MoMCE: Mixture of Modality and Cue Experts for Multimodal Deception Detection

Dongliang Zhu[1], Ruimin Hu[1]*, Zitong Yu[2,3]*, Xiaobao Guo[4], Mei Wang[1], Shuo Ye[2], Fei Ma[5], and Xiaochun Cao[6]

[1] Wuhan University
[2] Great Bay University
[3] Dongguan Key Laboratory for Intelligence and Information Technology
[4] Nanyang Technological University
[5] Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ)
[6] Sun Yat-sen University
zhudongliang@whu.edu.cn

Abstract. Audio-visual deception detection aims to predict whether a person is lying by integrating visual and acoustic modalities, which has two main challenges: 1) the modality conflict problem and 2) heterogeneous cue representation difficulty. However, existing approaches 1) often overlook the differences across modalities for different individuals; and 2) typically rely on a single encoder to handle diverse and individual-specific cues, which limits models' representation capacity for heterogeneous cues. To address these challenges, we propose **MoMCE**, a novel model with mixture of modality and cue experts for deception detection. It consists of two key components: 1) **Prompt-aware Mixture of Modality Experts**, which employs a learnable prompt routing mechanism to generate adaptive instance-aware modality weight distributions for dynamic modality adjustment. In addition, we propose a consistency-aware expert weighting loss. For samples with high cross-modal consistency, it encourages balanced contributions across modalities. In contrast, for samples with strong conflicts, it reduces the entropy of the modality weight distribution to focus on more reliable modalities. 2) **Prompt-aware Mixture of Cue Experts**, which captures heterogeneous and diverse deceptive cues within each modality. This module introduces multiple experts with distinct semantic biases on top of a shared backbone to model different deceptive patterns. Additionally, we introduce a cue expert diversity loss to balance learning across multiple cue experts, promoting effective representation of diverse deceptive cues. Extensive experiments demonstrate that MoMCE adapts to variations in both cross-modal contributions and cue heterogeneity, achieving substantial improvements in deception detection performance. The code is available at this link.

Keywords: multimodal deception detection · audio-visual

* Corresponding authors.

1 Introduction

Deception is defined as intentional misleading [10], with impacts on public safety [1], judicial fairness [15,37], and economic stability [5,11]. Therefore, improving the accuracy of deception detection is of great importance for safeguarding judicial fairness and maintaining social stability. With the development of deep learning [9, 28, 42], non-intrusive multimodal analysis has gradually become an important direction in deception detection. Such approaches aim to automatically extract and model deception-related features from speech and visual signals, thereby enabling the identification of potential deceptive behaviors [18, 27].

In multimodal deception detection tasks, it is often assumed that different modalities within the same video segment are consistent, meaning that a liar would reveal deceptive cues across all modalities. However, psychological studies [10] indicate that because deceivers attempt to disguise themselves, they often find it difficult to maintain complete consistency across modalities. For example, an individual may have a controlled facial expression, but their voice may reveal nervousness; conversely, their voice may remain steady, but their facial expressions may betray deception. This "modality conflict" phenomenon [10] results in different representations of deception across modalities, with the importance of each modality varying across specific samples. Moreover, even within the same modality, deception cues exhibit significant inter-individual variability [10, 47]. For example, in the visual modality, some individuals display noticeable eye movements when lying, while others primarily reveal information through subtle facial expressions. In the audio modality, some individuals speed up their speech or pause frequently, while others primarily express emotion through changes in voice tone. Such cue heterogeneity means that different individuals may rely on completely different behavioral patterns to convey deception cues within the same modality [31].

Existing methods primarily integrate multimodal information through common modality fusion techniques, such as feature concatenation [26, 27, 35], decision fusion [21,49], and adaptive fusion [11,18,46,51]. These approaches typically apply the same fusion strategy to all samples, ignoring differences in the contribution of each modality across samples. As illustrated in Figure 1, certain deceptive samples may not exhibit detectable cues in all modalities, leading to conflicting predictions. Relying on conventional fusion strategies under such conditions may leads to erroneous final decisions. On the other hand, to address deceptive cue heterogeneity, some studies attempt to introduce more handcrafted features [19, 20, 26, 35] or employ more powerful feature extractors [18, 24]. However, these methods still rely on a single feature space and lack the ability to dynamically adapt to individual differences, thus limiting the model's generalization performance across individuals.

In this paper, we propose a Mixture of Modality and Cue Experts (MoMCE) to jointly address the modality conflict and cue heterogeneity challenges. The model consists of two main components. First, to handle modality conflict, we design a prompt-aware mixture of modality experts. This module combines learnable prompt routing and gating mechanisms, treating each modality as

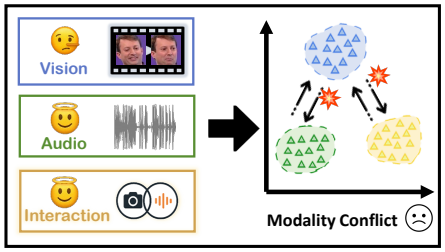

Fig. 1: Example of modality conflict in multimodal deception detection. The visual modality reveals deceptive cues, whereas the audio and interaction modalities remain consistent with genuine behavior. Such inter-modality inconsistency leads to conflicting evidence, directly degrading prediction confidence and accuracy.

an independent "expert." It adaptively assigns modality weights based on input features, thereby enabling dynamic modeling of modality contributions. Moreover, to mitigate performance degradation caused by incorrect routing, we introduce a consistency-aware expert weight loss: when the modalities are consistent (e.g., truthful samples or low-conflict deceptive samples), the weight distribution is constrained to remain balanced; when modality conflict is significant (high-conflict deceptive samples), the model is encouraged to reduce weight distribution entropy, leading to a more concentrated weight distribution and improved decision stability.

Second, to address cue heterogeneity, we propose a prompt-aware mixture of cue experts model. This model introduces multiple sets of prompts with different semantic priors to instantiate several backbone encoders with shared parameters, constructing cue experts at multiple levels. Each expert focuses on capturing different types of deceptive cues, thereby covering the diverse deceptive behavior patterns. In addition, during training, cue-level experts tend to overfit to the most salient cues, leading to insufficient learning of other potential cues [36, 43, 44, 48]. To address this, we design a cue diversity constraint that balances the learning process across cue experts, ensuring that various deception cues are effectively modeled. Our main contributions are as follows:

- To address modality conflicts, we propose a mixture of modality experts that dynamically generates sample-specific cross-modal weight distributions, enabling adaptive fusion of modality-level features. To further mitigate erroneous routing, we introduce a consistency-aware expert weighting loss to constrain the model's routing entropy and regulate the modality weights.
- We propose a mixture of cue experts that leverages prompts with different semantic priors to construct multiple cue-level experts for each modality, enabling the model to capture heterogeneous deception cues. To prevent expert collapse, we further design a cues diversity constraint that encourages experts to learn complementary and discriminative features.

– Extensive experiments on the DOLOS, Bag-of-Lies, and MU3D datasets show that our method outperforms state-of-the-art multi-modal deception detection approaches, demonstrating its effectiveness.

2 Related work

Audio-Visual Deception Detection. Audio-visual joint learning for multimodal deception detection aims to identify deceptive behaviors by simultaneously modeling audio and visual modalities [52]. Previous studies can be broadly categorized into two directions: optimizing modality fusion and enhancing deception cue representation. In the first category, early methods typically integrated handcrafted multimodal features through simple feature concatenation or decision-level fusion [2, 7, 21, 33, 38]. With the rise of deep learning, some approaches began leveraging attention mechanisms to fuse deep representations, including CNN [17, 27], ResNet [11], DNN [39], LSTM [22], GNN [51], ViT [58], and Wav2Vec [41, 58]. More recently, research has shifted towards efficient and adaptive methods. For example, PECL [18] introduces cross-modal adapters to learn latent alignments and relational features across modalities. However, such methods typically apply the same processing strategy to all input samples, ignoring potential modality conflicts. The second line of work [18,19,24,26,29,35], attempts to capture richer deception cues by introducing more hand-crafted features or training more powerful encoders. However, these approaches still rely on a single feature space and lack the ability to dynamically adapt to individual differences. In contrast, our approach performs dynamic modality weighting and employs a mixture of cue experts to model heterogeneous cues, effectively addressing modality conflicts and cue heterogeneity, thereby enabling a more flexible and adaptive multimodal deception detection framework.

Mixture-of-Experts. The core idea of a Mixture-of-Experts (MOE) is to dynamically select the most suitable "experts" based on the input, where each expert specializes in capturing a specific type of feature. Through a collaborative mechanism among experts, the MoE effectively integrates the strengths of each expert while maintaining computational efficiency, thereby improving the model's performance. In multimodal learning, recent studies have applied MoE to achieve several goals, including reducing computational cost [34,40], handling modality missingness [50], modeling modality interactions [3,6], dynamic modality fusion [8], cross-modal continual learning [23], and aligning modality-specific and modality-invariant features [14,16,55]. Inspired by these advances, we decouple inter- and intra-modality processing, represent both modalities and deception cues as experts, and introduce prompt-aware structures to automatically aware expert competence and learn their importance weights. Thereby improving the ability to personalized model individual deceptive behaviors.

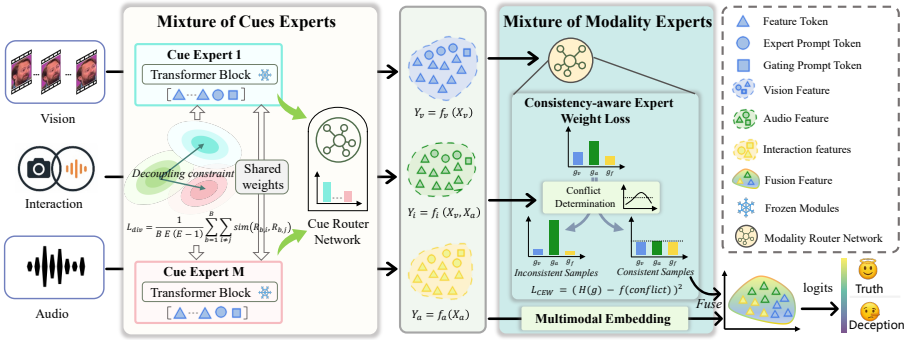

Fig. 2: Framework of MoMCE. The proposed framework consists of a prompt-aware MoCE for modeling diverse deceptive cues within each modality and a MoME for dynamically integrating visual, audio, and interaction representations.

3 Method

3.1 Overall Framework

Following typical multimodal deception detection approaches [18], this paper aims to identify deception in videos using both visual and audio signals. Given a video segment U, the inputs consist of a visual modality (v) and an audio modality (a). The visual and audio streams are encoded into embedding sequences through a two-dimensional convolutional neural network (2D-CNN) and a one-dimensional convolutional neural network (1D-CNN), respectively. These two modalities are then projected into a unified embedding space:

$$X_v \in \mathbb{R}^{L\times D}, \quad X_a \in \mathbb{R}^{L\times D}, \tag{1}$$

where L represents the sequence length and D denotes the embedding dimension. Motivated by [11], we model the interaction features as an independent modality and employ a cross-modal attention mechanism to align and integrate the visual and audio modalities, thereby obtaining the interaction embeddings.

Our overall framework is shown in Figure 2 and consists of two main parts. In the mixture of cue experts, each modality introduces multiple cue-level experts to capture diverse deception cues. These experts are then aggregated into modality-level features through a prompt-aware routing network.

$$Y_v = f_v(X_v), \quad Y_i = f_i(X_v, X_a), \quad Y_a = f_a(X_a), \tag{2}$$

where f_v, f_i, and f_a denote the visual, interaction, and audio mixture of cue experts, respectively. In the mixture of modality experts stage, the visual features Y_v, audio features Y_a, and interaction features Y_i are dynamically weighted by a modality-level prompt routing mechanism to evaluate the relative contribution of each modality. This produces a unified representation, which is then processed by a feed-forward network (FFN) to generate the prediction.

3.2 Prompt-aware Mixture of cue experts

To capture the diverse deceptive cues within each modality, we propose a prompt-aware mixture of cue experts module. This module consists of multiple cue experts and a dynamic prompt routing network. Each cue expert is built based on a pre-trained ViT [12] (for the visual modality) or W2V2 [4] (for the audio modality), and is supplemented with learnable prompts to ensure strong representational power and expert diversity. Formally, for visual or audio modalities, let the token sequence at the n-th Transformer block be

$$X^{(n)} = [x_1^{(n)}, \ldots, x_L^{(n)}]. \tag{3}$$

Inspired by Visual Prompt Learning [25,53,54], we introduce two types of learnable prompt tokens: gating prompts and expert prompts. These prompts are prepended to the input sequence and interact with visual or audio tokens through self-attention. After passing through the encoder layers, the gating prompts aggregate contextual information and are projected to generate routing scores, which are normalized to obtain the expert weight distribution. Based on our dual-layer MoE, the gating prompts are divided into cue-level prompts $p_{\text{cue}}^{(n,m)}$ and modality-level prompts $p_{\text{modal}}^{(n,m)}$, corresponding to cue routing and modality weighting, respectively. The expert prompts $p_{\text{exp}}^{(n,m)}$ interact with input features via the encoder layers, guiding the encoder to bias its representations toward different feature subspaces, thereby enabling experts to capture diverse deceptive cues. Let T_{g} and T_{e} denote the prompt lengths of the gating prompts and expert prompts, respectively. The input for the m-th expert at the n-th layer is:

$$\overline{Z}^{(n,m)} = \left[\, \mathbf{p}_{\text{modal}}^{(n,m)};\, \mathbf{p}_{\text{cue}}^{(n,m)};\, \mathbf{p}_{\text{exp}}^{(n,m)};\, \mathbf{X}^{(n)} \right], \tag{4}$$

where $\mathbf{p}_{\text{modal}}^{(n,m)} = (p_{\text{modal}}^{(t,n,m)})_{t=1}^{T_g}$, $\mathbf{p}_{\text{cue}}^{(n,m)} = (p_{\text{cue}}^{(t,n,m)})_{t=1}^{T_g}$, $\mathbf{p}_{\text{exp}}^{(n,m)} = (p_{\text{exp}}^{(t,n,m)})_{t=1}^{T_e}$, $\mathbf{X}^{(n)} = (x_\ell^{(n)})_{\ell=1}^{L}$. Subsequently, $\overline{Z}^{(n,m)}$ is passed through the frozen, pre-trained Transformer block to perform contextual representation learning

$$Z^{(n,m)} = \text{TransformerBlock}\big(\overline{Z}^{(n,m)}; \theta_{\text{shared}}^{(n)}\big). \tag{5}$$

Prompt-aware Routing Network. We extract $\overline{p}_{\text{cue}}^{(n,m)}$ from $Z^{(n,m)}$, which interacts with expert prompts and the original tokens through a self-attention. As a result, it is able to perceive each expert's response to the current features:

$$\overline{p}_{\text{cue}}^{(n,m)} = Z^{(n,m)}[\text{index of } p_{\text{cue}}^{(n,m)}]. \tag{6}$$

The routing score for each expert is then computed as

$$s^{(n,m)} = {w^n}^\top \overline{p}_{\text{cue}}^{(n,m)}, \qquad \alpha^{(n,m)} = \frac{\exp\big(s^{n,m}\big)}{\sum_{j=1}^{M} \exp\big(s^{n,j}\big)}, \tag{7}$$

where w^n is the learnable weight, $\alpha^{(n,m)}$ is the normalized routing weight obtained. we obtain $\tilde{Z}^{(n,m)}$ by removing the expert and cue gating prompts from

$Z^{(n,m)}$. The final fused feature of the MoCE module is obtained by

$$Y^n = \sum_{m=1}^{M} \alpha^{(n,m)} \cdot \tilde{Z}^{(n,m)}. \tag{8}$$

The visual and audio branches independently perform the above computations, resulting in visual features Y_{v}^n and audio features Y_{a}^n. For the interaction branch, let the m-th interaction expert at the n-th layer, the inputs are

$$\begin{aligned} Z_{\mathrm{a}}^{(n,m)} &= [\, p_{\mathrm{a,g}}^{(n,m)},\, p_{\mathrm{a,e}}^{(n,m)},\, Y_{\mathrm{a}}^{n}\,], \\ Z_{\mathrm{v}}^{(n,m)} &= [\, p_{\mathrm{v,g}}^{(n,m)},\, p_{\mathrm{v,e}}^{(n,m)},\, Y_{\mathrm{v}}^{n}\,], \end{aligned} \tag{9}$$

where $p_{\mathrm{a,g}}^{(n,m)}$ and $p_{\mathrm{a,e}}^{(n,m)}$ denote the learnable audio expert gating prompt and the audio expert prompt, respectively, and $p_{\mathrm{v,g}}^{(n,m)}$ and $p_{\mathrm{v,e}}^{(n,m)}$ are defined analogously. Within each interaction expert, cross-modal attention is applied to capture interactions between the two modalities:

$$\begin{aligned} \hat{Z}_{\mathrm{a}}^{(n,m)} &= \mathrm{MHSA}\big(Z_{\mathrm{a}}^{(n,m)},\, Z_{\mathrm{v}}^{(n,m)},\, Z_{\mathrm{v}}^{(n,m)}\big), \\ \hat{Z}_{\mathrm{v}}^{(n,m)} &= \mathrm{MHSA}\big(Z_{\mathrm{v}}^{(n,m)},\, Z_{\mathrm{a}}^{(n,m)},\, Z_{\mathrm{a}}^{(n,m)}\big), \end{aligned} \tag{10}$$

where MHSA denotes the multi-head self-attention mechanism. Subsequently, we extract the gating prompts and token representations from $\hat{Z}_{\mathrm{a}}^{(n,m)}$ and $\hat{Z}_{\mathrm{v}}^{(n,m)}$:

$$\begin{aligned} \overline{p}_{\mathrm{a,g}}^{(n,m)},\, \overline{Y}_{\mathrm{a}}^{(n,m)} &= \hat{Z}_{\mathrm{a}}^{(n,m)}\big[\text{index of } p_{\mathrm{a,g}}^{(n,m)},\, Y_{\mathrm{a}}^{n}\big], \\ \overline{p}_{\mathrm{v,g}}^{(n,m)},\, \overline{Y}_{\mathrm{v}}^{(n,m)} &= \hat{Z}_{\mathrm{v}}^{(n,m)}\big[\text{index of } p_{\mathrm{v,g}}^{(n,m)},\, Y_{\mathrm{v}}^{n}\big]. \end{aligned} \tag{11}$$

We compute both the interaction features $\overline{Y}_{\mathrm{i}}^{(n,m)}$ and the expert-level gating token $\overline{p}_{\mathrm{i,g}}^{(n,m)}$ as

$$\begin{aligned} \overline{Y}_{\mathrm{i}}^{(n,m)} &= \phi\big(W_f[\overline{Y}_{\mathrm{a}}^{(n,m)},\, \overline{Y}_{\mathrm{v}}^{(n,m)}]\big), \\ \overline{p}_{\mathrm{i,g}}^{(n,m)} &= \phi\big(W_c[\overline{p}_{\mathrm{a,g}}^{(n,m)},\, \overline{p}_{\mathrm{v,g}}^{(n,m)}]\big), \end{aligned} \tag{12}$$

where $[\cdot,\cdot]$ denotes concatenation, W_f and W_c are trainable weights, and $\phi(\cdot)$ is the ReLU activation. Finally, we select the experts and perform weighted fusion according to Eqs. (7)–(8), yielding the final interaction representation:

$$Y_{\mathrm{i}}^{n} = \text{MoE-Fusion}(\overline{Y}_{\mathrm{i}}^{(n,m)}, \overline{p}_{\mathrm{i,g}}^{(n,m)}). \tag{13}$$

Cue Expert Diversity Loss. The *Mixture of Cue Experts* aim to capture heterogeneous deceptive cues. However, during the training of multiple experts, a "dominant cue" phenomenon often arises: several experts tend to focus on the most salient and easy-to-learn deceptive cues, while other potential cues receive insufficient attention. To mitigate this issue, we introduce the *Cue Expert Diversity Loss* to balance the learning among experts. The loss is defined as:

$$\mathcal{L}_{\mathrm{div}} = \frac{1}{BM(M-1)} \sum_{b=1}^{B} \sum_{i \neq j} \cos(\mathbf{R}_{b,i}, \mathbf{R}_{b,j}), \tag{14}$$

where $\mathbf{R} \in \mathbb{R}^{B\times M\times D}$ denotes the expert representations for a batch of B samples. Here, $\cos(\mathbf{R}_{b,i}, \mathbf{R}_{b,j})$ denotes the cosine similarity between the i-th and j-th expert representations of the b-th sample, M denotes the number of experts. This encourages expert to provide complementary features, thereby reducing redundancy and improving the capture of diverse deceptive cues.

3.3 Prompt-aware Mixture of Modality Experts

After MoCE encoding, we obtain features for three modalities. These features contain both modality gating prompts and feature tokens. Similar to the prompt-aware routing network, we use modality gating prompts to generate the expert weight distribution. Specifically, we extract modality gating prompts $\overline{p}^{(m)}_{\text{modal}}$ and feature tokens $\overline{X}_k$:

$$\overline{p}^{(m)}_{\text{modal}}, \overline{X}_k = Y_k[\text{index of } p^{(m)}_{\text{modal}}, X], \quad k \in \{\text{vis}, \text{aud}, \text{inter}\}. \tag{15}$$

Afterwards, we compute modality weights:

$$u_k = \mathbf{w}^{\top}_{\text{modal}} \overline{p}^{(m)}_{\text{modal}}, \qquad \beta_k = \frac{\exp(u_k)}{\sum_{j\in\{\text{vis},\text{aud},\text{inter}\}} \exp(u_j)}, \tag{16}$$

where $\mathbf{w}_{\text{modal}}$ is a learnable weight, u_k is the score for modality k, β_k is the softmax weight for modality k. Finally, we obtain the fused multimodal representation Y_{c} by weighting:

$$Y_{\text{c}} = \sum_{k\in\{\text{vis},\text{aud},\text{inter}\}} \beta_k \overline{X}_k. \tag{17}$$

Based on this fused representation, we apply a feed-forward network (FFN) to obtain the prediction $\hat{y}$, and compute the loss function by comparing it with the ground-truth label y.

Consistency-aware Expert Weight Loss. Although the MoME can assign expert weights according to each expert's capability, due to the highly individualized nature of modality conflicts across different samples, the learned modality weight distribution may still rely on unreliable modalities. We introduce a consistency-aware expert weight loss, which constrains the modality weight distribution based on the conflict level. Specifically, let the conflict scores κ be:

$$\kappa = 1 - \frac{1}{3 * (3-1)} \sum_{k\neq l} \cos(\overline{X}_k, \overline{X}_l), \tag{18}$$

where k, l index the visual, audio, and interaction, and $\cos(\overline{X}_k, \overline{X}_l)$ computes the cosine similarity between the mean representations of modality k and l. Given a threshold c_{th}, which serves as a coarse partition of conflict intensity rather than a dataset-specific tuning parameter, we compare the samples' conflict scores κ with the threshold and divide them into three categories: *genuine samples*, *low-conflict deceptive samples*, and *high-conflict deceptive samples*. For genuine and

low-conflict deceptive samples, we adopt a more uniform weight distribution in the loss to encourage modeling modality diversity, while for high-conflict deceptive samples, we promote a more concentrated weight distribution to emphasize the most reliable modality. We measure the dispersion of the weights using the routing entropy [14], defined as:

$$\mathrm{H}(\beta^{(b)}) = -\sum_{k} \beta_k^{(b)} \log(\beta_k^{(b)} + \epsilon), \tag{19}$$

where ϵ is a small constant for numerical stability. The consistency-aware expert weight loss $\mathcal{L}_{\mathrm{CEW}}$ is defined as:

$$\mathcal{L}_{\mathrm{CEW}} = \frac{1}{B} \sum_{b=1}^{B} \left(\mathrm{H}(\beta^{(b)}) - \mathrm{H}_b^*\right)^2, \tag{20}$$

where H^* denotes the target entropy. For genuine and low-conflict deceptive samples, we set $\mathrm{H}^* = 1.0$ to encourage a relatively balanced modality distribution, while for high-conflict deceptive samples, we set $\mathrm{H}^* = 0.0$ to encourage the model to focus on the most reliable modality. These values represent two extreme routing prototypes rather than strict optimization targets.

3.4 Loss Functions

We combine the above loss functions to obtain the overall training objective:

$$\mathcal{L} = \mathcal{L}_{\mathrm{cls}} + \lambda_{\mathrm{aux}} \mathcal{L}_{\mathrm{CEW}} + \lambda_{\mathrm{div}} \mathcal{L}_{\mathrm{div}}, \tag{21}$$

where $\mathcal{L}_{\mathrm{cls}}$ denotes cross-entropy loss. λ_{aux} and λ_{div} are weights of the auxiliary losses. The overall workflow of MoMCE is detailed in Appendix.1.

4 Experimental Analysis

4.1 Implementation Details

For the visual modality, we use AlphaPose [13] to extract faces from videos. For each face clip, we uniformly sample 64 frames, normalize them, and resize them to 160×160 pixels. The face images are encoded using a 2D-CNN module to produce 256-dimensional features. For the audio data, the raw speech signal is resampled and encoded using a 1D-CNN module, resulting in a 512-dimensional feature vector for each audio sample. The visual and audio tokens are then projected into 768 dimensions through a linear projection layer. We train the model for 100 epochs using the Adam optimizer with a learning rate of 1×10^{-5}. The consistency-aware expert weight loss is inactive for the first 30 epochs to avoid early interference from unstable modality features. Model performance is evaluated in terms of Accuracy (ACC), F1-score (F1), and Area Under the Curve (AUC). All training and inference are conducted on two H100 GPUs. Unless otherwise specified, the number of experts M in MoCE is set to 8, with $T_{\mathrm{g}} = 1$ gating prompt and $T_{\mathrm{e}} = 4$ expert prompts. The hyperparameters λ_{div} and λ_{aux} are both set to 0.1, and the threshold is set to $c_{\mathrm{th}} = 0.5$.

Table 1: Comparison of methods on DOLOS, BOL, and MU3D. Missing entries are marked with "–", and bold values indicate the best results. All values are reported as percentages (%). Rows highlighted in light blue, green, and red correspond to Visual-only, Audio-only, and Fusion-based models, respectively. F, A, and L denote Face, Audio, and Limb modalities.

Method	DOLOS			BOL			MU3D		
	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC
LieNet (F) [27]	59.67	73.24	53.14	55.65	30.30	51.07	57.19	57.62	57.19
FCN (F) [35]	60.98	68.65	61.99	56.23	**63.26**	59.53	57.64	59.13	50.89
DDABG (F) [26]	55.47	62.52	52.13	56.66	55.17	57.89	55.63	51.99	52.20
PECL (F) [18]	61.44	69.42	58.89	58.94	43.06	58.44	54.68	47.19	54.64
AFFAKT (F) [24]	67.64	70.54	72.12	–	–	–	–	–	–
DLF-BRAM (F+L) [57]	78.86	82.27	77.72	60.05	55.63	**61.07**	56.63	58.42	55.74
LieNet (A) [27]	57.14	72.72	50.00	56.15	24.38	52.82	53.73	44.52	53.73
PECL (A) [18]	59.19	73.46	52.54	58.43	56.70	57.29	56.25	45.72	56.10
AFFAKT (A) [24]	61.98	68.22	63.91	–	–	–	–	–	–
LieNet (F+A) [27]	56.50	69.72	51.02	59.78	58.14	58.09	53.48	33.62	54.75
PECL (F+A) [18]	64.75	71.20	62.71	59.51	51.06	59.41	55.31	60.07	55.31
AFFAKT (F+A) [24]	68.10	70.73	72.26	–	–	–	–	–	–
MoMCE (F+A) (Our)	**79.10**	**82.38**	**78.09**	**61.09**	52.91	60.93	**58.58**	**60.35**	**58.69**

4.2 Dataset and Protocol

We evaluate the performance of MoMCE on three datasets: DOLOS [18], Bag-of-Lies (BOL) [21], and MU3D [30]. These datasets cover a variety of scenarios, including game shows, spontaneous lies, and real-world simulations, allowing for comprehensive evaluation of the model under diverse conditions. For the DOLOS dataset, we follow the three official train-test splits provided by the dataset and report the average results. For the Bag-of-Lies dataset, we use the B group containing all video samples. In addition, we follow the official participant-based split [21] and perform three-fold cross-validation. The MU3D dataset contains 320 videos from 80 participants across four demographic groups (20 Black males, 20 Black females, 20 White males, and 20 White females). We therefore adopt a 4-fold based on these groups to mitigate the influence of participant identity, gender, and race [32]. All compared methods within the same dataset follow the same evaluation protocol. Detailed dataset descriptions and example visualizations are provided in the Appendix.2.

4.3 In-Domain Evaluation

Under in-domain evaluation, we compare MoMCE with state-of-the-art automatic deception detection (ADD) methods, with the results on DOLOS, Bag-of-Lies, and MU3D shown in Tables 1. Recent ADD baselines mainly include

Table 2: Cross-dataset results on DOLOS (D), Bag-of-Lies (B), and MU3D (M); "X&Y→Z" denotes training on X and Y, testing on Z. All values are reported as percentages (%). Rows with light blue, green, and red backgrounds denote Face-only, Audio-only, and Fusion models, respectively.

Method	M&B → D			D&M → M			D&B → B			Average		
	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC
LieNet [27]	53.42	69.62	49.84	52.19	62.77	52.19	55.38	42.23	55.32	53.66	58.21	52.45
FCN [35]	48.84	31.68	50.63	52.56	48.97	51.26	57.10	61.28	55.99	52.84	47.31	52.63
PECL [18]	54.01	70.09	50.08	51.25	50.63	51.25	54.46	42.19	54.40	53.24	54.30	51.91
LieNet [27]	53.91	63.50	**52.29**	53.75	36.21	53.75	52.92	49.17	52.90	53.53	49.63	52.98
PECL [18]	54.63	68.80	51.36	51.56	**67.09**	51.56	57.54	**63.68**	57.59	54.58	**66.52**	53.50
LieNet [27]	54.40	68.23	51.53	54.69	50.51	54.69	51.08	59.75	51.14	53.39	59.50	52.45
PECL [18]	54.51	69.55	50.92	51.25	66.95	51.25	55.38	52.46	55.37	53.71	62.99	52.51
MoMCE	**57.80**	**72.88**	50.71	**56.25**	60.49	**55.26**	**59.20**	56.68	**59.17**	**57.75**	63.35	**55.05**

LieNet [27], FacialCueNet [35], DDABG [26], PECL [18], AFFAKT [24], and DLF-BRAM [57]. Most benchmark results are taken from prior work [57].

Results and Analysis. As shown in Table 1, MoMCE achieves significant improvements on accuracy (ACC) over existing methods: on the DOLOS dataset, ACC increases by 0.24% compared to DLF-BRAM [57] (78.86%→79.10%), and on the BOL dataset by 1.04% (60.05%→61.09%). Similar gains are observed in F1 and AUC on DOLOS. These improvements are consistent across Bag-of-Lies and MU3D. It is worth noting that our method yields slightly lower F1 than FacialCueNet [35], mainly because FacialCueNet leverages multiple handcrafted facial deception features to boost recall. However, handcrafted features often have limited generalization in complex scenarios, which explains why FacialCueNet underperforms in ACC and AUC. Moreover, DLF-BRAM [57] combines facial and limb cues, and the Bag-of-Lies dataset contains substantial body and hand movements, which gives it a slight advantage in terms of AUC. Overall, our method achieves the best performance on seven out of nine metrics across the three datasets, fully demonstrating its effectiveness. The further analysis of computational efficiency is provided in Appendix.7.

4.4 Cross-Domain Evaluation

To further assess the effectiveness of MoMCE, we conduct *cross-domain evaluation* in Table 2. Across three datasets and three metrics (12 metrics in total), MoMCE achieves the best performance on 8 metrics. Notably, the average precision is improved from 54.58% (PECL(A)) to 57.75%. Compared to the *in-domain setting*, the *cross-domain setting* is considerably more challenging due to domain shifts [45, 56]. For instance, DOLOS consists of drama scenes, whereas Bag-of-lies and MU3D mainly contain laboratory scenes, leading to different levels of

Table 3: Ablations on MoME (Mixture of Modality Experts) and MoCE (Mixture of Cue Experts). All values are reported as percentages (%).

MoME	MoCE	DOLOS			BOL			MU3D		
		ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC
✓		61.73	69.85	58.97	59.87	41.94	58.30	57.50	58.35	57.50
	✓	65.62	71.65	63.86	60.40	44.14	59.44	57.63	54.64	53.64
✓	✓	**79.10**	**82.38**	**78.09**	**61.09**	**52.91**	**60.93**	**58.58**	**60.35**	**58.69**

Table 4: Ablations on CEDL and CEWL loss functions. All values are reported as percentages (%).

Methods	DOLOS			BOL			MU3D		
	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC
w/o CEDL	75.65	79.62	74.44	60.09	33.81	57.02	57.81	43.01	57.35
w/o CEWL	76.52	79.55	76.00	60.56	46.87	57.95	58.44	49.26	58.44
Full	**79.10**	**82.38**	**78.09**	**61.09**	**52.91**	**60.95**	**58.58**	**60.35**	**58.69**

cue exposure. Despite these challenges, MoMCE consistently outperforms all competing methods, demonstrating its strong generalization ability.

4.5 Ablation Study and Analysis

Effectiveness of MoME and MoCE. We evaluate the key components of MoMCE, including MoME and MoCE, on three datasets, as shown in Table 3. MoME yields a significant performance improvement. This demonstrates that dynamically fusing modality features according to sample-level importance enables the model to leverage complementary information from each modality more effectively, leading to better multi-modal representation learning. Furthermore, we observe that integrating MoCE also leads to a significant performance improvement. This result highlights the necessity of modeling heterogeneous cues within each modality, and shows that MoCE can capture multiple informative cues, filling the gap left by a single-encoder design.

Effectiveness of CEDL and CEWL. We conduct ablation studies on CEDL and CEWL, as reported in Table 4. Both losses consistently improve performance, suggesting that without additional regularization, the model tends to over-focus on homogeneous cues and ignore discriminative information. CEDL explicitly encourages the model to learn diverse cues, while CEWL mitigates the modality conflict issue by enforcing balanced learning on conflict samples.

Hyper-parameter Analysis. To validate the robustness of MoMCE, we conduct a sensitivity analysis on its hyperparameters. We evaluate the overall performance of CEWL weight λ_{aux}, CEWL weight λ_{div}, different threshold c_{th},

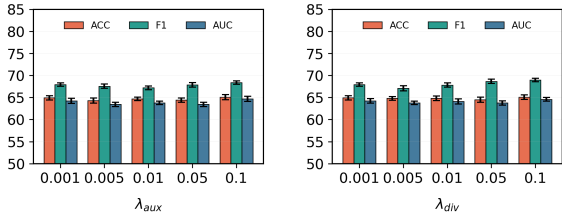

Fig. 3: Effect of λ_{aux} and λ_{div} on ACC, F1, and AUC.

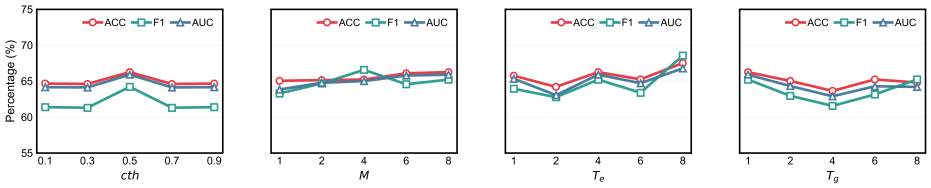

Fig. 4: Effects of varying c_{th}, M, T_{e} and T_{g} on model performance. ACC, F1, and AUC are plotted to illustrate trends and relative stability across hyperparameter settings.

numbers of experts M, numbers of expert prompts T_{e}, and numbers of gating prompts T_{g} on the three datasets. Specifically, the sensitivity analysis is performed by varying one target hyperparameter while keeping the others fixed at their baseline values used in the experiments. Figure 3 and Figure 4 show the average results of different hyperparameter settings across the datasets. The results indicate that the overall evaluation metrics remain relatively stable, indicating that the proposed method is not sensitive to the values of these hyperparameters.

Visualization of Feature Distributions. We also visualize the feature distributions of MoMCE, MoMCE without MoCE, and MoMCE without MoME in Figure 5, providing a qualitative comparison. We conduct the experiment on the test set of Protocol 1 from the DOLOS dataset, projecting sample features into a two-dimensional space using t-SNE. The results show that the feature distributions of MoMCE without MoCE and MoMCE without MoME are irregular, while the features of MoMCE are more compact and consistent, with clearer separability among classes. These results indicate that MoMCE enhances feature discriminability, leading to more accurate deception detection.

Visualization of Modal Weight Distribution. To evaluate the effectiveness of MoME on conflicting samples, we visualize the modality weights of several examples from the DOLOS dataset, as illustrated in Fig. 6. The upper part indicates the labels of beneficial and conflicting modalities, while the lower part presents the learned modality weights, where darker colors represent higher weights. When modality conflicts occur, MoME automatically assigns higher weights to the correct modality. Moreover, the learned weights exhibit strong sparsity, a behavior explicitly encouraged by the CEWL design. These observations demonstrate that MoME can reliably select trustworthy modalities and thereby effectively alleviate modality conflict.

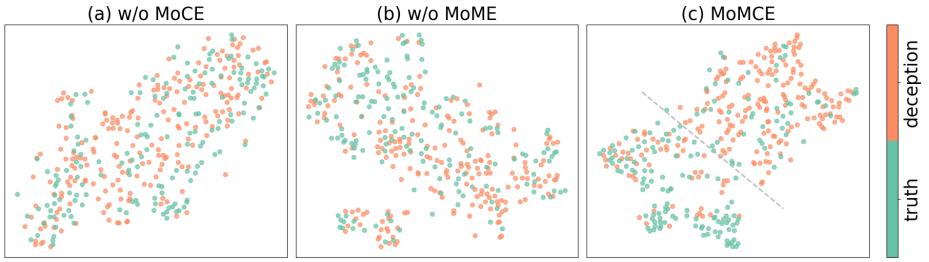

Fig. 5: t-SNE visualization of feature distribution on DOLOS.

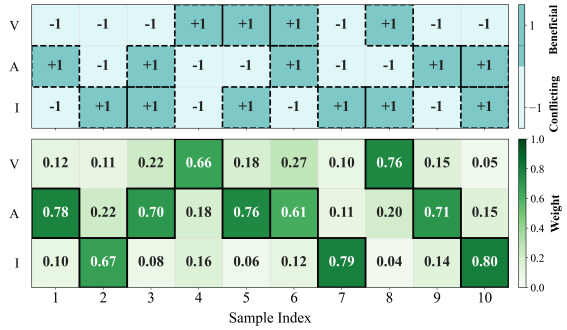

Fig. 6: Visualization of conflicting samples and modality weights on DOLOS. Darker colors indicate larger weights.

Expert Top-1 Statistics and CAM Visualization. To verify whether the model truly captures the heterogeneity of cues within the visual modality, we conduct analyses from two perspectives: (i) Top-1 expert activation statistics (see Figure 7), and (ii) CAM visualizations of expert groups (see Figure 8). (i) The Top-1 activation statistics reveal clear activation differences among experts across the visual, audio, and interaction branches, indicating that the model dynamically routes different samples to different experts and thus forms a stable division-of-labor pattern. (ii) The CAM visualizations further show that Visual Expert-7 predominantly attends to the eye and mouth regions and their surrounding areas, whereas Visual Expert-3 mainly focuses on the nasal bridge and its adjacent regions. The distinct spatial attention patterns between these two experts demonstrate that the model learns different types of cues within the visual modality. Above findings confirm that our MoMCE architecture can effectively model the heterogeneous cues embedded in the visual modality. The further qualitative analysis of audio and interaction experts is provided in Appendix.5.

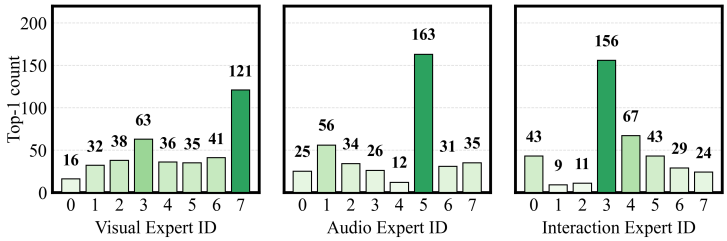

Fig. 7: Top-1 selection frequency of each expert across the 382 test samples in the DOLOS dataset for the three modalities.

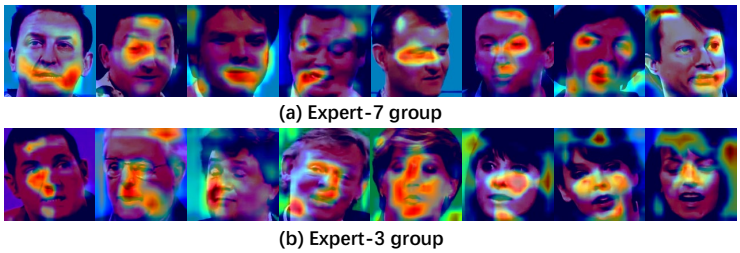

Fig. 8: Representative CAM visualizations of samples assigned to Visual Expert-7 and Visual Expert-3 group.

5 Conclusion

In this paper, we address the challenges of modality conflicts and cues heterogeneity in audiovisual deception detection. We propose a Mixture of Modality and Cue Experts, MoMCE, which consists of two key components: a prompt-aware mixture of modality experts that adaptively adjusts modality contributions with consistency-aware expert weight regularization, and a prompt-aware mixture of cue experts that captures diverse and heterogeneous cues within each modality with cue diversity regularization. Extensive experiments demonstrate that MoMCE achieves superior accuracy compared to existing methods.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (Nos. U22A2035, U1736206, U1803262, 62576076, and 62411540034), the Guangdong Basic and Applied Basic Research Foundation (Nos. 2023A1515140037 and 2026A1515010184), the Open Topics from the Lion Rock Labs of Cyberspace Security (Project No. LRL24009), the CCF-Tencent Rhino-Bird Open Research Fund, Guangdong Research Team for Communication and Sensing Integrated with Intelligent Computing (Project No. 2024KCXTD047), the SongShan Lake HPC Center (SSL-HPC) at Great Bay University.

References

1. Abouelenien, M., Pérez-Rosas, V., Mihalcea, R., Burzo, M.: Detecting deceptive behavior via integration of discriminative features from multiple modalities. IEEE Transactions on Information Forensics and Security **12**(5), 1042–1055 (2016)
2. Abouelenien, M., Pérez-Rosas, V., Mihalcea, R., Burzo, M.: Detecting deceptive behavior via integration of discriminative features from multiple modalities. IEEE Transactions on Information Forensics and Security **12**(5), 1042–1055 (2016)
3. Akbari, H., Kondratyuk, D., Cui, Y., Hornung, R., Wang, H., Adam, H.: Alternating gradient descent and mixture-of-experts for integrated multimodal perception. Advances in Neural Information Processing Systems **36**, 79142–79154 (2023)
4. Baevski, A., Zhou, Y., Mohamed, A., Auli, M.: wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in neural information processing systems **33**, 12449–12460 (2020)
5. Bajaj, N., Rajwadi, M., Constance, T.G., Wall, J., Moniri, M., Laird, T., Woodruff, C., Laird, J., Glackin, C., Cannings, N.: Deception detection in conversations using the proximity of linguistic markers. Knowledge-Based Systems **267**, 110422 (2023)
6. Cao, B., Sun, Y., Zhu, P., Hu, Q.: Multi-modal gated mixture of local-to-global experts for dynamic image fusion. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 23555–23564 (2023)
7. Chebbi, S., Jebara, S.B.: Deception detection using multimodal fusion approaches. Multimedia Tools and Applications **82**(9), 13073–13102 (2023)
8. Cheng, Y., Li, Y., He, J., Feng, R.: Mixtures of experts for audio-visual learning. Advances in Neural Information Processing Systems **37**, 219–243 (2024)
9. Cheng, Z., Chen, S., Yang, B., Guan, Y., Chen, J., Lian, Z., Peng, X., Ma, F., Cui, L., Tian, Q.: Omniopsd: Rationale-privileged on-policy self-distillation for affective computing. arXiv preprint arXiv:2606.15920 (2026)
10. DePaulo, B.M., Lindsay, J.J., Malone, B.E., Muhlenbruck, L., Charlton, K., Cooper, H.: Cues to deception. Psychological bulletin **129**(1), 74 (2003)
11. Ding, M., Zhao, A., Lu, Z., Xiang, T., Wen, J.R.: Face-focused cross-stream network for deception detection in videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7802–7811 (2019)
12. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
13. Fang, H.S., Xie, S., Tai, Y.W., Lu, C.: Rmpe: Regional multi-person pose estimation. In: Proceedings of the IEEE international conference on computer vision. pp. 2334–2343 (2017)
14. Fang, Y., Huang, W., Wan, G., Su, K., Ye, M.: Emoe: Modality-specific enhanced dynamic emotion experts. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 14314–14324 (2025)
15. Fornaciari, T., Poesio, M.: Automatic deception detection in italian court cases. Artificial intelligence and law **21**, 303–340 (2013)
16. Gao, Z., Hu, D., Jiang, X., Lu, H., Shen, H.T., Xu, X.: Enhanced experts with uncertainty-aware routing for multimodal sentiment analysis. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 9650–9659 (2024)
17. Gogate, M., Adeel, A., Hussain, A.: Deep learning driven multimodal fusion for automated deception detection. In: 2017 IEEE symposium series on computational intelligence. pp. 1–6 (2017)

18. Guo, X., Selvaraj, N.M., Yu, Z., Kong, A.W.K., Shen, B., Kot, A.: Audio-visual deception detection: Dolos dataset and parameter-efficient crossmodal learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22135–22145 (2023)
19. Guo, X., Yu, Z., Selvaraj, N.M., Shen, B., Kong, A.W.K., Kot, A.C.: Benchmarking cross-domain audio-visual deception detection. arXiv preprint arXiv:2405.06995 (2024)
20. Guo, Y., Nie, G., Zou, X., Zhu, D., Gao, W., Liao, M.: Styles energy consumption analysis of lane-changing maneuvers in autonomous vehicles: The role of driving styles. International Journal of Electrical Power & Energy Systems **165**, 110489 (2025)
21. Gupta, V., Agarwal, M., Arora, M., Chakraborty, T., Singh, R., Vatsa, M.: Bag-of-lies: A multimodal dataset for deception detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. pp. 0–0 (2019)
22. Hsiao, S.W., Sun, C.Y.: Attention-aware multi-modal rnn for deception detection. In: 2022 IEEE International Conference on Big Data (Big Data). pp. 3593–3596 (2022)
23. Huai, T., Zhou, J., Wu, X., Chen, Q., Bai, Q., Zhou, Z., He, L.: Cl-moe: Enhancing multimodal large language model with dual momentum mixture-of-experts for continual visual question answering. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 19608–19617 (2025)
24. Ji, Z., Tian, X., Liu, Y.: Affakt: A hierarchical optimal transport based method for affective facial knowledge transfer in video deception detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 1336–1344 (2025)
25. Jia, M., Tang, L., Chen, B.C., Cardie, C., Belongie, S., Hariharan, B., Lim, S.N.: Visual prompt tuning. In: European conference on computer vision. pp. 709–727. Springer (2022)
26. Kang, J., Qu, W., Cui, S., Feng, X.: Deception detection algorithm based on global and local feature fusion with multi-head attention. In: 2024 3rd International Conference on Image Processing and Media Computing (ICIPMC). pp. 162–168. IEEE (2024)
27. Karnati, M., Seal, A., Yazidi, A., Krejcar, O.: Lienet: A deep convolution neural network framework for detecting deception. IEEE transactions on cognitive and developmental systems **14**(3), 971–984 (2021)
28. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. nature **521**(7553), 436–444 (2015)
29. Lin, R., Xiong, A., He, Q., Mai, S., Huang, L., Hu, H.: From uni-to multi-modal: Progressive multi-source domain adaptation with gradient disentangling for audio-visual deception detection. Machine Intelligence Research **23**(2), 484–500 (2026)
30. Lloyd, E.P., Deska, J.C., Hugenberg, K., McConnell, A.R., Humphrey, B.T., Kunstman, J.W.: Miami university deception detection database. Behavior research methods **51**(1), 429–439 (2019)
31. Ma, F., Yuan, Y., Xie, Y., Ren, H., Liu, I., He, Y., Ren, F., Yu, F.R., Ni, S.: Generative technology for human emotion recognition: A scoping review. Information Fusion **115**, 102753 (2025)
32. Mambreyan, A., Punskaya, E., Gunes, H.: Dataset bias in deception detection. In: 2022 26th International Conference on Pattern Recognition (ICPR). pp. 1083–1089. IEEE (2022)

33. Mathur, L., Matarić, M.J.: Introducing representations of facial affect in automated multimodal deception detection. In: Proceedings of the 2020 international conference on multimodal interaction. pp. 305–314 (2020)
34. Mustafa, B., Riquelme, C., Puigcerver, J., Jenatton, R., Houlsby, N.: Multimodal contrastive learning with limoe: the language-image mixture of experts. Advances in Neural Information Processing Systems **35**, 9564–9576 (2022)
35. Nam, B., Kim, J.Y., Bark, B., Kim, Y., Kim, J., So, S.W., Choi, H.Y., Kim, I.Y.: Facialcuenet: unmasking deception-an interpretable model for criminal interrogation using facial expressions: Iy kim et al. Applied Intelligence **53**(22), 27413–27427 (2023)
36. Peng, X., Wei, Y., Deng, A., Wang, D., Hu, D.: Balanced multimodal learning via on-the-fly gradient modulation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8238–8247 (2022)
37. Pérez-Rosas, V., Abouelenien, M., Mihalcea, R., Burzo, M.: Deception detection using real-life trial data. In: Proceedings of the 2015 ACM on international conference on multimodal interaction. pp. 59–66 (2015)
38. Rill-Garcia, R., Jair Escalante, H., Villasenor-Pineda, L., Reyes-Meza, V.: High-level features for multimodal deception detection in videos. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. pp. 0–0 (2019)
39. Şen, M.U., Perez-Rosas, V., Yanikoglu, B., Abouelenien, M., Burzo, M., Mihalcea, R.: Multimodal deception detection using real-life trial data. IEEE Transactions on Affective Computing **13**(1), 306–319 (2020)
40. Shen, S., Yao, Z., Li, C., Darrell, T., Keutzer, K., He, Y.: Scaling vision-language models with sparse mixture of experts. arXiv preprint arXiv:2303.07226 (2023)
41. Shi, Y., Gao, Y., Lai, Y., Wang, H., Feng, J., He, L., Wan, J., Chen, C., Yu, Z., Cao, X.: Shield: An evaluation benchmark for face spoofing and forgery detection with multimodal large language models. Visual Intelligence **3**(1), 9 (2025)
42. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
43. Wang, A., Sun, X., Xie, R., Li, S., Zhu, J., Yang, Z., Zhao, P., Han, J., Kang, Z., Wang, D., et al.: Hmoe: Heterogeneous mixture of experts for language modeling. arXiv preprint arXiv:2408.10681 (2024)
44. Wang, L., Gao, H., Zhao, C., Sun, X., Dai, D.: Auxiliary-loss-free load balancing strategy for mixture-of-experts. arXiv preprint arXiv:2408.15664 (2024)
45. Wang, M., Hu, R., Zhu, X., Zhu, D., Wang, X.: Learning with noisy labels for robust fatigue detection. Knowledge-Based Systems **300**, 112199 (2024)
46. Wang, M., Zhu, X., Hu, R., Zhu, D., Liao, L., Ye, M.: Mdfg: Multi-dimensional fine-grained modeling for fatigue detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 12730–12738 (2025)
47. Wang, T., Lin, X., Xu, Y., Ye, Q., Guo, D., Escalera, S., Khoriba, G., Yu, Z.: Micro-gesture recognition: A comprehensive survey of datasets, methods, and challenges. Machine Intelligence Research **23**(2), 308–330 (2026)
48. Wei, Y., Li, S., Feng, R., Hu, D.: Diagnosing and re-learning for balanced multimodal learning. In: European Conference on Computer Vision. pp. 71–86. Springer (2024)
49. Wu, Z., Singh, B., Davis, L., Subrahmanian, V.: Deception detection in videos. In: Proceedings of the AAAI conference on artificial intelligence. vol. 32 (2018)

50. Yun, S., Choi, I., Peng, J., Wu, Y., Bao, J., Zhang, Q., Xin, J., Long, Q., Chen, T.: Flex-moe: Modeling arbitrary modality combination via the flexible mixture-of-experts. Advances in Neural Information Processing Systems **37**, 98782–98805 (2024)
51. Zhang, H., Ding, Y., Cao, L., Wang, X., Feng, L.: Fine-grained question-level deception detection via graph-based learning and cross-modal fusion. IEEE Transactions on Information Forensics and Security **17**, 2452–2467 (2022)
52. Zhang, J., Lin, X., Huang, J., Ye, S., Guo, X., Zhu, D., Hu, R., Guo, D., Liang, Y., Yu, Z., et al.: Multimodal deception detection: A survey. Machine Intelligence Research **23**(2), 284–307 (2026)
53. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16816–16825 (2022)
54. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International journal of computer vision **130**(9), 2337–2348 (2022)
55. Zhu, D., Hu, R., Song, S., Guo, X., Li, X., Wang, Z.: Cross-illumination video anomaly detection benchmark. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 2516–2525 (2023)
56. Zhu, D., Niu, Z., Zhao, B., Huang, J., Ye, S., Lin, X., Ma, H., Wang, T., Zhang, J., Zhu, C., et al.: Svc 2026: the second multimodal deception detection challenge and the first domain generalized remote physiological measurement challenge. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1498–1506 (2026)
57. Zhu, D., Zhang, C., Hu, R., Wang, M., Liao, L., Ye, M.: Detecting deceptive behavior via learning relation-aware visual representations. IEEE Transactions on Information Forensics and Security (2025)
58. Zhuo, Y., Baskaran, V.M., Kiaw, L.W.Y., Phan, R.C.W.: Video deception detection through the fusion of multimodal feature extraction and neural networks. In: 2024 International Joint Conference on Neural Networks (IJCNN). pp. 1–8. IEEE (2024)