts conduct re-annotation. During the re-annotation process, the annotators refer to the annotation results of LLMs. They repeatedly watch the videos and carefully read the texts, analyzing the emotion and personality information contained in the data from multiple perspectives to make comprehensive judgments. To ensure the consistency and accuracy of the annotations, cross-validation is carried out on the annotation results of human annotators. Each data is independently annotated by at least two experts. Their results are then compared, and discrepancies are resolved through discussion until a consensus is reached. After annotation, we calculate the kappa score, achieving a score of **0.85**. We finally obtain the DDEP dataset, with labels illustrated in Figure 4.

4 Adaptive Reliability-weighted Fusion Network

4.1 Feature Extraction

Text extraction. Text modal feature extraction aims to obtain key information such as semantics and context in the text. Following Cai et al. (2024), we adopt Baichuan (Bai) (Yang et al., 2023) for text feature extraction $\mathbf{h}_t$.

Video extraction. Following Cai et al. (2024), we use the multimodal models CLIP (CLB) (Radford et al., 2021) and ViT (Dosovitskiy et al., 2020) to extract video features $\mathbf{h}_v$.

Audio extraction. Following Cai et al. (2024), we select Wav2vec-base (W2B) (Baevski et al., 2020), HUBERT-base (HBB) (Hsu et al., 2021), and WavLM-base (WMB) (Chen et al., 2022) to extract audio features $\mathbf{h}_a$.

4.2 Uncertainty Estimation Module

In the multimodal joint detection of deception, emotion, and personality, given the complexity of multimodal data and the inherent uncertainty of model predictions, accurately estimating uncertainty is of great importance. For the features $\mathbf{h}_m$ ($m \in \{t, v, a\}$) extracted from each modality, we map them to a high-dimensional Gaussian distribution space $N(\mu_m, \sigma_m)$ to quantify the uncertainty.

$$\mu_m = GRU_\mu(\mathbf{h}_m), \sigma_m = GRU_\sigma(\mathbf{h}_m) \quad (4)$$

Task	Model	Acc.	AUC	P	R	F1
Deception	ViT-HBB-Bai-MDPE	64.00	67.50	47.15	52.91	49.86
	ViT-HBB-Bai-DDEP	65.12	68.22	59.26	57.82	57.57
	ViT-HBB-Bai-Ours	66.35	69.45	61.06	59.41	59.30
	ViT-WMB-Bai-MDPE	63.59	67.20	46.89	54.35	50.35
	ViT-WMB-Bai-DDEP	64.78	67.93	59.84	58.55	58.45
	ViT-WMB-Bai-Ours	65.42	69.15	61.07	60.10	60.19
	CLB-HBB-Bai-MDPE	64.66	68.70	49.48	52.33	50.87
	CLB-HBB-Bai-DDEP	65.72	69.68	59.32	58.33	58.30
	CLB-HBB-Bai-Ours	**66.63**	**70.21**	**61.49**	**60.71**	**60.83**
Emotion	ViT-HBB-Bai	77.16	73.17	57.58	58.95	58.26
	ViT-HBB-Bai-Ours	80.23	79.02	60.71	60.43	60.57
	ViT-WMB-Bai	78.39	75.25	57.82	59.37	58.58
	ViT-WMB-Bai-Ours	80.48	78.51	60.97	61.26	61.11
	CLB-HBB-Bai	77.32	74.04	58.01	59.62	58.81
	CLB-HBB-Bai-Ours	**80.73**	**79.49**	**61.16**	**61.78**	**61.47**
Personality	ViT-HBB-Bai	82.51	72.11	39.51	38.22	38.98
	ViT-HBB-Bai-Ours	85.55	83.41	51.01	47.23	49.15
	ViT-WMB-Bai	83.89	74.91	40.78	39.89	40.30
	ViT-WMB-Bai-Ours	85.58	82.99	51.63	47.55	49.71
	CLB-HBB-Bai	83.07	75.02	40.92	40.33	40.60
	CLB-HBB-Bai-Ours	**85.93**	**83.81**	**51.81**	**47.86**	**49.90**

Table 1: Experimental results on deception, emotion, and personality detection tasks.

This uncertainty estimation module clarifies model prediction reliability, enabling rational weighting and fusion based on per-modality uncertainty.

4.3 Reliability-weighted Fusion Module

Considering the differences in reliability of different modalities in the task, we introduce an reliability-weighted fusion module. Based on the uncertainty estimation σ_m of each modality, we calculate the fusion weight w_m.

$$w_m = \frac{\frac{1}{\sigma_m}}{\sum_{j=t,v,a} \frac{1}{\sigma_j}} \tag{5}$$

This formula ensures that the modality with lower uncertainty has a higher fusion weight. Then, we calculate the fused feature $\mathbf{h}_f$ by taking a weighted sum of each modality's features $\mathbf{h}_m$.

$$\mathbf{h}_f = w_t \mathbf{h}_t + w_v \mathbf{h}_v + w_a \mathbf{h}_a \tag{6}$$

4.4 Alignment Module

The alignment module aims to align the model's uncertainty estimation with its actual prediction error. We define the alignment loss function L_{ali}, and use the mean squared error (MSE) to measure the difference between the uncertainty estimation σ_m and the prediction error ϵ_m. For each modality m, the prediction error ϵ_m is calculated by the cross-entropy loss between the predicted probability distribution $\mathbf{p}_m$ and the one-hot encoded distribution of the ground truth label y, that is:

$$\begin{aligned} \epsilon_m &= CrossEntropyLoss(\mathbf{p}_m, \mathbf{y}_{one-hot}) \\ L_{ali} &= \sum_{m=t,v,a} MSE(\sigma_m, \epsilon_m) \end{aligned} \tag{7}$$

4.5 Sorting Constraint Module

The sorting constraint module enforces consistency between the uncertainty estimates of different modalities and the ranking of their fusion weights. We define the ordering loss L_{ord}, constructed by comparing the uncertainty estimates σ_m and σ_j of distinct modalities against the relative order of their corresponding fusion weights w_m and w_j $(m, j \in \{t, v, a\})$.

$$L_{sor} = \sum_{m \neq j} max(0, (\sigma_m - \sigma_j) - (\beta(w_m - w_j))) \tag{8}$$

4.6 Prediction and Training

The fused feature $\mathbf{h}_f$ passes through three fully connected layers FC_{final} and a softmax function, obtaining final predictions $\hat{y}_d$, $\hat{y}_e$, $\hat{y}_p$ for deception, emotion, and personality tasks.

$$\hat{y}_{d/e/p} = softmax(FC_{final}(\mathbf{h}_f^{d/e/p})) \tag{9}$$

We employ the cross-entropy loss function L_{cls} to measure the difference between the final prediction result $\hat{y}_{d/e/p}$ and the ground-truth label $y_{d/e/p}$. By comprehensively considering the classification

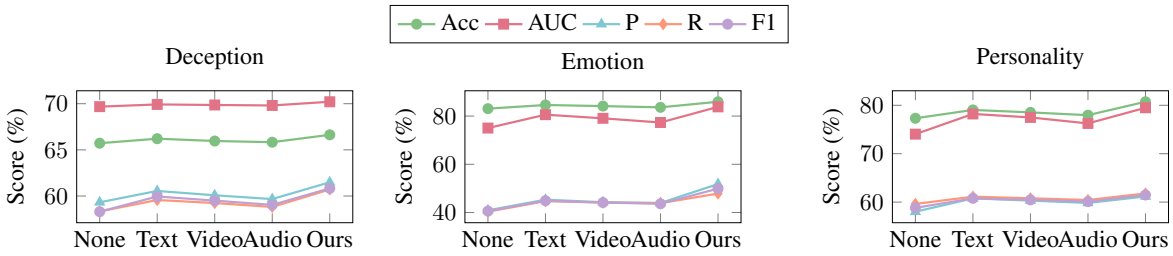


Figure 6: Influence of the reliability-weight fusion mechanism on the DDEP dataset.

loss, alignment loss, and sorting loss, the overall loss function is formulated as:

$$L = L_{cls} + \lambda_1 L_{ali} + \lambda_2 L_{sor} \tag{10}$$

5 Experiments

5.1 Experimental Setup

Dataset. We evaluate the effectiveness of our model on the widely used MDPE dataset (Cai et al., 2024) and our newly established DDEP dataset.

Baseline Systems. To validate the effectiveness of our model, we compare it with SoTA baselines tailored to each task. **Deception Detection**: Two baseline categories are defined by dataset and feature strategy: *MDPE-based baselines*: Fixed emotion/personality feature models on the MDPE dataset (`-MDPE` in Table 1). *DDEP-based baselines*: Dynamic emotion/personality feature models on our self-constructed DDEP dataset with only unimodal encoders (`-DDEP` in Table 1). **Emotion and Personality Detection**: All experiments are performed on the DDEP dataset (MDPE lacks corresponding labels), using competitive unimodal encoder-based models as baselines (details in Section 4.1). For all tasks, `-Ours` in Table 1 denotes DDEP-evaluated models integrated with our reliability-weight fusion strategy.

5.2 Main Results

We conduct comprehensive experiments on the MDPE and DDEP datasets to systematically evaluate our method against SoTA baselines. As shown in Table 1, our method outperforms the SoTA baselines across all three tasks. In the deception detection task, our method with dynamic emotion and personality labels far outperforms the fixed emotion and personality features at the subject-level. The F1 score increases by 2.53%. This indicates that by annotating each sample with personalized emotion and personality labels, more clues related to deceptive behavior can be unearthed, thus significantly enhancing the performance of deception detection. In emotion detection, our model shows a 2.66% increase in the F1 score compared to the SoTA. In personality detection, the improvement of our model is even more remarkable. Our model has a 9.3% increase in the F1 score compared to the SoTA baseline. This demonstrates that our reliability-weighted fusion can more rationally integrate multimodal data and effectively capture the key information related to emotions and personalities in different modalities.

Task	Model	Acc	AUC	P	R	F1
Dep	Ours	**66.63**	**70.21**	**61.49**	**60.71**	**60.83**
	w/o Ali	66.32	70.04	60.89	60.24	60.58
	w/o Sort	66.21	69.97	60.53	59.86	60.15
	w/o Rel	65.97	69.85	60.19	59.53	59.66
Emo	Ours	**80.73**	**79.49**	**61.16**	**61.78**	**61.47**
	w/o Ali	79.68	77.92	60.41	61.12	60.59
	w/o Sort	79.12	77.03	60.26	60.96	60.43
	w/o Rel	78.74	76.69	59.84	60.31	60.10
Per	Ours	**85.93**	**83.81**	**51.81**	**47.86**	**49.90**
	w/o Ali	85.29	80.68	49.62	45.91	47.60
	w/o Sort	84.81	79.25	48.92	45.28	47.11
	w/o Rel	84.47	78.31	46.79	43.84	45.21

Table 2: Ablation study on multimodal deception detection task.

5.3 Ablation Study

To further explore the contributions of each component of our model, we conduct ablation experiments by removing alignment module, sorting constraint module, and reliability-weighted fusion module respectively. The results are shown in Table 2. Removing the alignment module causes consistent performance degradation across all metrics, underscoring its role in aligning uncertainty estimates with actual prediction errors. Similarly, eliminating the sorting constraint module reduces performance, demonstrating its necessity for ensuring modality uncertainty estimates reflect their importance in deception detection. Removing the reliability-weighted fusion module leads to the most significant decline, this highlights the module's ability to

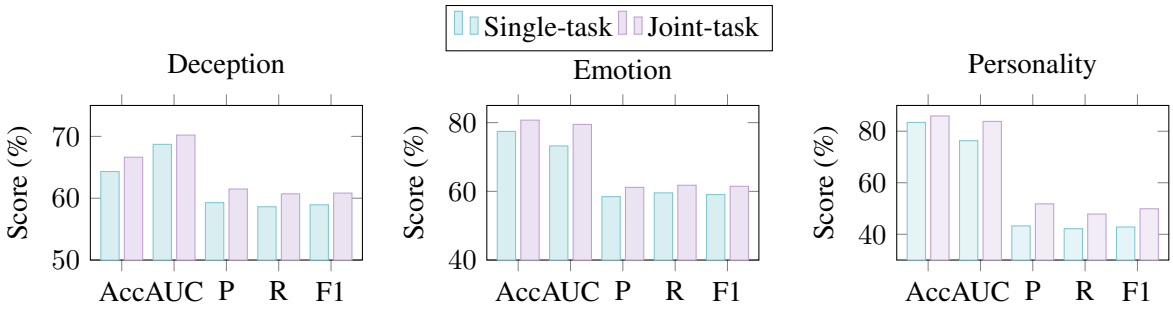


Figure 7: Comparative results of joint task and single task.

assign weights based on modality reliability, effectively emphasizing high reliability information.

5.4 The Impact of Reliability-Weighted Fusion Module

To evaluate the impact of the reliability-weighted fusion module, Figure 6 compares five strategies: reliability-weighted, no weights, and modality-dominant (text/audio/video) approaches. No weights fusion performs poorest, as equal treatment of modalities ignores their varying contributions, leading to information loss. Among modality-dominant baselines, text dominance yields the best results compared to audio dominant and video dominant approaches. This indicates that text plays a dominant role in the detection of these three tasks. However, compared with the fusion approach using reliability weights, fixing a single modality as dominant still results in poor performance. Reliability-weighted dynamically assigns weights based on modality reliability, maximizing complementary strengths and enhancing robustness.

5.5 The Effectiveness of Joint Tasks

We design a set of single-task experiments for each task to thoroughly explore the effectiveness of the joint tasks. The experimental results are shown in Figure 7. The results demonstrate that, in the three tasks of deception detection, emotion detection, and personality detection, the performance of the joint task significantly outperforms that of single tasks. This indicates that there are close inherent connections and synergistic effects among emotions, personality, and deception. In the joint-task mode, the information related to each task can complement and promote one another, thereby enhancing the overall performance.

6 Related Work

Deception Detection. Deceptive behaviors are widespread in numerous fields such as social interactions, business, and security (Damstra et al., 2021; Kumar et al., 2021; Zheng et al., 2025b,d). Abouelenien et al. (2016) integrate multiple modalities of information to construct a comprehensive deception detection system. Mathur and Matarić (2020) focus on analyzing the discriminative abilities of visual, speech, and language modality features and explore their impact mechanisms on deception detection. Regarding dataset construction, previous studies (Gupta et al., 2019; Pérez-Rosas et al., 2015; Vance et al., 2022) have contributed a series of valuable deception detection datasets. Although these researches achievements are remarkable, they only focus on the single task of deception. Considering that personality factors and emotion cues play crucial roles in deception detection, Cai et al. (2024) propose a multimodal deception dataset with personality and emotion features.

LLMs Assisted Data Annotation. The rapid development of machine learning highly depends on a large amount of labeled training data. Yet, annotation is costly and time-consuming. A promising solution combines automatic pre-annotation with human refinement. Wang et al. (2021) re-annotate low-confidence LLM-generated instances, while (Gilardi et al., 2023; He et al., 2023; Zheng et al., 2024b) show ChatGPT outperforms humans in multiple tasks, validating LLMs' task-specific strengths. Ding et al. (2022) use GPT-3 as an annotator for classification and token-level task experiments. Wang et al. (2024) propose a multi-stage human-LLM collaboration method to optimize annotation workflows. Prior works (Kneusel and Mozer, 2017; Lai and Tan, 2019) further confirm that model outputs enhance human decision-making, highlighting LLMs' annotation support value.

7 Conclusion

In this paper, we propose an innovative method for joint detection of multimodal deception, emotion, and personality, which integrates dynamic emotion and personality annotation and an adaptive reliability-weighted fusion framework. By implementing a multi-model multi-prompt annotation strategy and a label quality evaluation standard, we construct the high-quality DDEP dataset. Our Rel-DDEP quantifies uncertainties and conducts reliability-weighted fusion, effectively enhancing the performance of the model. Numerous experiments on the MDPE and our DDEP datasets show that our framework achieves state-of-the-art results.

Limitations

Our work uses a basic shared encoder for the joint recognition of emotion, personality, and deception tasks. While this design ensures computational efficiency and simplifies training by reusing cross-task feature extraction layers, it inherently limits deep interactions among the three tasks. Specifically, the shared encoder learns general-purpose features uniformly applied to all tasks, failing to capture task-specific nuances and context-aware cross-task correlations. In future research, we will explore advanced multi-task learning frameworks to explicitly model task-specific dependencies, dynamically allocate feature extraction resources, and enhance adaptability to diverse task interactions, thereby further improving joint recognition performance.

Acknowledgments

This work was funded by Kuaishou. Fei Li and Donghong Ji are co-corresponding authors.

References

Mohamed Abouelenien, Verónica Pérez-Rosas, Rada Mihalcea, and Mihai Burzo. 2016. Detecting deceptive behavior via integration of discriminative features from multiple modalities. *IEEE Transactions on Information Forensics and Security*, 12(5):1042–1055.

Jalal S Alowibdi, Ugo A Buy, S Yu Philip, and Leon Stenneth. 2014. Detecting deception in online social networks. In *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014)*, pages 383–390. IEEE.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Cong Cai, Shan Liang, Xuefei Liu, Kang Zhu, Zhengqi Wen, Jianhua Tao, Heng Xie, Jizhou Cui, Yiming Ma, Zhenhua Cheng, et al. 2024. Mdpe: A multimodal deception dataset with personality and emotional characteristics. *Advances in Neural Information Processing Systems*.

Lile Cai, Xun Xu, Jun Hao Liew, and Chuan Sheng Foo. 2021. Revisiting superpixels for active learning in semantic segmentation with realistic annotation costs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10988–10997.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Alyt Damstra, Hajo G Boomgaarden, Elena Broda, Elina Lindgren, Jesper Strömbäck, Yariv Tsfati, and Rens Vliegenthart. 2021. What does fake look like? a review of the literature on intentional deception in the news and on social media. *Journalism Studies*, 22(14):1947–1963.

Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Shafiq Joty, Boyang Li, and Lidong Bing. 2022. Is gpt-3 a good data annotator? *arXiv preprint arXiv:2212.10450*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Joseph P Gaspar and Maurice E Schweitzer. 2013. The emotion deception model: A review of deception in negotiation and the role of emotion in deception. *Negotiation and Conflict Management Research*, 6(3):160–179.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Viresh Gupta, Mohit Agarwal, Manik Arora, Tanmoy Chakraborty, Richa Singh, and Mayank Vatsa. 2019. Bag-of-lies: A multimodal dataset for deception detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0.

Christopher N Gutierrez, Saurabh Bagchi, H Mohammed, and Jeff Avery. 2015. Modeling deception in information security as a hypergame–a primer. In *Proceedings of the 16th Annual Information Security Symposium*, page 41. CERIAS-Purdue University.

Xiao Han, Nizar Kheir, and Davide Balzarotti. 2018. Deception techniques in computer security: A research perspective. *ACM Computing Surveys (CSUR)*, 51(4):1–36.

Jinglue Hang, Xiangbo Lin, Tianqiang Zhu, Xuanheng Li, Rina Wu, Xiaohong Ma, and Yi Sun. 2024. Dexfuncgrasp: A robotic dexterous functional grasp dataset constructed from a cost-effective real-simulation annotation system. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 10306–10313.

Xingwei He, Zhenghao Lin, Yeyun Gong, Alex Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, Weizhu Chen, et al. 2023. Annollm: Making large language models to be better crowdsourced annotators. *arXiv preprint arXiv:2303.16854*.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460.

Ronald T Kneusel and Michael C Mozer. 2017. Improving human-machine cooperative visual search with soft highlighting. *ACM Transactions on Applied Perception (TAP)*, 15(1):1–21.

Srijan Kumar, Chongyang Bai, VS Subrahmanian, and Jure Leskovec. 2021. Deception detection in group video conversations using dynamic interaction networks. In *Proceedings of the international AAAI conference on web and social media*, volume 15, pages 339–350.

Vivian Lai and Chenhao Tan. 2019. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 29–38.

Sarah I Levitan, Guzhen An, Mandi Wang, Gideon Mendels, Julia Hirschberg, Michelle Levine, and Andrew Rosenberg. 2015. Cross-cultural production and detection of deception from speech. In *Proceedings of the 2015 ACM on workshop on multimodal deception detection*, pages 1–8.

Timothy J Luke, Erik Mac Giolla, Amina Memon, Sara Landström, Pär Anders Granhag, and Saul Kassin. 2023. What have we learned about cues to deception? a survey of expert opinions. *Psychology, Crime & Law*, pages 1–20.

Leena Mathur and Maja J Matarić. 2020. Introducing representations of facial affect in automated multimodal deception detection. In *Proceedings of the 2020 international conference on multimodal interaction*, pages 305–314.

Verónica Pérez-Rosas, Mohamed Abouelenien, Rada Mihalcea, Yao Xiao, CJ Linton, and Mihai Burzo. 2015. Verbal and nonverbal clues for real-life deception detection. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2336–2346.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Nathan Vance, Jeremy Speth, Siamul Khan, Adam Czajka, Kevin W Bowyer, Diane Wright, and Patrick Flynn. 2022. Deception detection and remote physiological monitoring: A dataset and baseline experimental results. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 4(4):522–532.

Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. Want to reduce labeling cost? gpt-3 can help. *arXiv preprint arXiv:2108.13487*.

Xinru Wang, Hannah Kim, Sajjadur Rahman, Kushan Mitra, and Zhengjie Miao. 2024. Human-llm collaborative annotation through effective verification of llm labels. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–21.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Li Zheng, Boyu Chen, Hao Fei, Fei Li, Shengqiong Wu, Lizi Liao, Donghong Ji, and Chong Teng. 2024a. Self-adaptive fine-grained multi-modal data augmentation for semi-supervised muti-modal coreference resolution. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 8576–8585.

Li Zheng, Hao Fei, Ting Dai, Zuquan Peng, Fei Li, Huisheng Ma, Chong Teng, and Donghong Ji. 2025a. Multi-granular multimodal clue fusion for meme understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 26057–26065.

Li Zheng, Hao Fei, Fei Li, Bobo Li, Lizi Liao, Donghong Ji, and Chong Teng. 2024b. Reverse multi-choice dialogue commonsense inference with graph-of-thought. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19688–19696.

Li Zheng, Donghong Ji, Fei Li, Hao Fei, Shengqiong Wu, Jingye Li, Bobo Li, and Chong Teng. 2025b. Ecqed: emotion-cause quadruple extraction in dialogs. *IEEE Transactions on Audio, Speech and Language Processing*.

Li Zheng, Tengyue Song, Yuzhe Ding, Xiaorui Wu, Fei Li, Dongdong Xie, Jinbo Li, Chong Teng, and Donghong Ji. 2025c. Improving emotion and intent understanding in multimodal conversations with progressive interaction. *IEEE Transactions on Affective Computing*.

Li Zheng, Sihang Wang, Hao Fei, Zuquan Peng, Fei Li, Jianming Fu, Chong Teng, and Donghong Ji. 2025d. Enhancing hyperbole and metaphor detection with their bidirectional dynamic interaction and emotion knowledge. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL'25)*.