DPMMA: Dual-Prompt MultiModal Alignment for Deception Detection in Dialogues

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*—Deception Detection in Dialogues (DDD) aims to recognize deceptive communication by analyzing multimodal behaviors. However, existing methods often treat modalities independently before shallow fusion, failing to bridge the inherent semantic gap between high-level linguistic concepts and low-level visual signals. To address this, we propose Dual-Prompt for Multimodal Alignment (DPMMA). Unlike conventional approaches, DPMMA introduces a Dual-Phase Alignment strategy: it integrates LLaMA3-generated "hard prompts" for explicit semantic translation with learnable "soft tokens" for latent geometric adaptation. Furthermore, a Semantic Guidance Module explicitly encodes psychological priors (e.g., cognitive load indicators) to highlight subtle deceptive micro-cues. Finally, a Progressive Fusion Transformer equipped with Gated Residual Injection ensures deep integration while preserving modality-specific details. Experiments on the Box of Lies dataset demonstrate that DPMMA not only outperforms strong baselines but also significantly mitigates class bias, achieving robust detection for both deceptive and truthful statements.**

***Index Terms*—deception detection, multimodal alignment, dual-prompts, semantic**

I. Introduction

Deception, defined as the act of deliberately misleading others, is a pervasive phenomenon in various forms of interpersonal communication, particularly in high-stakes settings, such as courtrooms [1] and competitive games [2], where dishonesty may be incentivized. Contrary to popular belief, studies indicate that people generally overestimate their ability to detect deception, with actual accuracy barely exceeding chance levels [3]. Consequently, Deception Detection in Dialogues (DDD) has emerged as a multidisciplinary research field aiming to automate veracity assessment.

As shown in Fig. 1, DDD aims to predict the truthfulness of statements by synergizing verbal and non-verbal clues. In this study, we specifically target the alignment between textual semantics and visual behaviors.

Despite the significance of the task, current approaches face substantial limitations. While early methods relied on manual feature engineering or single-modal data [4]–[8], the advent of deep learning has introduced more powerful architectures ranging from LSTMs [9] and CNNs [10], [11] to Transformers [12]. However, even recent multimodal frameworks [13]–[16] exhibit two primary deficiencies: (i) Inadequate Fusion and Semantic Gap: Most strategies process modalities in isolation before applying shallow fusion (e.g., concatenation) [17]. This loose coupling fails to align the disparate geometric spaces of high-level linguistic concepts and low-level visual signals, resulting in the dilution of critical deceptive cues. (ii) Neglect of Priors and Class Imbalance: Existing general-purpose models typically ignore the severe class imbalance inherent in DDD datasets [18] and fail to incorporate domain-specific psychological priors, limiting their ability to distinguish genuine emotional leakage from background noise.

Role	Content	System
person A	This is not the first time I have said that.	lie
person B	You said that.	truth
person A	Ok, I am just kidding ok anyway	lie
person B	It looks like you're looking, like we're looking in the mirror.	truth
person A	I know. God. When did I get so ugly?	truth
person B	My face! My face! My beautiful face!	truth
person A	Ok, it is, um, a Rubik's cube inside jello.	lie

Fig. 1. Example of multi-modal deception detection in dialog.

To address these challenges, we propose dual-Prompt for Multimodal Alignment (DPMMA), a novel framework designed to bridge the semantic gap and integrate psychological insights. Unlike previous approaches that rely on implicit alignment, DPMMA introduces a Dual-Phase Alignment strategy: it utilizes LLaMA3-generated "hard prompts" for explicit semantic translation and learnable "soft prompts" for latent geometric adaptation. Moreover, we propose a progressive fusion network that mitigates feature loss and promotes robust integration. Our code will be available at https://anonymous.4open.science/r/DDD-052F upon paper acceptance. Our contributions are summarized as follows:

- We propose a Dual-Phase Alignment framework that integrates LLaMA3-generated hard prompts with learnable soft anchors. This strategy effectively projects heteroge-

neous features into a shared semantic space. Additionally, a Semantic Guidance Module explicitly encodes psychological priors, highlighting deceptive micro-cues while suppressing background noise.

- We introduce a Progressive Fusion Transformer (PFT) equipped with a Gated Residual Injection mechanism. Unlike traditional early or late fusion, this architecture iteratively refines multimodal representations by adaptively re-injecting original semantic contexts, ensuring deep integration without diluting modality-specific details.
- We achieve state-of-the-art performance on the DDD dataset. Specifically, our proposed data augmentation and semantic guidance strategies effectively resolve the long-standing bias issue in the dataset, demonstrating robust performance on both deceptive and truthful classes.

II. Related Work

A. Deception Detection in Dialogues

Early approaches to DDD relied predominantly on hand-crafted features or single-modal analysis. Researchers utilized acoustic prosody, linguistic markers, and micro-expressions to identify falsehoods [4]–[6], [8]. For instance, [7] established benchmarks for the Box of Lies (BoL) dataset by combining manually annotated behavioral cues with linguistic features.

Deep learning architectures, including LSTMs [9], [13], CNNs [10], [11], and Transformers [12], have automated feature extraction to better capture temporal and semantic dependencies. However, these methods typically treat deception detection as a standard classification task, neglecting severe class imbalance [18] and consequently producing models biased toward the majority class and insensitive to the subtle psychological inconsistencies of high-stakes lying.

B. Multimodal Fusion

To leverage complementary information, recent studies have shifted towards multimodal architectures [14], [15]. Despite these advancements, the effective integration of heterogeneous modalities remains an open problem. Standard approaches often employ Late Fusion or Simple Concatenation, where unimodal features are encoded independently and merged only at the classification stage [16], [17]. While computationally efficient, this strategy assumes that features from different modalities (e.g., text and video) naturally reside in a compatible geometric space. In practice, this independence leads to a significant **semantic gap**, where the nuanced relationship between a spoken lie and a subtle visual tic is lost due to a lack of explicit alignment mechanisms [19].

Recent advancements in Prompt Learning [20] suggest that Large Language Models (LLMs) can serve as effective bridges between modalities. By translating visual signals into linguistic descriptions, models leverage the reasoning capabilities of LLMs to align disparate features. Unlike previous methods, our work targets semantic misalignment in DDD by introducing prompt-based semantic projection and progressive fusion.

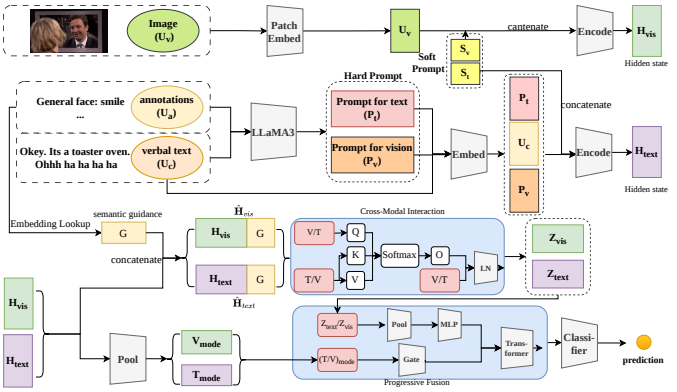

Fig. 2. Overview of the proposed DPMMA model.The model consists of three key stages: (1) Dual-Phase Alignment (Section III-A), which bridges the semantic gap using LLaMA3-generated "hard prompts" and learnable "soft tokens"; (2) Semantic Guidance Injection (Section III-B, which encodes psychological priors via a learnable lookup table; and (3) Progressive Fusion (Section III-C), which integrates cross-modal interactions through a Transformer equipped with Gated Residual Injection.

TABLE I
TEMPLATE FOR LLAMA3 TO GENERATE PROMPT

Role	Content
system	You're good at characterizing people by their features. You're an expert at facial expression analysis. Integrate the given independent features and comments into complete sentences with words begin with 'In the photo'. Ensure correct subject-verb agreement regarding the number of people.
user	Guest verbal: okey. It's a toaster oven. Ohh yaa hahaha; Guest Head: Move Backward; Guest Gaze: Towards object; Guest Mouth Openness: Open mouth; Guest Mouth Lips: Retracted; image name: Box of Lies with Jennifer Lawrence. speaker role: Guest
assistant	In the photo Box of Lies with Jennifer Lawrence, the guest says 'okey. It's a toaster oven. Ohh yaa hahaha' with head move backward. The guest's gaze is towards object, the mouth is open, the mouth lips is retracted.
user	Input content
assistant	Generated content

III. Methodology

Given utterances $Q = \{u_1, ..., u_M\}$ from speakers $S = \{s_1, ..., s_N\}$, the DDD task aims to determine their veracity. As illustrated in Fig. 2, DPMMA operates through a structured pipeline. We clarify its key components: (i) **Hard Prompts**: LLaMA3-generated texts translating behavioral annotations into semantic context. (ii) **Soft Prompts**: Learnable continuous anchors for global cross-modal alignment. (iii) **Semantic Guidance**: Learnable embeddings encoding psychological priors to highlight deceptive cues. (iv) **Cross-Attention Interaction**: Dual-stream mechanism capturing fine-grained dependencies. (v) **Gated Residual Injection**: Adaptive re-injection of aligned features to prevent deep-layer dilution.

A. Data Processing and Prompt Generation

For each utterance u, we extract the verbal text U_c, the corresponding visual frames U_v, and the behavioral annotations

U_a provided by the BoL dataset.

Unlike controlled lab datasets, BoL features dynamic environmental noise (e.g., camera cuts) that compromises automatic feature extraction. Following standard protocols [7], we employ high-fidelity manual annotations (U_a) to isolate multimodal reasoning from low-level extraction errors. These annotations serve as clean "semantic anchors," enabling LLaMA3 to interpret behavioral cues and bridge the gap between psychological concepts and visual signals.

We treat behavioral annotations (U_a) as a structured descriptive text stream rather than auxiliary labels. Accordingly, the task is formulated as aligning verbal (U_c), visual (U_v), and behavioral (U_a) signals. This setup mirrors real-world human-in-the-loop scenarios (e.g., judicial analysis), where expert annotation often precedes automated reasoning.

Prompt Generation: To operationalize this, we utilize LLaMA3-8B to generate descriptive "hard prompts". We construct a composite input $[U_c, U_a]$ containing the verbal claim and behavioral descriptions. Employing the instruction template in Table I, LLaMA3 synthesizes this information into coherent natural language contexts:

$$[P_t, P_v] = \mathrm{LLaMA3}(U_c, U_a) \tag{1}$$

The resulting prompts (P_t, P_v) translate the abstract behavioral tags into fluent descriptive text, providing a unified semantic context that aligns the subsequent textual and visual encoders.

B. Feature Extraction and Alignment

Since deception often manifests as semantic incongruence between verbal claims and non-verbal behaviors (e.g., a sad statement accompanied by a smile), we employ a dual-stream cross-attention mechanism to capture these fine-grained intermodal conflicts. This process involves constructing prompt-augmented inputs, encoding them via Transformer backbones, and injecting psychological guidance.

Encoder Inputs with Dual Prompts. While LLaMA3 provides semantic context, discrete text prompts alone may not optimally align the continuous feature spaces. Drawing inspiration from recent advancements in multimodal prompt tuning [20]–[22], which demonstrate that optimizing continuous vectors can significantly enhance vision-language integration, we introduce learnable "soft prompt" tokens.

We denote these tokens as $\mathbf{S}_t \in \mathbb{R}^{L_s \times D}$ and $\mathbf{S}_v \in \mathbb{R}^{L_s \times D}$ for the textual and visual branches respectively, where L_s is the prompt length and D is the embedding dimension. These continuous vectors serve as adaptive anchors and are prepended to the input sequences.

For the textual branch, we utilize the pre-trained word embedding layer of RoBERTa [23], denoted as $\mathrm{Emb}(\cdot)$. The input sequence $\mathbf{X}_{text}$ is constructed by concatenating the soft tokens with the embeddings of the LLaMA3-generated hard prompts (P_t, P_v) and the verbal utterance (U_c):

$$\mathbf{X}_{text} = [\mathbf{S}_t; \mathrm{Emb}(P_t); \mathrm{Emb}(U_c); \mathrm{Emb}(P_v)] \tag{2}$$

For the visual branch, let U_v denote the sequence of raw visual frames aligned with the utterance. Following the standard Vision Transformer (ViT [24]) protocol, we apply a patch embedding layer, $\mathrm{PatchEmb}(\cdot)$, which flattens the images into 2D patches and projects them linearly to dimension D. The visual input $\mathbf{X}_{vis}$ is formed by concatenating the soft tokens with these patch embeddings:

$$\mathbf{X}_{vis} = [\mathbf{S}_v; \mathrm{PatchEmb}(U_v)] \tag{3}$$

Encoding and Global Alignment. The constructed sequences are processed by their respective encoders to capture contextual dependencies. This yields the hidden state sequences $\mathbf{H}_{text}$ and $\mathbf{H}_{vis}$:

$$\mathbf{H}_{text} = \mathrm{RoBERTa}(\mathbf{X}_{text}), \quad \mathbf{H}_{vis} = \mathrm{ViT}(\mathbf{X}_{vis}) \tag{4}$$

To extract global aligned representations representing the unified semantic geometry, we pool the output states corresponding to the positions of the soft prompt tokens. Let $\mathbf{H}[\mathbf{S}]$ denote the slicing operation that selects the hidden states at indices corresponding to $\mathbf{S}$. The aligned vectors $\mathbf{T}_{mode}$ and $\mathbf{V}_{mode}$ are obtained as:

$$\mathbf{T}_{mode} = \mathrm{Pool}(\mathbf{H}_{text}[\mathbf{S}_t]), \quad \mathbf{V}_{mode} = \mathrm{Pool}(\mathbf{H}_{vis}[\mathbf{S}_v]) \tag{5}$$

where $\mathrm{Pool}(\cdot)$ denotes the mean-pooling operation. These vectors encapsulate the aligned global context essential for the subsequent fusion gating.

Semantic Guidance Injection. Psychological studies suggest that deceivers often exhibit specific cognitive and emotional cues, such as increased pauses, gaze aversion, or rigid facial expressions, closely associated with cognitive load [25]. To minimize potential loss of these subtle clues during feature encoding, we design the Semantic Guidance Module to explicitly encode these psychological priors.

We first extract discrete behavioral attributes (e.g., "Gaze: Averted") from the manual annotations U_a. Since these attributes are categorical, we map them to continuous representations using a **learnable embedding lookup operation**. Specifically, we maintain a matrix $\mathbf{E}_{guide}$, where each row corresponds to a behavioral category:

$$\mathbf{G} = \mathrm{Lookup}(U_a; \mathbf{E}_{guide}) \tag{6}$$

where $\mathbf{G}$ represents the sequence of semantic guidance tokens. To force the encoders to attend to these psychological priors, we append $\mathbf{G}$ to the encoded feature sequences along the sequence dimension. The final logically augmented representations $\hat{\mathbf{H}}_{text}$ and $\hat{\mathbf{H}}_{vis}$ are defined as:

$$\hat{\mathbf{H}}_{text} = [\mathbf{H}_{text}; \mathbf{G}], \quad \hat{\mathbf{H}}_{vis} = [\mathbf{H}_{vis}; \mathbf{G}] \tag{7}$$

These augmented sequences serve as the input for the cross-modal interaction module.

C. Multimodal Integration and Prediction

Cross-Modal Interaction: We employ a dual-stream Cross-Attention mechanism to capture fine-grained dependencies. The augmented sequences $\hat{\mathbf{H}}_{text}$ and $\hat{\mathbf{H}}_{vis}$ serve as inputs. In the Text-Guided Vision branch (T $\rightarrow$ V), the textual features act as the Query (Q), while visual features serve as Key (K) and Value (V):

$$Q_t = \hat{\mathbf{H}}_{text}\mathbf{W}_Q, \quad K_v = \hat{\mathbf{H}}_{vis}\mathbf{W}_K, \quad V_v = \hat{\mathbf{H}}_{vis}\mathbf{W}_V \tag{8}$$

The interactive output $\mathbf{O}_{t\leftarrow v}$ is computed via scaled dot-product attention:

$$\mathbf{O}_{t\leftarrow v} = \text{Softmax}\left(\frac{Q_t K_v^\top}{\sqrt{d_k}}\right) V_v \tag{9}$$

We apply residual connections and Layer Normalization (LN) to obtain the final interactive representations $\mathbf{Z}_{text}$ and $\mathbf{Z}_{vis}$:

$$\mathbf{Z}_{text} = \text{LN}(\hat{\mathbf{H}}_{text} + \mathbf{O}_{t\leftarrow v}), \quad \mathbf{Z}_{vis} = \text{LN}(\hat{\mathbf{H}}_{vis} + \mathbf{O}_{v\leftarrow t}) \tag{10}$$

Progressive Fusion and Prediction: To prevent the dilution of modality-specific signals in deep networks, we propose the PFT. Formally, we first obtain global interactive representations by performing mean-pooling on the cross-attention outputs. These are projected into a shared latent space via an MLP to initialize the fusion stream:

$$Y_0 = \text{MLP}([\text{Pool}(\mathbf{Z}_{text}); \text{Pool}(\mathbf{Z}_{vis})]) \tag{11}$$

The integrated features Y_0 serve as the input to the PFT. To strictly enforce the semantic geometry anchored by soft prompts throughout the fusion depth, we employ a Gated Residual Injection mechanism. Instead of injecting raw embeddings which lack contextual semantics, we utilize the global aligned representations $(\mathbf{T}_{mode}, \mathbf{V}_{mode})$ derived in Eq. 5, as they encapsulate the aggregated information of prompts and utterances. The update at layer k is expressed as:

$$Y_{k+1} = \text{Transformer}(Y_k + \lambda \cdot \mathcal{G}(\mathbf{T}_{mode}, \mathbf{V}_{mode})) \tag{12}$$

where λ is a learnable scaling factor. The function $\mathcal{G}(\cdot)$ acts as an adaptive adapter. Since the dimension of the aligned encoders $(\mathbf{T}_{mode}, \mathbf{V}_{mode})$ differs from the fusion hidden state Y_k, $\mathcal{G}$ performs a linear projection on the concatenated inputs to align dimensions and fuse the heterogeneous signals into a unified injection vector. This mechanism ensures that the original semantic context acts as a consistent guide without disrupting the feature space of the deep fusion layers.

IV. Experiments

A. Dataset

We conduct experiments on BoL [7], the only publicly available multimodal resource for DDD. Derived from a televised game, BoL serves as a valuable proxy for high-stakes deception, capturing genuine psychological stress and improvisation comparable to real-world scenarios. Unlike trivial falsehoods, these interactions involve cognitive complexity, challenging models to detect subtle incongruences between verbal claims and non-verbal behaviors.

The dataset comprises 25 videos yielding 862 deceptive and 187 truthful utterances, annotated with transcripts and behavioral labels. To address class imbalance, we augment the truthful class by pairing original visual features with LLaMA3-generated semantic paraphrases, thereby expanding the total dataset to approximately 2,000 utterances. This strategy relies on the premise that non-verbal indicators of truthfulness (e.g., direct gaze) persist across linguistic variations of the same claim.

TABLE II
Performance of DDD models on BoL. Acc= Accuracy, P=Precision, R=Recall, F=F1-score.

Models	Acc	Lie			Truth		
		P	R	F	P	R	F
RF1	0.65	0.64	0.67	0.66	0.66	0.63	0.65
RF2	0.73	0.75	0.69	0.71	0.71	0.77	0.74
Atten-Mixer	0.66	0.82	0.73	0.78	0.55	0.58	0.56
AVDDCL	0.73	0.82	0.83	0.83	0.65	0.62	0.63
BiModel CNN	0.80	0.83	0.86	0.84	0.76	0.73	0.74
DPMMA	**0.84**	**0.84**	**0.89**	**0.86**	**0.84**	**0.79**	**0.81**

B. Experimental Settings

Following standard protocols [7], we utilize a 5-fold cross-validation scheme. Regarding input modalities, we adopt the Oracle Feature Setting by incorporating verbal transcripts, visual frames, and manual behavioral annotations. This strategy isolates the model's multimodal reasoning capability from upstream visual recognition errors, allowing for a rigorous evaluation of our Dual-Phase Alignment. During prompt generation and data augmentation, LLaMA3 is configured with a temperature of 0.7 and a top-p of 0.9 to ensure diverse yet semantic-preserving outputs. In contrast, we explicitly exclude audio modalities due to excessive background noise (e.g., audience laughter) in the BoL dataset. Consequently, this study focuses exclusively on robust Visual-Textual analysis.

Given the limited prior work on BoL, we adopt the metrics proposed by its creators [7] and compare DPMMA against the following reproduced baselines. All reproduced models are implemented strictly following the hyperparameter configurations detailed in their respective original papers:

- **RF1** [7]: A Random Forest (RF) model trained on manually annotated nonverbal features.
- **RF2** [16]: A RF variant trained on automatically extracted acoustic, visual, and linguistic features.
- **BiModal CNN** [15]: A fusion network originally designed for linguistic and physiological modalities. We adapt this method to BoL by replacing physiological signals with the dataset's behavioral annotations while preserving the original linguistic input.
- **Atten-Mixer** [26]: A fusion method combining MLP-Mixer and self-attention layers. We reproduce this architecture to capture intra- and inter-modal interactions.
- **AVDDCL** [27]: A framework that extracts cognitive load features from audio-visual data and utilizes focal loss. We reproduce their method for comparison.

C. Overall Results and Analysis

Table II presents comparative performance of DPMMA and baseline models on the BoL dataset. The RF baselines (RF1, RF2) achieve accuracies of 0.65 and 0.73, respectively. While RF2 demonstrates the value of incorporating multimodal signals over manual features, its shallow architecture fails to capture deep semantic correlations across modalities.

Among deep learning approaches, Atten-Mixer [26] and AVDDCL [27] yield suboptimal performance (Acc=0.66 and 0.73). We attribute this to two limitations: (1) Unbridged

TABLE III
PERFORMANCE OF ABLATION STUDY. ACC= ACCURACY, P=PRECISION, R=RECALL, F=F1-SCORE

Models	Acc	Lie			Truth		
		P	R	F	P	R	F
DPMMA	**0.84**	**0.84**	**0.89**	**0.86**	**0.84**	**0.79**	**0.81**
w/o integration	0.77	0.78	0.85	0.81	0.76	0.69	0.72
w/o prompt	0.78	0.80	0.85	0.82	0.77	0.71	0.74
w/o V-prompt	0.79	0.79	0.87	0.83	0.79	0.71	0.75
w/o T-prompt	0.80	0.81	0.87	0.84	0.79	0.73	0.76

Semantic Gap: These models attempt to fuse raw audio-visual signals directly with text without explicit alignment. Lacking a mechanism like our Dual-Phase Alignment, the high-level semantic abstractions of deception fail to align with low-level sensory features, leading to noisy cross-modal interactions. (2) Noise Sensitivity: Their reliance on acoustic features (which we exclude) introduces interference from the noisy TV show environment. The BiModal CNN [15] achieves a competitive accuracy of 0.80. However, it relies on a static fusion strategy. As deception often involves subtle and transient cues, its rigid architecture likely struggles to adaptively weigh conflicting signals across modalities.

In contrast, DPMMA achieves state-of-the-art performance with 0.84 accuracy and 0.86 F1-score for deceptive utterances. Crucially, it maintains balanced performance with a Truth F1-score of 0.81, significantly outperforming baselines that collapse on the minority class. This robustness validates our core methodological contributions:

- Addressing the Semantic Gap: By using LLaMA3-generated "hard prompts" and learnable "soft anchors", DPMMA projects heterogeneous features into a unified semantic geometry before fusion, ensuring that visual cues are semantically intelligible to the text encoder.
- Preserving Psychological Priors: The high recall on Truth samples suggests that our Semantic Guidance Module successfully encoded behavioral priors (via the lookup table), allowing the model to identify truthful signals (e.g., direct gaze) even within an imbalanced dataset.
- Deep Integration without Dilution: The superior accuracy indicates that the PFT effectively synthesized these clues. Unlike BiModal CNN, our Gated Residual Injection mechanism ensures that original modality-specific details are not washed out in deep layers, preserving critical micro-expressions needed for the final verdict.

D. Ablation Study

To verify the contribution of each component in DPMMA, we conducted an ablation study as detailed in Table III.

Impact of Progressive Fusion (w/o integration): Substituting PFT with simple concatenation sharply degrades performance (Accuracy $0.84 \rightarrow 0.77$; Truth F1 $0.81 \rightarrow 0.72$), validating the Gated Residual Injection mechanism. Concatenation causes "feature dilution", where subtle clues are overwhelmed by dominant signals. PFT mitigates this by iteratively re-injecting aligned features ($\mathbf{T}_{mode}, \mathbf{V}_{mode}$), preserving direct access to the original semantic geometry.

(a) Overt Leakage (b) Feigned Composure

	Utterance	GT	BiModal	DPMMA
(a)	This is not the first time I have said that.	Lie	Lie (✓)	Lie (✓)
(b)	When the time I go to Coachella, nobody wants...	Lie	**Truth** (✗)	**Lie** (✓)

Fig. 3. Case study comparing DPMMA with the BiModal CNN baseline. GT denotes Ground Truth. (✓) indicates correct prediction, while (✗) indicates failure.

Impact of Dual-Phase Alignment (w/o prompt): Removing the alignment framework (both hard and soft prompts) causes accuracy to fall to 0.78. This confirms that without the Explicit Semantic Translation provided by LLaMA3, the model struggles to bridge the representational gap between visual perception and linguistic semantics. Further analysis reveals that the visual prompt (**w/o V-prompt**) is more critical than the textual prompt. Removing the visual prompt leads to a lower accuracy (0.79) compared to removing the textual prompt (0.80). We attribute this to the inherent nature of the data: visual features reside in a latent space significantly distant from textual semantics. Consequently, the LLaMA3-generated descriptions act as a critical "translator", converting abstract pixels into semantically aligned concepts that the Transformer can process effectively.

Collectively, these results demonstrate that DPMMA's performance stems not from a single module, but from the synergistic effect of aligning features in input space (Prompts) and preserving them in feature space (Progressive Fusion).

E. Case Study

To demonstrate DPMMA's robustness against subtle deceptive strategies, we visualize two distinct scenarios in Fig. 3.

Case (a): Overt Deception (Easy Sample). In the left image, the speaker exhibits classic "leakage cues" commonly cited in psychological literature: she averts her gaze downward and displays a nervous smile while making a false claim. These are high-intensity visual signals. Consequently, both the baseline (BiModal CNN) and our DPMMA correctly capture this obvious incongruence between her facial expression and the context, predicting "Lie".

Case (b): Masked Deception (Hard Sample). The right image presents a more sophisticated deceptive tactic. Here, the speaker adopts a strategy of "feigned composure"—he explicitly closes his eyes and maintains a rigid, serious facial expression to suppress emotional leakage.

- **Baseline Failure:** Misled by surface-level calmness, the BiModal CNN processes visual features independently, interpreting the absence of overt nervousness (e.g., no fidgeting) as truthfulness, yielding a false negative.
- **DPMMA Success:** In contrast, DPMMA correctly predicts "Lie". This success validates the synergy of our Dual-Phase Alignment. First, the LLaMA3-generated hard prompt explicitly translates the visual act of "closing eyes" into a semantic description, bridging the gap between pixels and concepts. Second, the Semantic Guidance Module identifies this behavior as a psychological indicator of high cognitive load (used to fabricate a story) rather than sincerity. By forcing the model to attend to these subtle "masked" cues via guidance tokens, DPMMA penetrates the speaker's disguise and identifies the underlying deception.

F. Limitations

Despite its efficacy, DPMMA presents three primary limitations. First, it operates in an "oracle" setting relying on manual behavioral annotations. While this isolates our fusion mechanism from upstream visual errors, end-to-end performance with automated feature extractors remains unverified. Second, evaluation is restricted to the relatively small BoL dataset, limiting verifiable generalizability to broader high-stakes contexts. Finally, acoustic modalities were explicitly excluded due to the severe background noise inherent in the televised data.

V. Conclusion

This paper introduces DPMMA, a framework integrating Dual-Phase Alignment and Progressive Fusion to bridge the semantic gap in DDD. Unlike conventional methods, DPMMA utilizes LLaMA3-generated hard prompts and learnable soft tokens to align heterogeneous visual signals within a unified semantic geometry. Additionally, a Semantic Guidance Module is used to encode psychological priors to identify subtle "leakage cues" in high-stakes, noisy-intensive environments. Empirical results on the BoL dataset demonstrate state-of-the-art accuracy, effectively mitigating class imbalance. Future work will focus on integrating robust automated feature extractors for end-to-end deployment, exploring audio denoising techniques to incorporate acoustic modalities, and extending the framework to broader high-stakes scenarios (e.g., judicial trials).

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] M. Jaiswal, S. Tabibu, and R. Bajpai, "The truth and nothing but the truth: Multimodal analysis for deception detection," in *ICDMW*, 2016.
[2] K. E. Sip, M. Lynge, M. Wallentin, W. B. McGregor, C. D. Frith, and A. Roepstorff, "The production and detection of deception in an interactive game," *Neuropsychologia*, 2010.
[3] C. F. Bond Jr and B. M. DePaulo, "Accuracy of deception judgments," *Personality and social psychology Review*, 2006.
[4] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock, "Finding deceptive opinion spam by any stretch of the imagination," in *ACL*, Jun. 2011.
[5] V. Pérez-Rosas and R. Mihalcea, "Experiments in open domain deception detection," in *EMNLP*, 2015.
[6] B. de Ruiter and G. Kachergis, "The mafiascum dataset: A large text corpus for deception detection," 2019. [Online]. Available: https://arxiv.org/abs/1811.07851
[7] F. Soldner, V. Pérez-Rosas, and R. Mihalcea, "Box of lies: Multimodal deception detection in dialogues," in *NAACL*, 2019.
[8] H. Bingol and B. Alatas, "Chaos enhanced intelligent optimization-based novel deception detection system," *Chaos, Solitons & Fractals*, 2023.
[9] D. Peskov, B. Cheng, A. Elgohary, J. Barrow, C. Danescu-Niculescu-Mizil, and J. Boyd-Graber, "It takes two to lie: One to lie, and one to listen," in *ACL*, 2020.
[10] M. K. Camara, A. Postal, T. H. Maul, and G. H. Paetzold, "Can lies be faked? comparing low-stakes and high-stakes deception video datasets from a machine learning perspective," *Expert Systems with Applications*, 2024.
[11] M. Karnati, A. Seal, A. Yazidi, and O. Krejcar, "Lienet: A deep convolution neural network framework for detecting deception," *IEEE Transactions on Cognitive and Developmental Systems*, 2022.
[12] A. Velutharambath and R. Klinger, "UNIDECOR: A unified deception corpus for cross-corpus deception detection," in *WASSA*, 2023.
[13] B. Rodriguez-Meza, R. Vargas-Lopez-Lavalle, and W. Ugarte, "Recurrent neural networks for deception detection in videos," in *ICAT*, 2021.
[14] S. Venkatesh, R. Ramachandra, and P. Bours, "Robust algorithm for multimodal deception detection," in *MIPR*, 2019.
[15] 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.
[16] J. Zhang, S. I. Levitan, and J. Hirschberg, "Multimodal deception detection using automatically extracted acoustic, visual, and lexical features," in *Proc. Interspeech*, 2020.
[17] S. Shankar, L. Thompson, and M. Fiterau, "Progressive fusion for multimodal integration," 2022. [Online]. Available: https://arxiv.org/abs/2209.00302
[18] S. A. Prome, N. A. Ragavan, M. R. Islam, D. Asirvatham, and A. J. Jegathesan, "Deception detection using machine learning (ml) and deep learning (dl) techniques: A systematic review," *Natural Language Processing Journal*, 2024.
[19] C. Cai, S. Liang, X. Liu, K. Zhu, Z. Wen, J. Tao *et al.*, "Mdpe: A multimodal deception dataset with personality and emotional characteristics," 2025. [Online]. Available: https://arxiv.org/abs/2407.12274
[20] M. Jia, L. Tang, B.-C. Chen, C. Cardie, S. Belongie, B. Hariharan *et al.*, "Visual prompt tuning," in *ECCV*, 2022.
[21] M. U. Khattak, H. Rasheed, M. Maaz, S. Khan, and F. S. Khan, "Maple: Multi-modal prompt learning," in *CVPR*, 2023.
[22] Y. Xin, J. Du, Q. Wang, K. Yan, and S. Ding, "Mmap: Multi-modal alignment prompt for cross-domain multi-task learning," *AAAI*, 2024.
[23] 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
[24] 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.
[25] A. Vrij, P. A. Granhag, S. Mann, and S. Leal, "Outsmarting the liars: Toward a cognitive lie detection approach," *Current Directions in Psychological Science*, 2011.
[26] 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
[27] 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