MDFDU: A Multi-Task Deceptive-Factual Dialogue Understanding Framework

Yanqing Liu, Zhong Qian*, Peifeng Li, Qiaoming Zhu
School of Computer Science and Technology, *Soochow University*, Suzhou, China
20235227053@stu.suda.edu.cn, {qianzhong, pfli, qmzhu}@suda.edu.cn

***Abstract*—Trustworthiness in dialogue systems fundamentally relies on Deception Detection in Dialogue (DDD), defined as the discernment of subjective deceptive intent, and Dialogue-based Fact-Checking (DialFC), which entails verifying objective factual accuracy. However, existing methodologies predominantly treat these tasks in isolation, neglecting their shared mechanism of detecting information incongruence. This separation impedes generalization, particularly within the DDD domain, where data scarcity frequently leads to overfitting. To mitigate these limitations, we propose Multi-Task Deceptive-Factual Dialogue Understanding (MDFDU), a unified framework designed to jointly model subjective and objective trustworthiness. Key to our approach is a Prompt-Modulated Shared Reasoning mechanism, which leverages prompts generated by Large Language Models (LLMs) to dynamically bias the attention of a shared encoder, effectively decoupling task-specific semantics from shared reasoning capabilities. Additionally, we incorporate an uncertainty-weighted consistency regularization to align the latent representations of semantically corresponding labels (i.e., mapping "Lie" to "Refutes" and "Truth" to "Supported"), thereby enabling transfer from the resource-rich DialFC task to the data-constrained DDD domain. Extensive experiments on the Box of Lies and DialFact benchmarks validate the efficacy of our method, demonstrating that MDFDU achieves state-of-the-art performance on both tasks.**

***Index Terms*—trustworthiness, Deception Detection in Dialogue, Dialogue-based Fact-Checking, Prompt-Modulated Shared Reasoning**

I. Introduction

Ensuring trustworthiness in dialogue systems requires addressing two complementary challenges: discerning subjective deceptive intent and verifying objective factual accuracy. This reliability depends on two distinct yet interconnected tasks: **Deception Detection in Dialogue (DDD)**, which analyzes multimodal behavioral cues to identify intentional falsification [1], and **Dialogue-based Fact-Checking (DialFC)**, which evaluates the logical consistency of claims against external evidence [2]. While DDD focuses on the *how*—detecting incongruence among vocal, visual, and verbal signals—DialFC focuses on the *what*—identifying contradictions between statements and established facts.

Figure 1 illustrates this distinction. In the DDD scenario (Left), a speaker describes an object as *"Inside my box, it is very red"* yet displays micro-expressions of hesitation and gaze aversion. Although the sentence is linguistically fluent, the **incongruence** between verbal confidence and non-verbal anxiety signals deceptive intent. Conversely, in the DialFC scenario (Right), a speaker confidently claims, *"Rock and Hotel Music was released back in 1955."* However, retrieved evidence states the release date was *"January 1956"* revealing a factual **contradiction**. Despite differences in modality (behavioral vs. textual) and scope (subjective vs. objective), both tasks fundamentally rely on detecting **inconsistent information**—either across modalities or against external knowledge.

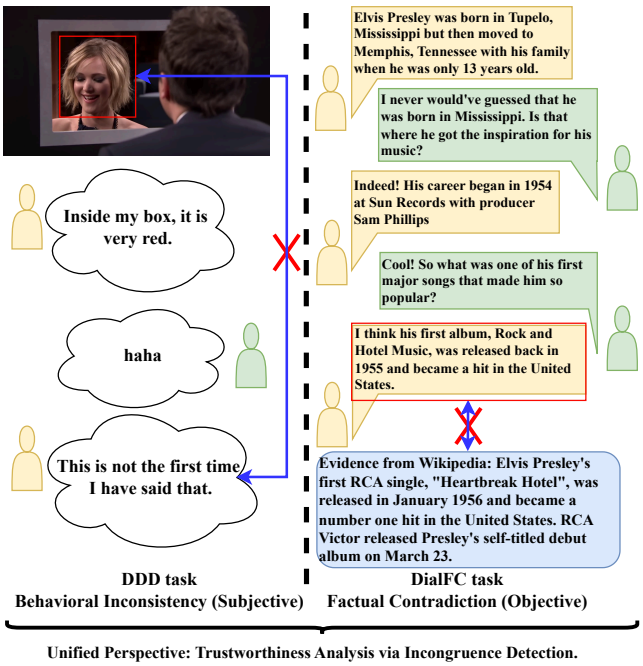

Fig. 1. Examples for dialogue trustworthiness. Despite differing modalities, both tasks fundamentally require reasoning about *incongruence*—either cross-modal or evidence-based.

Current methodologies predominantly treat these tasks as independent problems. DDD approaches typically employ modality fusion networks [3]–[5] or rely on handcrafted features [6] to capture behavioral anomalies. While effective on specific benchmarks, these models often function as pattern matchers, overfitting to surface-level heuristics (e.g., nervousness) rather than reasoning about semantic conflict. Similarly, DialFC methods often adapt textual entailment architectures originally designed for static claims [7]–[9], frequently neglecting the pragmatic nuances of conversational context.

This separation introduces two critical bottlenecks. First, it ignores the shared cognitive basis of *incongruence*. Both tasks fundamentally rely on detecting contradictions, whether

between modalities (DDD) or against evidence (DialFC). Second, significant data disparity exists between domains. While DialFC is data-rich, DDD suffers from scarcity and overfitting risks. Segregating these tasks prevents transferring the clearer reasoning signals from fact-checking to enhance the generalization of deception detection.

However, unifying these tasks introduces significant challenges, particularly the heterogeneity of modalities and the risk of negative transfer during joint optimization. To address these issues, we propose Multi-Task Deceptive-Factual Dialogue Understanding (MDFDU). Our methodology draws inspiration from recent unified representation paradigms like Visual Autoregressive Modeling (VAR) [10], which maps multi-scale visual predictions into a unified latent space. We adapt these insights specifically for the trustworthy dialogue domain to enable robust cross-task learning

First, inspired by Parameter-Efficient Fine-Tuning [11], we introduce a Prompt-Modulated Shared Reasoning mechanism. Unlike standard textual prompt tuning, we extend this paradigm to the multimodal domain. We utilize LLM-generated prompts as dynamic control signals for attention modulation to modulate the attention biases of a Shared Reasoning Encoder. This design decouples task-specific semantics from shared reasoning, enabling positive transfer while avoiding the interference common in hard parameter sharing. Second, building on uncertainty-weighted optimization [12], we integrate a novel cross-task consistency regularization. We hypothesize that "deception" and "refutation" share structural alignment in the latent space. By enforcing consistency between these representations (weighted by task uncertainty), we transfer structural knowledge from the resource-rich DialFC task to the data-scarce DDD task, effectively mitigating data scarcity. Our code will be available at https://anonymous.4open.science/r/MT-1455.

Our main contributions are summarized as follows:

- We propose MDFDU, a unified framework that bridges DDD and DialFC. To the best of our knowledge, this is the first work to explicitly model the shared cognitive mechanism of *incongruence* across these distinct domains.
- We introduce a Prompt-Modulated Shared Reasoning mechanism that employs adaptive attention biases to decouple task-specific semantics from shared reasoning, facilitating robust cross-task transfer.
- We design an uncertainty-aware optimization strategy with consistency regularization, which effectively mitigates negative transfer and achieves state-of-the-art accuracy on both the Box of Lies and DialFact benchmarks.

II. Related Work

a) DDD: Early research employed handcrafted features (e.g., gaze, acoustic jitter) [6], [13], while recent deep fusion methods [3], [14] target latent cross-modal signals. However, the scarcity of authentic deception data frequently causes overfitting to surface heuristics.

b) DialFC: Distinct from static verification [7], DialFC targets dynamic conversations [2]. Current methods adapt textual entailment or retrieval frameworks [8], [15] but treat verification in isolation, ignoring speaker intent.

c) Multi-Task Learning (MTL) and Prompt Tuning (PT): While MTL risks negative transfer from conflicting gradients [16], uncertainty weighting [12] effectively mitigates this issue. Concurrently, Prompt Tuning (PT) [11] efficiently adapts frozen backbones.

III. Method

We propose MDFDU, a unified framework for joint DDD and DialFC. Our approach is based on the hypothesis that while these tasks differ in output semantics, they share a fundamental underlying mechanism: the identification of *incongruence*—whether between multimodal behavioral cues (in DDD) or between claims and evidence (in DialFC). MDFDU unifies these domains by decoupling task-specific semantic priors from a shared reasoning engine. As illustrated in Fig. 2, the framework comprises three core components: (1) Modality-Aware Prompt Conditioning (Section III-A), (2) a Shared Reasoning Encoder (SRE, Section III-B) , and (3) Uncertainty-Weighted MTL (Section III-C and III-D).

A. Modality-Aware Prompt Conditioning

To adapt the model to the distinct requirements of DDD (denoted as task D) and DialFC (task F), we employ static task-specific prompts (P_t), pre-generated by an (Large Language Model) LLM, to inject task-specific reasoning priors. Let $t \in \{D, F\}$. We generate a structured prompt P_t focusing on distinct cues: P_D targets intent and nonverbal alignment (e.g., *"Do facial expressions contradict the statement?"*), while P_F targets logical coherence (e.g., *"Does the evidence support the claim reasoning?"*).

We formulate distinct conditioning strategies to handle the heterogeneity of textual and visual modalities:

a) Textual Prefix Tuning: For the dialogue text x_T, we treat the prompt P_t as a prefix sequence. The text encoder processes the concatenated sequence $[P_t; x_T]$, allowing the self-attention mechanism to directly model token-level interactions between the reasoning instructions and the dialogue content:

$$E_T = \text{TextEncoder}([P_t : x_T]) \tag{1}$$

b) Visual Cross-Attention: Visual inputs x_V (present only in DDD) are processed by a frozen vision encoder to yield feature map V. To avoid semantic misalignment between discrete text and continuous visuals, we project the prompt P_t into a latent vector p_t via the text encoder. We then employ a cross-attention mechanism where the prompt acts as the query to extract task-relevant regions from the visual features:

$$p_t = \text{TextEncoder}(P_t), \quad V_{\text{raw}} = \text{VisionEncoder}(x_V) \tag{2}$$

$$E_V = \text{Attention}(Q = p_t, K = V_{\text{raw}}, V = V_{\text{raw}}) \tag{3}$$

This formulation ensures that the visual representation E_V is dynamically filtered to emphasize regions relevant to the

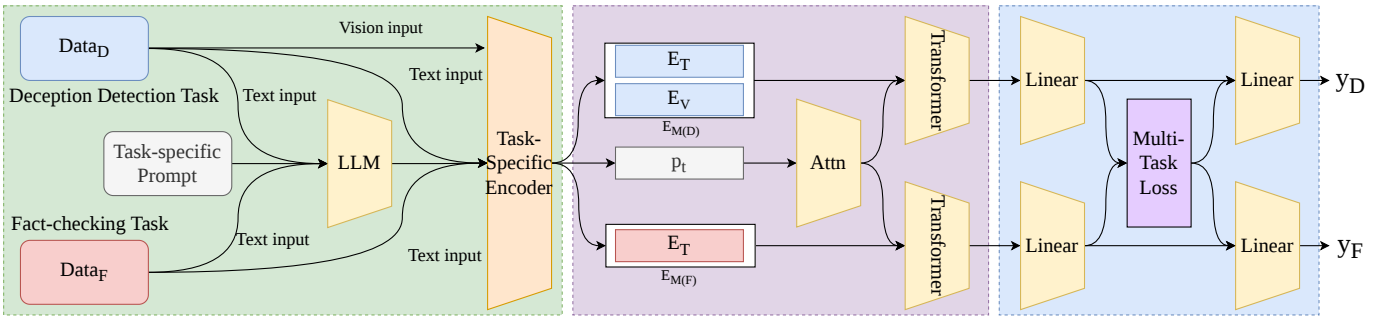

Fig. 2. **Architecture of the MDFDU framework. Task DDD denoted as D, task DialFC denoted as F.** The model utilizes an LLM to generate task-specific prompts (P_t) which condition the feature encoding process. Textual inputs use prefix-tuning, while visual inputs utilize prompt-to-image cross-attention (Modality-Aware Prompt Conditioning, Section III-A). A Shared Reasoning Encoder (SRE, Section III-B, E denote inputs after encoding) processes the multimodal features, modulated by a learnable prompt bias to toggle between task-specific reasoning modes. Finally, task-decoupled heads yield predictions, optimized via a joint loss comprising task-specific objectives and a cross-task consistency regularization (Uncertainty-Weighted Multi-Task Learning, Section III-C and III-D).

prompt's reasoning query (e.g., facial micro-expressions) without altering the pre-trained visual backbone. For the text-only DialFC task, E_V is omitted.

B. Prompt-Modulated Shared Reasoning

The core of MDFDU is the SRE, a Transformer-based module with parameters fully shared across tasks. To prevent task interference while maintaining shared parameters, we introduce a *prompt-modulation mechanism*. Rather than treating the task prompt as an input token, we use the latent prompt embedding p_t to modulate the attention maps of the SRE.

Let $E_M = [E_T; E_V]$ denote the multimodal input sequence (or $E_M = E_T$ for the textual-only case). We inject task-specific inductive biases directly into the attention mechanism via a prompt-dependent bias term $b(p_t)$. The attention computation in the SRE is defined as:

$$\text{Attn}(Q, K, V, p_t) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}} + b(p_t)\right)V \quad (4)$$

$$b(p_t) = W_b p_t \quad (5)$$

where W_b is a learnable projection matrix. The output is:

$$H = \text{Transformer}(E_M; \text{bias} = b(p_t)) \quad (6)$$

By adding the bias term $b(p_t)$ to the attention logits, we shift the attention distribution based on the task context. This allows a single set of Transformer weights to execute distinct reasoning modes—switching focus between behavioral inconsistency (DDD) and logical contradiction (DialFC)—while sharing the underlying feature extraction capabilities.

C. Task-Decoupled Inference

The shared representation H is passed to task-specific heads to map the features to the label spaces. This decouples shared incongruence extraction from task-specific label management:

$$\hat{y}_D = \sigma(W_D H + \beta_D), \quad \hat{y}_F = \sigma(W_F H + \beta_F) \quad (7)$$

where WD, F and $\beta D, F$ are the weights and biases for the deception and fact-checking heads, respectively.

D. Uncertainty-Weighted Optimization

Training involves optimizing two classification objectives alongside a regularization term.

a) Consistency Regularization: We hypothesize that semantically corresponding labels, i.e. mapping "Lie" to "Refutes" and "Truth" to "Supported", share a latent space defined by *incongruence* versus *coherence*. To enforce this alignment, we apply a consistency loss between the shared representations of paired samples:

$$\mathcal{L}_{cons} = ||f_{align}(H^{(D)}) - f_{align}(H^{(F)})||_2^2 \quad (8)$$

where $H^{(D)}$ and $H^{(F)}$ denote the shared reasoning representations corresponding to semantically related samples from the deception and fact verification tasks, respectively, and f_{align} is a lightweight projection network that maps representations into a common alignment space.

b) Joint Loss: A rigid summation of losses is suboptimal due to the differing convergence rates of the two tasks. We adopt homoscedastic uncertainty weighting [12], introducing learnable scalar noise parameters σ_D and σ_F to dynamically balance the task gradients:

$$\mathcal{L} = \sum_{t \in D,F} \left(\frac{1}{2\sigma_t^2}\mathcal{L}_t + \log \sigma_t\right) + \frac{\lambda}{\sigma_D \sigma_F}\mathcal{L}_{\text{cons}} \quad (9)$$

$\mathcal{L}_t$ represents the cross-entropy loss for task t. The term $\log \sigma_t$ serves as a regularizer to prevent the model from ignoring a difficult task by driving its σ_t to infinity. This dynamic weighting scheme serves a dual purpose: it automatically down-weights the supervision from the task with higher variance to stabilize convergence, while simultaneously reducing the consistency penalty when uncertainty is high. Consequently, cross-task alignment is enforced progressively, contingent on the reliability of the constituent task representations.

IV. Experiment

A. Dataset

We evaluate our framework on two benchmarks (summarized in Table I), which notably represent the only publicly

TABLE I
LABEL DISTRIBUTION FOR DIALFACT AND BOL

Dataset	S / Truth	R / Lie	NEI
DialFact	7,281	7,298	7,666
BoL	187	862	–

available datasets for the DDD and DialFC tasks, respectively. For the DDD task, we use Box of Lies (BoL) [6], a multimodal dataset of televised game show interactions annotated for veracity. As shown in the table, BoL exhibits severe class imbalance, comprising 862 *Lie* and 187 *Truth* video segments.

For the DialFC task, we employ DialFact [2], a large-scale dataset containing approximately 22,000 conversational claims that require evidence-based verification. Each instance is classified into one of three distinct labels: *Supported* (S), *Refuted* (R), or *Not Enough Information* (NEI).

B. Experimental Settings

Backbone Architectures. We encode text using RoBERTa-base [17] and visual inputs (DDD) via a frozen ViT [18]. LLaMA3-8B generates the task-specific reasoning prompts (P_t). We freeze the ViT to prevent overfitting on the data-scarce BoL dataset ($\sim$1k samples), prioritizing training stability over fine-grained temporal resolution.

Training Strategy. To mitigate the scale disparity between DialFact and BoL, we oversample BoL. An epoch traverses DialFact fully while randomly resampling BoL to ensure consistent supervision. Our uncertainty-weighted loss dynamically prevents gradient domination by the larger task.

Evaluation Metrics. Following official protocols [2], [6], we report Accuracy (Acc) and Macro-F1 (F1). We prioritize Macro-F1 to reliably assess minority class discrimination under severe data imbalance. All reported gains over baselines are statistically significant ($p < 0.05$, paired t-test).

C. Baselines

We compare against existing methods following the official protocols for each dataset. All baselines are reproduced under our experimental settings to ensure a fair comparison.

Baselines for the DialFC Task:

DB-Few [15]: Adapts BART0 [19] via instruction tuning on 48 dialogue tasks, augmented with 100 target-domain examples for few-shot transfer. **Aug-Wow** [2]: Optimizes a pre-trained model for DialFact by appending preceding dialogue turns to the claim to incorporate conversational context. **Auto-FC** [8]: Transfers fact-checking architectures pre-trained on FEVER [7] and VitaminC [20] to the dialogue domain using evidence retrieval and query reformulation modules. **ATEB** [21]: Fine-tunes Gemma-2B on a label-enhanced reformulation of MNLI [22] to sharpen discriminative boundaries for verification. **GPT-4-pipeline** [9]: A baseline using GPT-4 with Chain-of-Thought (CoT) prompting to elicit reasoning paths in zero-shot and few-shot settings. We omit fine-tuning due to GPT-4's proprietary nature and prohibitive cost.

Baselines for the DDD Task:

TABLE II
PERFORMANCE COMPARISON ON DIALFC AND DDD TASKS. THE EVALUATION METRICS ARE ACCURACY (ACC) AND MACRO-F1 (F1).

DialFC			DDD		
Model	**Acc**	**F1**	**Model**	**Acc**	**F1**
ATEB [21]	0.36	0.37	RF1 [6]	0.65	0.66
DB-Few [15]	0.43	0.43	RF2 [13]	0.73	0.72
Aug-WoW [2]	0.52	0.51	Atten-Mixer [14]	0.66	0.55
Auto-FC [8]	0.64	0.63	AVDDCL [23]	0.73	0.60
GPT-4-pipeline [9]	0.73	0.73	BiModal CNN [3]	0.80	0.69
MDFDU (Ours)	**0.75**	**0.76**	**MDFDU (Ours)**	**0.83**	0.75

RF1 [6]: A Random Forest classifier trained on manually curated non-verbal features, serving as a benchmark for human-annotated cues. **RF2** [13]: An automated variant of RF1 utilizing algorithmically extracted acoustic, visual, and linguistic features. **BiModal CNN** [3]: A dual-branch convolutional network fusing linguistic and visual data. We adapt the architecture to ingest BoL visual annotations in place of the original physiological inputs. **Atten-Mixer** [14]: Integrates MLP-Mixer blocks with self-attention to capture high-order dependencies both within and across modalities. **AVDDCL** [23]: Extracts audio-visual cognitive load features and applies focal loss to address class imbalance.

D. Results and Analysis

Table II summarizes the performance of MDFDU against competitive baselines on both the DialFC and DDD benchmarks. Our framework achieves state-of-the-art results on both tasks, recording an Accuracy of **0.75** on DialFact and **0.83** on BoL. More importantly, MDFDU demonstrates consistent and statistically significant gains in Macro-F1, validating its robustness against class imbalance.

1) Analysis of DialFC task: On the DialFact benchmark, MDFDU outperforms both specialized small models and large-scale LLM pipelines. Baselines like **ATEB** (0.36 Acc) and **DB-Few** (0.43 Acc) struggle significantly. These models rely heavily on transfer learning from NLI tasks without explicitly modeling the conversational context, leading to poor generalization on dialogue-specific claims. **Auto-FC** (0.64 Acc) improves upon this by incorporating evidence retrieval, yet it treats verification as a standard sequence classification task, failing to explicitly model the *reasoning steps* required to link claims to evidence. MDFDU outperforms the GPT-4-pipeline (0.76 vs. 0.73 F1), demonstrating the efficacy of specialized reasoning over brute-force scale. Despite using a backbone orders of magnitude smaller (RoBERTa-base, $\sim$125M parameters), MDFDU achieves superior results through precise inductive biases. This confirms that for specialized verification tasks, compact, task-aligned models offer a more reliable and computationally efficient alternative to generalist LLMs.

2) Analysis of DDD task: In the multimodal setting, MDFDU substantially outperforms existing methods, particularly in Macro-F1. Random Forest baselines (**RF1, RF2**) yield moderate performance (Acc: 0.65–0.73), indicating that shallow models and handcrafted features struggle to cap-

TABLE III
ABLATION STUDY ON THE IMPACT OF DIFFERENT COMPONENTS IN MDFDU. ACC: ACCURACY, F1: MACRO-F1. W/O DENOTES "WITHOUT".

Variant	DialFC		DDD	
	Acc	F1	Acc	F1
MDFDU (Full Model)	**0.75**	**0.76**	**0.83**	**0.75**
Architecture Variants:				
(1) w/o MTL	0.69	0.68	0.76	0.64
(2) w/o PC	0.72	0.71	0.79	0.69
Optimization Variants:				
(3) w/o CR ($\mathcal{L}_{cons}$)	0.73	0.74	0.80	0.71
(4) w/o UW	0.72	0.70	0.78	0.67

ture complex deceptive nuances. Conversely, deep learning approaches achieve higher Accuracy but suffer significant Acc-F1 discrepancies. **BiModal CNN** reaches 0.80 Acc but only 0.69 F1, indicating overfitting to the dominant "Lie" class. **Atten-Mixer** and **AVDDCL** face similar imbalance issues (F1: 0.55 and 0.60), despite AVDDCL's focal loss. MDFDU overcomes this, achieving an F1 of 0.75 (+6% over the strongest baseline). Attributable to **Prompt-Modulated Reasoning**, our injected inductive bias forces the model to detect incongruence rather than relying on class priors. This enables MDFDU to identify subtle "Truth" signals missed by others, yielding a balanced precision-recall profile. While computationally heavier than lightweight baselines like RF2, MDFDU's multimodal reasoning gains justify the cost.

3) Impact of Joint Learning: A key observation is that MDFDU achieves these results using a **SRE**. The performance gains validate our hypothesis that deception (DDD) and factual refutation (DialFC) share a common cognitive signature: *incongruence*. By training jointly with uncertainty weighting, MDFDU leverages the larger DialFact dataset to stabilize the reasoning representation for the smaller BoL dataset, effectively transferring the ability to detect semantic inconsistency from text-only to multimodal contexts.

E. Ablation Study

To validate the effectiveness of each component in MDFDU, we conduct an ablation study by selectively removing key modules from the full framework. The results are summarized in Table III for clear comparison.

Experimental Setup for Variants: We compare the full MDFDU model against four distinct variants:

1) **w/o MTL:** We train two separate models independently. One is trained exclusively on DialFact (DialFC task) and the other on BoL (DDD task). This variant removes the shared encoder and joint optimization, isolating the benefit of explicit cross-task knowledge transfer.
2) **w/o Prompt Conditioning (PC):** We remove the LLM-generated task prompts (P_t). For textual input, the model encodes raw dialogue without the prefix. For visual input (in DDD), the cross-attention mechanism is replaced by standard concatenation of global visual features with text embeddings. SRE operates without the prompt-dependent bias term $b(p_t)$.
3) **w/o Consistency Regularization (CR, $\mathcal{L}_{cons}$):** We remove the consistency loss term from the objective function ($\lambda = 0$). The model is trained solely on the summation of task-specific classification losses, removing the explicit constraint to align the latent representations of deception and refutation.
4) **w/o Uncertainty Weighting (UW):** We replace the learnable uncertainty parameters σ_t with fixed scalar weights ($\sigma_D = \sigma_F = 1$). This forces a static 1:1 balance between the two task losses throughout training, ignoring the differences in convergence difficulty.

Analysis of Results: As shown in Table III, the full MDFDU framework achieves the highest performance across all metrics. The degradation in F1 scores is particularly revealing of the robustness provided by each component.

- Impact of Multi-Task Learning: The most significant performance drop occurs in Variant 1 (Single Task). While DDD Accuracy drops by 7%, Macro-F1 plummets by 11% (0.75 $\rightarrow$ 0.64). This contrast confirms that the data-scarce DDD task relies on the reasoning patterns learned from DialFC to generalize to minority classes. Without auxiliary supervision from DialFC, the model overfits to the majority class in BoL.
- Efficacy of Prompt Conditioning: Removing prompt modulation (Variant 2) leads to a decline of 6% in DDD F1. Without explicit reasoning priors (e.g., specific instructions to look for facial contradictions), the shared encoder struggles to disentangle task-specific feature patterns, leading to "negative transfer" where visual noise overwhelms the decision process.
- Necessity of Optimization Strategies: **Consistency Regularization:** Removing $\mathcal{L}_{cons}$ (Variant 3) causes a drop in F1 (0.75 $\rightarrow$ 0.71 on DDD). This indicates that explicit alignment is required to map "deception" and "refutation" to a shared latent space, which is crucial for distinguishing hard negatives. **Uncertainty Weighting:** Variant 4 (Fixed Weights) shows a sharp decline in DDD F1 (0.67). Without dynamic weighting, the larger DialFact dataset dominates gradient updates. The model optimizes for the "easier" global average, neglecting the fine-grained boundaries needed for the imbalanced BoL task. Uncertainty weighting prevents this by down-weighting the dominant task when its variance increases.

The ablation results verify that MDFDU's superior performance stems from the combination of its components. While MTL provides the foundation for transfer, Prompt Conditioning and Uncertainty Weighting are essential for translating that potential into consistently and empirically robust performance on imbalanced, hard-to-classify samples, as evidenced by the sustained high F1 scores.

F. Error Analysis

We analyzed 100 misclassified samples to identify the limitations of MDFDU. The errors primarily fall into three categories, reflecting challenges in fine-grained perception, semantic discrimination, and logical reasoning:

- **Subtle Non-verbal Leakage (DDD):** Approximately 30% of DDD errors occur in "poker face" scenarios. While our prompt mechanism attends to visual cues, the frozen ViT backbone lacks sufficient temporal resolution to capture high-frequency micro-expressions, leading to failures when visual contradictions are extremely subtle. **For instance, the model failed to detect a fleeting smirk (lasting $<$200ms) that contradicted a sorrowful narrative, incorrectly classifying the speaker's generally neutral expression as truthful.** This indicates that static frame-level representations are insufficient for modeling rapid affective fluctuations.
- **NEI vs. Refutation Ambiguity (DialFC):** The most pervasive error (45%) is misclassifying *Not Enough Info* as *Refutes*. The model consistently mistakes topical similarity for semantic conflict, blurring the line between missing and negative evidence. **For instance, if a claim cites a movie's *cast* but the evidence outlines the *plot*, the model predicts *Refutes* based on entity overlap, despite the lack of contradictions.** This highlights a critical deficiency in evaluating sufficiency: the model is biased toward detecting conflict in high-overlap scenarios, failing to recognize when context is merely associative rather than dispositive.
- **Complex Multi-hop Reasoning:** Across both tasks, the model proves unreliable when detecting inconsistencies that require symbolic or numerical deduction. **For instance, verifying a claim that "sales doubled" against evidence showing an increase from** 100 **to** 150 **requires precise arithmetic calculation verification (**$150/100 \neq 2$**), which our model often misclassifies based on the general positive trend.** This limitation suggests that standard attention mechanisms, while highly effective for semantic retrieval, cannot fully substitute for the structured logical inference required to distinguish between qualitative similarity and quantitative exactness.

V. Conclusion

We introduced MDFDU, a unified framework bridging DDD and DialFC by targeting their shared mechanism of *incongruence*. Our Prompt-Modulated Shared Reasoning decouples task semantics from core reasoning, enabling a single encoder to process both multimodal cues and textual conflicts. Furthermore, an uncertainty-weighted consistency loss transfers structural knowledge from DialFC to the data-scarce DDD domain. Achieving state-of-the-art results on BoL and DialFact, our findings confirm that objective reasoning supervision enhances subjective intent detection. Future work will explore fine-grained temporal modeling and CoT integration.

Acknowledgment

The authors would like to thank all the anonymous reviewers for their comments on this paper. This research was supported by the National Natural Science Foundation of China (Nos. 62276177, 62006167 and 62376181), the Natural Science Foundation of the Jiangsu Higher Education Institutions of China (Nos. 24KJB520036 and 25KJA521001), Key R & D Plan of Jiangsu Province (BE2021048), and Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions.

References

[1] W. Khan, K. Crockett, J. O'Shea, A. Hussain, and B. M. Khan, "Deception in the eyes of deceiver: A computer vision and machine learning based automated deception detection," *Expert Systems with Applications*, 2021.

[2] P. Gupta, C.-S. Wu, W. Liu, and C. Xiong, "DialFact: A benchmark for fact-checking in dialogue," in *ACL*, 2022.

[3] P. Li, M. Abouelenien, R. Mihalcea, Z. Ding, Q. Yang, and Y. Zhou, "Deception detection from linguistic and physiological data streams using bimodal convolutional neural networks," in *ISPDS*, 2024.

[4] S. Shankar, L. Thompson, and M. Fiterau, "Progressive fusion for multimodal integration," 2022. [Online]. Available: https://arxiv.org/abs/2209.00302

[5] M. Ding, A. Zhao, Z. Lu, T. Xiang, and J.-R. Wen, "Face-focused cross-stream network for deception detection in videos," in *CVPR*, 2019, pp. 7794–7803.

[6] F. Soldner, V. Pérez-Rosas, and R. Mihalcea, "Box of lies: Multimodal deception detection in dialogues," in *NAACL*, 2019.

[7] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, "FEVER: a large-scale dataset for fact extraction and VERification," in *NAACL*, 2018.

[8] E. Chamoun, M. Saeidi, and A. Vlachos, "Automated fact-checking in dialogue: Are specialized models needed?" in *EMNLP*, 2023.

[9] A. Rangapur and A. Rangapur, "The battle of llms: A comparative study in conversational qa tasks," 2024. [Online]. Available: https://arxiv.org/abs/2405.18344

[10] K. Tian, Y. Jiang, Z. Yuan, B. PENG, and L. Wang, "Visual autoregressive modeling: Scalable image generation via next-scale prediction," in *NeurIPS*, 2024. [Online]. Available: https://openreview.net/forum?id=gojL67CfS8

[11] X. L. Li and P. Liang, "Prefix-tuning: Optimizing continuous prompts for generation," *ACL-IJCNLP*, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:230433941

[12] R. Cipolla, Y. Gal, and A. Kendall, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7482–7491.

[13] J. Zhang, S. I. Levitan, and J. Hirschberg, "Multimodal deception detection using automatically extracted acoustic, visual, and lexical features," in *Proc. Interspeech*, 2020.

[14] X. Guo, Z. Yu, N. M. Selvaraj, B. Shen, A. W.-K. Kong, and A. C. Kot, "Benchmarking cross-domain audio-visual deception detection," 2025. [Online]. Available: https://arxiv.org/abs/2405.06995

[15] P. Gupta, C. Jiao, Y.-T. Yeh, S. Mehri, M. Eskenazi, and J. Bigham, "InstructDial: Improving zero and few-shot generalization in dialogue through instruction tuning," in *EMNLP*, 2022.

[16] C. Fifty, E. Amid, Z. Zhao, T. Yu, R. Anil, and C. Finn, "Efficiently identifying task groupings for multi-task learning," in *NeurIPS*, 2021.

[17] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen *et al.*, "Roberta: A robustly optimized bert pretraining approach," 2019. [Online]. Available: https://arxiv.org/abs/1907.11692

[18] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2021.

[19] B. Y. Lin, K. Tan, C. S. Miller, B. Tian, and X. Ren, "Unsupervised cross-task generalization via retrieval augmentation," in *NeurIPS*, 2022.

[20] T. Schuster, A. Fisch, and R. Barzilay, "Get your vitamin C! robust fact verification with contrastive evidence," in *NAACL*, 2021.

[21] S. Han, F. P. Gomez, T. Vu, Z. Li, D. Cer, H. Zeng *et al.*, "Ateb: Evaluating and improving advanced nlp tasks for text embedding models," 2025. [Online]. Available: https://arxiv.org/abs/2502.16766

[22] A. Williams, N. Nangia, and S. Bowman, "A broad-coverage challenge corpus for sentence understanding through inference," in *NAACL*, 2018.

[23] H. Ji, M. Lee, and E. Park, "Enhancing deception detection with cognitive load features: An audio-visual approach," 2025. [Online]. Available: https://openreview.net/forum?id=WNZNsyzcaB