DecepGPT: Schema-Driven Deception Detection with Multicultural Datasets and Robust Multimodal Learning

Jiajian Huang[1†], Dongliang Zhu[2†], Zitong Yu[1,4✉], Hui Ma[1✉], Jiayu Zhang[1], Chunmei Zhu[3], and Xiaochun Cao[3]

[1] Great Bay University, China
{yuzitong,mah}@gbu.edu.cn
[2] Wuhan University, China
[3] Sun Yat-sen University, China
[4] Dongguan Key Laboratory for Intelligence and Information Technology, China

Abstract. Multimodal deception detection aims to identify deceptive behavior by analyzing audiovisual cues for forensics and security. In these high-stakes settings, investigators need verifiable evidence connecting audiovisual cues to final decisions, along with reliable generalization across domains and cultural contexts. However, existing benchmarks provide only binary labels without intermediate reasoning cues. Datasets are also small with limited scenario coverage, leading to shortcut learning. We address these issues through three contributions. First, we construct reasoning datasets by augmenting existing benchmarks with structured cue-level descriptions and reasoning chains, enabling models to output auditable reports. Second, we release T4-Deception, a multicultural dataset based on the unified "To Tell the Truth" television format implemented across four countries. With 1695 samples, it is the largest non-laboratory deception detection dataset. Third, we propose two modules for robust learning under small-data conditions. Stabilized Individuality-Commonality Synergy (SICS) refines multimodal representations by combining learnable global priors with sample-adaptive residuals and applying polarity-aware recalibration. Distilled Modality Consistency (DMC) aligns modality-specific predictions with the fused multimodal predictions via knowledge distillation to prevent unimodal shortcut learning. Experiments on three established benchmarks and our novel dataset demonstrate that our method achieves state-of-the-art performance in both in-domain and cross-domain scenarios, while exhibiting superior transferability across diverse cultural contexts. The datasets and code are available at this link.

Keywords: Multimodal Deception Detection · Auditable Reasoning · Multicultural Dataset · Stabilized Representation · Modality Consistency

† Equal contribution.
✉ Corresponding authors.

1 Introduction

Multimodal deception detection (MDD) aims to identify deceptive behavior by analyzing audio and visual cues [25, 34, 39], which provide objective decision support in high-stakes social analysis, such as forensic investigation and security screening [6], where human judgment is often subject to cognitive bias [2].

Recent progress in MDD has evolved from handcrafted behavioral descriptors [26] to end-to-end audiovisual deep learning models [22, 37]. However, as illustrated in Fig. 1a, traditional MDD methods are predominantly label-centric, focusing on optimizing binary classification accuracy. While achieving competitive performance, these methods typically provide only a final binary prediction. In forensic and legal contexts, a standalone label is insufficient. Investigators need to understand why a sample is flagged as deceptive, with evidence connecting behavioral cues such as micro-expressions and voice prosody to the final decision. Furthermore, MDD methods must demonstrate generalization across diverse cultural contexts to satisfy the requirements of real-world applications [27].

Existing benchmarks [6, 8] provide only binary labels without intermediate reasoning cues, preventing models from producing verifiable evidence. Moreover, existing datasets have limited scenario coverage. Important factors, such as identity pretense and cross-cultural variations, are still not well studied, which limits the generalization ability of MDD methods. In addition, the small scale of available data often causes models to learn spurious correlations [5] during training.

We address these issues through three contributions. First, we construct a reasoning dataset by augmenting existing benchmarks with structured cue-level descriptions and reasoning chains, enabling the generation of auditable reports as shown in Fig. 1b. Second, we release T4-Deception, a multicultural dataset covering identity pretense across four countries (the U.S., Germany, Vietnam, and Bulgaria) under a unified 'To Tell the Truth' television format. With 1695 samples, it is the largest non-laboratory deception benchmark to date. Third, we propose two modules for robust learning under small-data conditions. Stabilized Individuality-Commonality Synergy (SICS) refines multimodal features through a polarity-aware adjustment mechanism, which synergizes a learnable global prior with sample-adaptive residuals to enhance or suppress specific feature dimensions. Meanwhile, Distilled Modality Consistency (DMC) introduces a consistency regularizer that aligns unimodal predictions with multimodal teacher predictions via knowledge distillation to mitigate shortcut learning.

In summary, the main contributions of this article are as follows:

- **Reasoning Dataset.** We provide a standardized pipeline to enrich existing benchmarks with structured audiovisual cues and reasoning chains, enabling the generation of auditable reports for verifiable decision-making.
- **Multicultural Dataset.** We release T4-Deception (To Tell the Truth across 4 cultures), a large-scale dataset covering identity pretense across four countries under a unified television format. With 1695 samples, it is currently the largest non-laboratory dataset in the field.

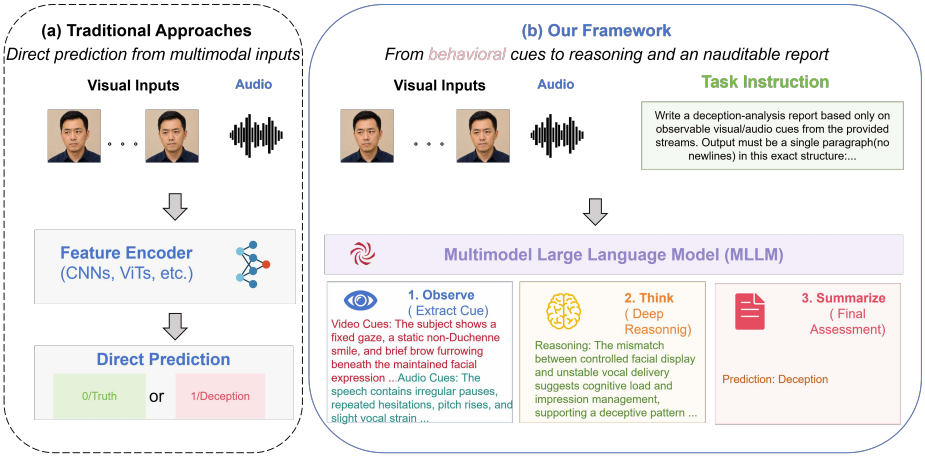


Fig. 1: Comparison between the traditional paradigm and our auditable framework. (a) Traditional methods directly map behavioral signals to binary labels, leaving the decision process opaque. (b) Our method generates structured reports with audiovisual cues and reasoning paths, making the final judgment auditable.

– **Robust Multimodal Learning Modules.** We propose Stabilized Individuality-Commonality Synergy (SICS) for polarity-aware feature refinement and Distilled Modality Consistency (DMC) for modality consistency distillation, which effectively improve both in-domain and cross-domain performance.

2 Related Work

Multimodal Deception Detection. Early MDD studies relied on handcrafted behavioral cues, including facial expressions, gestures, prosody, and linguistic indicators [26]. Later methods adopted deep audiovisual models to capture fine-grained deception patterns, using CNN/ResNet-based visual encoders [4, 14, 34] and attention-, graph-, or video-based multimodal architectures [10, 36, 40, 41]. Recent works further improve cross-modal learning [6] and generalization [7, 16]. AFAFFAKT [13] explores affective knowledge transfer to mitigate the small-data challenge in deception detection. However, most existing methods still formulate MDD as binary truthful/deceptive classification, offering limited evidence-level reasoning or auditability. Although MLLMs provide new opportunities for semantic reasoning and interpretable decision support [17], they remain vulnerable to hallucinated rationales in deception scenarios [21]. We address this issue by augmenting deception detection datasets with structured reasoning chains and training model to standardize evidence extraction for audition.

MLLM Reasoning. Recent MLLMs can generate free-form chain-of-thought (CoT) or rationale-style explanations for visual and video reasoning tasks [20,38]. However, in small-scale multimodal deception detection, such reasoning may be vulnerable to overfitting, unimodal shortcuts, and hallucinated rationales, especially when subtle behavioral cues such as facial dynamics, micro-gestures [33],

and vocal patterns are weak or ambiguous. To address this, we formulate MLLM reasoning as schema-constrained audiovisual evidence tracing, and introduce SICS to stabilize multimodal representations and DMC to reduce unimodal shortcut learning.

Behavioral Decoupling and Stabilized Refinement. Identity- or sample-specific variations may obscure shared behavioral cues [9]. Conventional decomposition [30] and modulation paradigms [1] provide ways to separate features or adaptively weight samples. Motivated by this, SICS focuses on stabilizing multimodal representations by jointly modeling behavioral commonality and sample-specific individuality. It combines a learnable global prior with a sample-adaptive residual through gating, and applies polarity-aware adjustment to enhance informative dimensions while suppressing noisy ones.

Mitigating Unimodal Dominance. Multimodal optimization often suffers from imbalanced gradients, leading to unimodal shortcuts [24]. Representative approaches alleviate this issue through modality dropout [23] or adaptive reweighting [11]. DMC addresses this issue during training by attaching auxiliary prediction heads to the visual and audio branches and aligning their predictions with the schema-conditioned multimodal prediction. This guides each modality branch to learn evidence consistent with the final audiovisual judgment, rather than relying on isolated unimodal shortcuts.

3 Method

To provide transparency for sensitive decision-making, we propose an auditable MDD framework. As shown in Fig. 2, we first enrich existing benchmarks with structured behavioral descriptions and reasoning chains. We then introduce T4-Deception (Table 1 and Fig. 3) to cover identity pretense across multicultural scenarios. Finally, we integrate two robust encoding modules (Fig. 4): SICS, which stabilizes representations by balancing shared behavioral commonalities and sample-specific variations, and DMC, which mitigates unimodal shortcuts through consistency distillation. Together, these components enable schema-constrained audit reports grounded in audiovisual cues.

3.1 Data Construction: Structured Multi-Cue Reasoning Supervision

Schema Design. We use a structured output format: `[Video Cues; Audio Cues; Reasoning; Prediction]`. The model first identifies deception-related audio and visual cues, then performs cross-modal reasoning based on these cues to derive the final prediction. This enforces an evidence-to-conclusion inference flow, where the prediction is derived from explicit behavioral observations through intermediate reasoning steps.

HITL Generation Pipeline. As shown in Fig. 2, our pipeline uses multiple specialized assistants. The Audio Assistant (Qwen-Omni) extracts audio cues such as prosody and speech patterns. The Video Assistant (GPT-4o) extracts

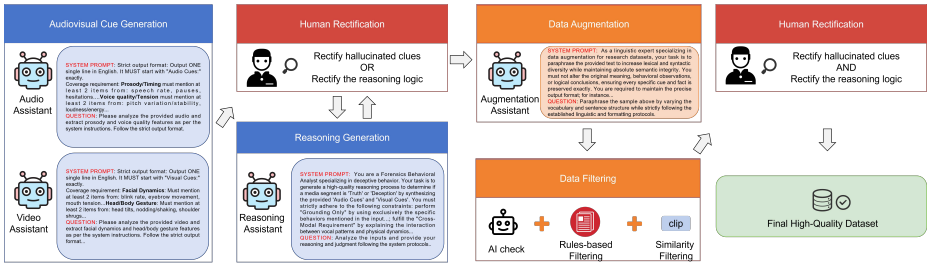


Fig. 2: Overview of reasoning dataset generation pipeline. The pipeline adopts a Human-in-the-Loop (HITL) framework to ensure high-quality, auditable structured reports. It begins with AI-driven audiovisual cue extraction, followed by human-guided rectification of hallucinations. A reasoning assistant then synthesizes these cues into forensic judgments. The data is further enriched through semantic augmentation and a multi-tiered filtering stage (comprising AI, rules-based, and CLIP-similarity checks) to produce the final high-fidelity benchmark.

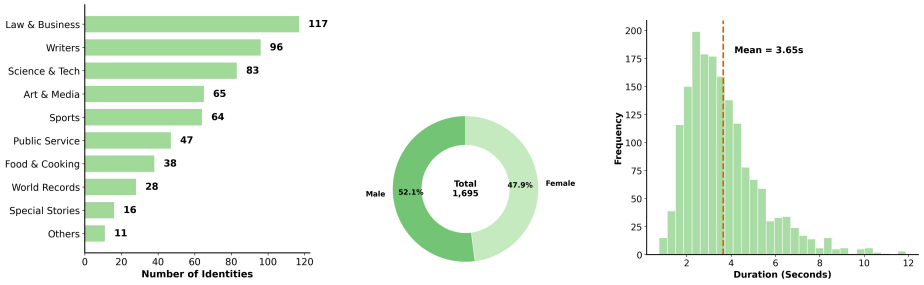
Table 1: Comparison of multimodal deception detection datasets. T4-Deception (To Tell the Truth across 4 cultures) dataset is the largest non-laboratory benchmark, featuring a unified identity pretense task across multiple cultural contexts.

Dataset	Subject	Clip	Deceptive	Truthful	Setting	Deceptive Task
Real Life Trials [25]	56	121	61	60	Courtroom	False Testimony
Bag of Lies [8]	35	325	162	163	Laboratory	False Image-Narration
MU3D [18]	80	320	160	160	Laboratory	False Social-Evaluation
Box of Lies [28]	26	1049	862	187	Game Show	False Object-Description
DOLOs [6]	213	1675	899	776	Game Show	False Story-Telling
T4-Deception (Ours)	1695	**1695**	**1130**	**565**	**Game Show**	**False Professional-Identity**
— *U.S. Edition*	876	876	584	292	Game Show	(Unified Multi-Culture)
— *German Edition*	702	702	468	234	Game Show	
— *Vietnam Edition*	66	66	44	22	Game Show	
— *Bulgarian Edition*	51	51	34	17	Game Show	

visual cues including facial dynamics and body language. A Reasoning Assistant (GPT-4o) then synthesizes these cues into a cross-modal judgment. Human annotators review the outputs from these three assistants to correct hallucinations or logical inconsistencies. An Augmentation Assistant (GPT-4o) then paraphrases the text to increase lexical variation. The augmented samples pass through automated filters before a final human review.

Multi-stage Filtering. After the data augmentation process, we apply three rigorous filtering stages to the augmented samples. First, AI-based checks automatically remove contradictory cue-reasoning pairs to ensure logical consistency. Second, rule-based filtering strictly ensures full format compliance across all entries. Third, similarity filtering effectively prevents data redundancy by eliminating near-duplicate samples. Finally, the filtered samples undergo a comprehensive final human review to confirm overall data quality and reliability.

Annotation Reliability. Two annotators performed correction and cross-validation after three pre-annotation calibration rounds. Inter-annotator agreement is com-



(a) Identity **(b)** Gender **(c)** Duration

Fig. 3: Dataset statistics of T4-Deception. We illustrate: (a) distribution of identities, where each one of total 565 identities is shared by one truthful and two deceptive participants; (b) balanced gender distribution; and (c) numerous short-term deceptive segments with an average temporal duration of 3.65s.

puted on discrete cue/reasoning audit labels derived from human-corrected annotations. **Cohen's κ is 0.73**, indicating substantial agreement.

3.2 T4-Deception (To Tell the Truth across 4 cultures) Dataset

To address cross cultures data scarcity in deception detection, we present T4-Deception dataset, which is also the largest non-laboratory benchmark for multimodal deception detection. T4-Deception studies identity pretense across the United States, Germany, Vietnam, and Bulgaria under a unified game-show format. Detailed source collection, filtering criteria, and label verification protocols are provided in Appendix 1. As detailed in Table 1, T4-Deception differs from existing game-show datasets that focus on object fabrication (Box of Lies [28]) or story fabrication (DOLOs [6]). It requires subjects to maintain a fabricated professional identity during interpersonal confrontation, eliciting complex visual and acoustic behavioral markers. With 1,695 samples from four countries and diverse professional identities (Fig. 3), the dataset supports cross-cultural analysis of identity-pretense deception. Its short segments, averaging 3.65s, further encourage models to focus on immediate cues to interpersonal confrontation.

3.3 Stabilized Individuality-Commonality Synergy (SICS)

The goal of SICS (visualized in the bottom-left inset of Fig. 4) is to adapt multimodal features while balancing shared behavioral commonality and sample-specific individuality. In deception detection, global behavioral patterns provide stable cues, whereas each sample may also contain personalized visual-acoustic variations. Therefore, SICS combines a learnable global vector with context-derived residuals to generate sample-adaptive adjustment weights. Positive and negative adjustment branches then enhance useful feature dimensions and suppress noisy or misleading ones, respectively.

Given visual tokens v_i and acoustic tokens a_i, we first perform cross-attention between the two modalities to obtain multimodal fusion tokens $x_i \in \mathbb{R}^{L\times d}$, where

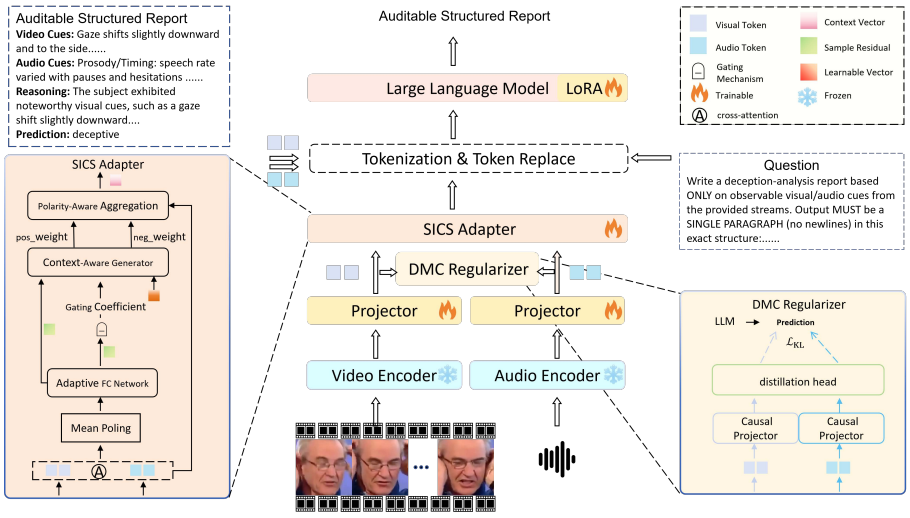


Fig. 4: Overview of our auditable audiovisual deception detection framework. The video and audio encoders extract modality features, which are fused into a robust representation. Within the encoder/fusion stage, SICS stabilizes representations by combining a shared baseline with sample-specific residuals, while DMC mitigates unimodal dominance through agreement regularization on modality-specific prediction distributions. The report generator then produces a single-line schema-constrained audit artifact: `Video Cues; Audio Cues; Reasoning; Prediction`.

L denotes the sequence length. The module then calculates the temporal mean of x_i to obtain a context vector $c_i \in \mathbb{R}^d$, which is processed by the adaptive FC network to generate the sample-adaptive residual $\Delta z_i \in \mathbb{R}^d$:

$$\Delta z_i = \mathbf{W}_2\big(\tanh(\mathbf{W}_1 c_i + \mathbf{b}_1)\big) + \mathbf{b}_2 \tag{1}$$

where $\mathbf{W}_1, \mathbf{W}_2, \mathbf{b}_1, \mathbf{b}_2$ are learnable weights and biases.

Subsequently, a gating coefficient g_i is computed to determine the fusion ratio between a learnable global vector $\mathbf{b}_{global} \in \mathbb{R}^d$ and the generated residual Δz_i: $g_i = \sigma(\mathbf{W}_g \Delta z_i + b_g)$. The Context-Aware Generator then produces positive and negative adjustment weights as follows:

$$w_i = \tanh\big(g_i \cdot \mathbf{b}_{global} + (1-g_i)\cdot \Delta z_i\big), \quad w_i^+ = \mathbf{W}^+ w_i + \mathbf{b}^+, \quad w_i^- = \mathbf{W}^- w_i + \mathbf{b}^- \tag{2}$$

where $\mathbf{W}^+$, $\mathbf{W}^-$, $\mathbf{b}^+$, and $\mathbf{b}^-$ are independent learnable parameters.

During polarity-aware aggregation, the input features are refined as: $x_i' = x_i \odot \mathrm{ReLU}(w_i^+) - x_i \odot \mathrm{ReLU}(w_i^-)$, where $\odot$ denotes element-wise multiplication. Then, the output is calculated as a weighted sum of the refined features and the original input:

$$\mathrm{Output}_i = \lambda \cdot x_i' + (1-\lambda)\cdot x_i \tag{3}$$

where λ is a balancing hyperparameter and we empirically set $\lambda = 0.2$.

3.4 Distilled Modality Consistency (DMC)

To mitigate unimodal shortcut learning, we introduce the DMC regularizer (visualized in the bottom-right inset of Fig. 4), which encourages visual and audio modalities to produce consistent predictions during training.

As shown in Fig. 4, the DMC module consists of modality-specific Causal Projectors followed by a shared distillation head. These components map the frozen visual and audio tokens to the label space $\mathcal{Y} = \{\texttt{deceptive}, \texttt{truthful}\}$. Specifically, the unimodal prediction p_m is generated as:

$$p_m(y) = \mathrm{softmax}\big(\Phi_{\mathrm{distill}}(\mathrm{Proj}_m(h_m))\big)(y), \quad m \in \{v, a\},\ y \in \mathcal{Y}, \tag{4}$$

where h_m represents the modality tokens, Proj_m is the Causal Projector, and Φ_{distill} denotes the distillation head.

The MLLM decoder produces a teacher distribution $q(y)$ at the final `Prediction` position when conditioning on the full multimodal context and the schema constraint. We minimize the KL divergence between the unimodal predictions from the shared distillation head and this MLLM decoder teacher:

$$\mathcal{L}_{\mathrm{distill}} = \sum_{m \in \{v,a\}} \mathrm{KL}\big(q \,\|\, p_m\big). \tag{5}$$

DMC regularizer encourages both modality projection heads to function effectively, which is discarded during inference.

3.5 Auditable Report Generation with Schema Constraint

We generate the auditable report directly with an MLLM. Given video V_i and audio A_i, the MLLM encoder produces modality-specific hidden states and a fused audiovisual representation:

$$H_{v,i},\ H_{a,i},\ H_{va,i} = \mathrm{Enc}_{\mathrm{MLLM}}(V_i, A_i), \tag{6}$$

where $H_{v,i}$ and $H_{a,i}$ capture modality-specific features for visual and acoustic signals, respectively. $H_{va,i}$ captures the cross-modal interactions between them.

Conditioned on these representations, the MLLM decoder then generates a structured, single-line schema-constrained report that provides the essential behavioral evidence required for a comprehensive and auditable deception analysis:

$$R_i = \mathrm{Dec}_{\mathrm{MLLM}}\big(H_{v,i},\ H_{a,i},\ H_{va,i};\ \texttt{schema}\big). \tag{7}$$

The schema enforces a fixed field order and semicolon delimiters: `Video Cues; Audio Cues; Reasoning; Prediction`. We supervise the MLLM to match the target report text, and require the final `Prediction` field to match the ground-truth label (`deceptive`/`truthful`).

3.6 Training Objective

We jointly optimize the schema-constrained report generation task alongside regularizers. The overall objective function is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{rep}} + \alpha \mathcal{L}_{\text{distill}}, \tag{8}$$

where $\mathcal{L}_{\text{rep}}$ denotes the token-level cross-entropy loss for generating the structured audit report. The term $\mathcal{L}_{\text{distill}}$ represents the consistency distillation loss derived from our DMC module. We empirically set $\alpha = 0.1$ to balance the accuracy of the generated reports and the mitigation of unimodality reliance.

4 Experiments

We evaluate our method on six aspects: (i) in-domain effectiveness across diverse deceptive tasks, (ii) cross-domain generalization under dataset shift, (iii) cross-cultural robustness in identity pretense, (iv) component contributions via ablation studies, (v) visualization analysis of SICS adapter and DMC regularizer, (vi) reasoning capability analysis.

4.1 Implementation Details

We use AffectGPT [15] as the base MLLM, which consists of visual/audio encoders, audiovisual projection modules, and an LLM decoder, initialized with the corresponding pre-trained weights. During fine-tuning, the encoders remain frozen; we apply LoRA to the LLM while fully training the projectors and our proposed components. Optimization is performed for 200 epochs using AdamW [19] ($LR = 5 \times 10^{-5}$). Training is conducted on a single NVIDIA H100 (80GB) with a batch size of 4, sampling 8 frames per video at 224×224 resolution.

4.2 Datasets and Protocols

To verify the effectiveness of our method, we evaluate our model on three established benchmarks and our newly introduced dataset, T4-Deception:

- **Bag-of-Lies (BoL)** [8]: 325 lab-collected samples on false image-narration.
- **MU3D** [18]: 320 lab-collected samples on false social-evaluation.
- **DOLOs** [6]: 1,675 game show samples on false story-telling.
- **T4-Deception (Ours)**: 1,695 samples across four cultural contexts: U.S. (876), Germany (702), Vietnam (66), and Bulgaria (51), collected under a unified false professional-identity task.

For the established benchmarks, we conduct in-domain evaluations following official protocols (3-fold for BoL and DOLOs; 4-fold for MU3D), alongside cross-domain assessments. We also perform in-cultural evaluations via 3-fold cross-validation and pairwise cross-cultural tests.

4.3 MLLM Configuration

We compare our method against two types of MLLMs:
Commercial MLLMs (Zero-shot). Commercial models, including GPT-4o [12] and so on, are evaluated via API calls in a zero-shot setting due to limited parameter access. These represent the reasoning capabilities of general-purpose models without task-specific training.
Open-source MLLMs (Fine-tuned). We fine-tune representative open-source MLLMs, including Qwen3-Omni [35], VideoLLaMA2 [3] and so on. These models are trained using LoRA on the same training sets as our method.
Unified Prompting and Output Schema. All models use the same schema-constrained prompt format. Models output a single-line structured record: `Video Cues; Audio Cues; Reasoning; Prediction`.

4.4 In-Domain Evaluation

Table 2: Comprehensive In-domain Benchmark. We evaluate performance across three datasets: DOLOs, Bag-of-Lies (BoL), and MU3D. **Set.**: ZS=Zero-shot, LoRA=Fine-tuning, Full=Full Training. **Mod.**: A=Audio, V=Video, AV=Audio-Visual.

				DOLOs		BoL		MU3D	
Method/Model	**Source**	**Set.**	**Mod.**	**Acc.**	**F1**	**Acc.**	**F1**	**Acc.**	**F1**
Task-specific Deep Learning Methods									
LieNet [14]	TCDS'22	Full	AV	56.50	69.72	59.78	58.14	53.48	33.62
FacialCueNet [22]	AI'23	Full	V	60.98	68.65	56.23	63.26	57.64	59.13
PECL [6]	ICCV'23	Full	AV	64.75	71.20	59.51	51.06	55.31	60.07
AFFAKT [13]	AAAI'25	Full	AV	68.10	70.73	–	–	–	–
Large Multimodal Models (LMMs)									
GPT-4o [12]	Comm.	ZS	V	66.38	64.21	57.14	56.35	54.12	52.47
Qwen-Omni [35]	Comm.	ZS	AV	51.72	49.56	45.27	35.82	50.23	41.38
VideoLLaMA2 [3]	Open	LoRA	AV	53.48	56.12	47.65	49.34	51.84	54.62
SALMONN2 [29]	Open	LoRA	AV	52.63	55.44	46.82	48.51	51.06	53.79
Qwen2.5-7B-VL [35]	Open	LoRA	AV	46.10	52.37	44.24	41.15	53.86	56.24
DecepGPT	**ours**	**LoRA**	**AV**	**73.23**	**76.13**	**63.46**	**63.72**	**61.25**	**67.22**

As shown in Table 2, our method achieves state-of-the-art performance across all three benchmarks. First, compared to zero-shot commercial giants (e.g., GPT-4o [12]), our method yields consistent accuracy gains of +6.85% on DOLOs, +6.32% on BoL, and +7.13% on MU3D. Second, compared to fine-tuned open-source MLLMs (e.g., VideoLLaMA2 [3]), our method yields consistent accuracy gains of +19.75% on DOLOs, +15.81% on BoL, and +7.39% on MU3D. Finally, our approach even surpasses specialized task-specific models (e.g., PECL [6], AFFAKT [13]) by margins of +5.13% (DOLOs), +3.68% (BoL), and +3.61% (MU3D). These results are achieved via efficient LoRA fine-tuning while generating structured audit reports, offering both higher accuracy and better interpretability than baselines that output simple labels. Additional results on

Table 3: Linguistic evaluation of generated reasoning reports. Higher scores indicate better quality.

Method	METEOR	ROUGE-L	BERTScore	GPT-Score
GPT-4o (Zero-shot)	0.284	0.312	0.765	5.42
VideoLLaMA2 (LoRA)	0.215	0.246	0.792	5.15
Qwen2.5-7B-VL (LoRA)	0.242	0.285	0.804	5.28
DecepGPT	**0.342**	**0.384**	**0.812**	**6.96**

Table 4: Cross-Domain Evaluation on DOLOs (D), Bag-of-Lies (B), and MU3D (M). "X&Y→Z" denotes training on X and Y, testing on Z.

Method	M&B → D		D&M → B		D&B → M		Average	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
LieNet [14]	54.40	68.23	54.69	50.51	51.08	59.75	53.39	59.50
PECL [6]	54.51	**69.55**	51.25	66.95	55.38	52.46	53.71	62.99
Ours	**63.46**	61.79	**59.62**	**72.00**	**57.24**	**60.38**	**60.11**	**64.72**

in-the-wild dataset are provided in Appendix 2. Beyond classification accuracy, we rigorously evaluate the quality of the generated auditable reports in Table 3. DecepGPT achieves superior performance across all metrics compared to general-purpose MLLMs. The high BERTScore (0.812) and GPT-Score (6.96) indicate that our schema-driven approach effectively grounds the reasoning in forensic evidence rather than generating generic hallucinations.

4.5 Cross-Domain Evaluation

We evaluate our method under dataset shift by training on source datasets and testing on target domains without target-domain fine-tuning. Table 4 demonstrates that our method consistently outperforms task-specific methods (e.g., PECL) on all transfer paths in accuracy: +8.95% (M&B → D), +4.93% (D&M → B), and +1.86% (D&B → M), with a +6.40% average gain. Regarding F1 score, our method also achieves improvements: +5.05% (D&M → B) and +0.63% (D&B → M), with a +1.73% average gain.

4.6 Evaluation on T4-Deception

We conduct a comprehensive evaluation on the T4-Deception dataset, including both **in-cultural** and **cross-cultural** assessments across four distinct regions: U.S., Germany, Vietnam, and Bulgaria. As shown in Table 5, we train models on each source culture and evaluate them on all four targets (including the source itself) in a zero-shot manner. In-cultural Performance: When trained and tested on the same culture (diagonal entries), our method achieves stable baseline performance, with an average accuracy of 62.58% (e.g., 64.65% for U.S. and 61.11%

Table 5: Comprehensive Evaluation on T4-Deception. We report Accuracy (Acc) and F1 Score for both in-cultural (diagonal, **bold**) and cross-cultural (off-diagonal) settings. Models are trained on the row culture and tested on the column culture.

Train ↓ / Test →	U.S.		Germany		Vietnam		Bulgaria	
	Acc	**F1**	**Acc**	**F1**	**Acc**	**F1**	**Acc**	**F1**
U.S.	**64.65**	**73.28**	59.50	52.14	57.58	61.20	63.00	71.62
Germany	57.00	57.26	**61.11**	**53.33**	56.06	59.14	60.78	62.45
Vietnam	55.48	61.05	54.87	48.92	**63.10**	**68.42**	58.82	64.10
Bulgaria	58.19	63.38	56.45	50.18	54.55	58.76	**61.46**	**66.25**

for Germany). This confirms the model's ability to capture culture-specific deception markers effectively. More importantly, the model demonstrates strong generalization capabilities when transferred to unseen cultures (off-diagonal entries). For instance, the model trained on the U.S. region retains 59.50% accuracy when tested on Germany, showing only a marginal drop compared to the in-domain setting. Across all 12 cross-cultural transfer pairs, the average performance remains stable at 57.69%, with an average relative degradation of only 4.89% compared to in-cultural results. Further comparative analysis of cross-cultural deception is provided in Appendix 1.

4.7 Ablation Study

Main Ablation Study. We evaluate the contribution of each core component across **all three datasets** in Table 6. Both SICS and DMC consistently improve over the baseline. On DOLOs, SICS provides a significant accuracy gain of +5.44%, while DMC adds a further +1.53%. Similarly, on Bag-of-Lies, SICS boosts accuracy by +4.81%, with DMC contributing an additional +1.93%. For MU3D, we observe a substantial improvement of +6.40% from SICS and +1.25% from DMC. The full model, combining both modules, achieves the best performance across all benchmarks, yielding an average accuracy gain of +6.42% over the baseline. These results confirm that SICS and DMC address complementary aspects of robust multimodal learning. The cross-domain transfer performance of these ablated variants is provided in Appendix 3.

Component Analysis and Backbone Comparison. To further validate the effectiveness of the internal mechanisms, we conduct detailed ablation studies on the SICS adapter components and compare different backbone networks. Due to space limitations, we report the results on the **DOLOs** dataset here as a representative example; the complete results for Bag-of-Lies and MU3D are detailed in Appendix 4. As shown in Table 7, removing any component of the SICS adapter (global prior, polarity-aware adjustment, or gating mechanism) leads to a clear drop in performance, confirming their necessity. Furthermore, we investigate the impact of the backbone architecture (Table 8). Replacing the LLM backbone with simpler structures (MLP or Transformer) results in

Table 6: Main ablation study across three datasets. Base refers to the model without the SICS adapter and DMC regularization.

	DOLOs		Bag-of-Lies		MU3D	
Variant	**Acc.**	**F1**	**Acc.**	**F1**	**Acc.**	**F1**
Base (no SICS/DMC)	67.24	71.18	57.69	54.26	53.75	65.34
Base + DMC	68.77	71.94	59.62	58.15	55.00	66.27
Base + SICS	72.68	76.02	62.50	61.43	60.15	67.08
Full (SICS + DMC)	**73.23**	**76.13**	**63.46**	**63.72**	**61.25**	**67.22**

Table 7: Ablation on SICS components.

Variant	Base	+Glo.	+Pol.	+Gat.	+Glo.&Pol.	+Glo.&Gat.	+Pol.&Gat.	**Full**
Acc. (%)	67.24	68.45	68.81	68.62	71.85	72.06	71.42	**73.23**
F1 Score	71.18	72.03	72.45	72.16	74.94	75.28	74.31	**76.13**

significantly lower performance. This suggests that the LLM backbone provides superior semantic reasoning capabilities essential for this task.

5 Visualization and Interpretability Analysis

We analyze the internal mechanisms of the SICS Adapter and DMC Regularizer from two aspects: (i) unimodal shortcut mitigation in the DMC Regularizer and (ii) representation stabilization in the SICS Adapter. We also assess the fidelity of the generated auditable evidence, verifying that it is grounded in the extracted behavioral cues to provide a transparent audit trail for the final prediction.

5.1 Visualization Analysis of the DMC Regularizer

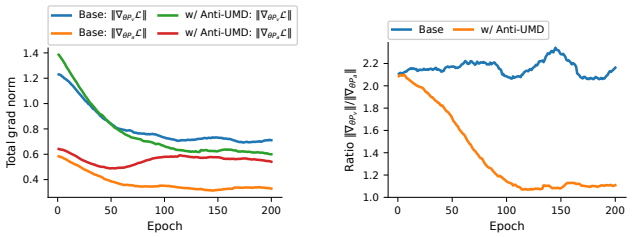
Fig. 5 illustrates how the DMC Regularizer affects cross-modal feature utilization. Fig. 5(a) tracks the gradient norms backpropagated to the modality-specific projectors. In the base setting, the visual projector receives larger gradients, indicating a bias toward visual cues. With DMC, the acoustic-side gradients are gradually strengthened, suggesting that the model incorporates more audio evidence during training. Fig. 5(b) shows the dynamics via the ratio $r = \|\nabla_{\theta P_v}\|/\|\nabla_{\theta P_a}\|$. The base model shows visual dominance ($r > 1$), while the DMC Regularizer reduces this ratio closer to unity. These results indicate that the DMC Regularizer helps balance gradient flow across modalities, supporting more balanced multimodal learning.

5.2 Visualization Analysis of the SICS Adapter

We visualize the adaptive gating distribution, the denoising effect, and the polarity-aware weights to illustrate the SICS Adapter as shown in Fig. 6.

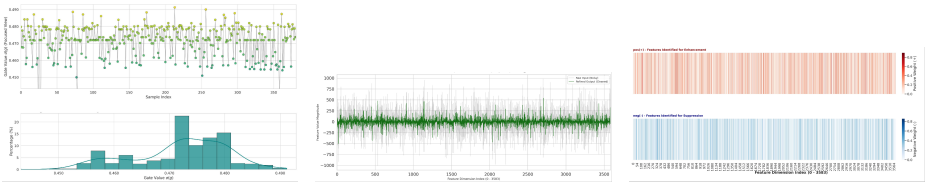
Table 8: Backbone comparison.

Backbone	Acc.	F1
MLP [31]	58.42	60.17
Trans. [32]	64.75	68.34
Ours	**73.23**	**76.13**



(a) Gradient norm. **(b)** Gradient ratio.

Fig. 5: Projector gradient dynamics analysis during training.



(a) Gating distribution $\sigma(g)$. **(b)** SICS feature stabilization. **(c)** Polarity-aware weights.

Fig. 6: Visualization analysis of the SICS Adapter. **(a)** Dynamic variation of gating values on the fusion of global priors and local residuals. **(b)** Comparison between volatile raw features and stabilized features processed by SICS. **(c)** Adjustment where positive weights enhance informative feature and negative weights suppress noise features.

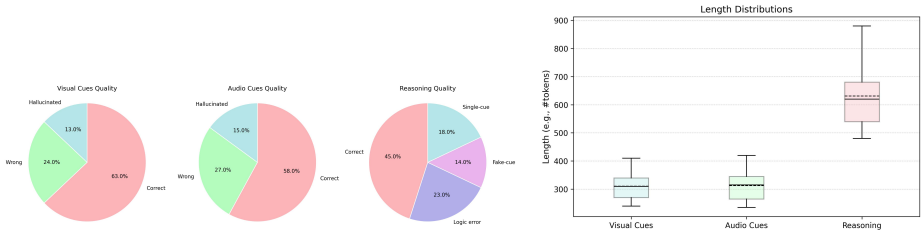
Dynamic gating behavior. Fig. 6a shows the gating coefficient $\sigma(g)$ varies across samples (mostly within $[0.443, 0.493]$). This indicates that the fusion is not a static offset but adapts to each sample. This supports the individuality–commonality synergy: the adapter maintains stable features for typical cases while allowing flexibility to model persona-specific residuals for outliers.

Feature stabilization and noise suppression. In small-data regimes, audiovisual learning can be affected by high-magnitude persona-driven noise (e.g., individual behaviors or recording conditions). Fig. 6b compares feature magnitudes before and after SICS Adapter: raw features show spiky, volatile dimensions, while refined features are smoother and more centered. This suggests that the SICS Adapter reduces such variations, improving numerical stability.

Polarity-aware enhancement vs. suppression. Fig. 6c visualizes the learned positive (enhancement) and negative (suppression) weights. The adjustment from polarity-aware may help produce more discriminative features for the MLLM.

5.3 Validity Analysis of Auditable Reasoning

We use structured, schema-constrained reports (`Video Cues`; `Audio Cues`; `Reasoning`; `Prediction`) as audit artifacts to identify hallucinations and shortcut rationales. We audit these reports along two axes with an error taxonomy: (i) Visual/Acoustic Cues, categorized as Correct, Counterfactual, or Non-existent (hallucinated); (ii) Reasoning Quality, classified as Correct, False-cue (logical but counterfactual), Incoherent, or Single-cue (modal collapse). Since textual



(a) Audit taxonomy distribution. **(b)** Response length distributions.

Fig. 7: Quantitative audit analysis. (a) distribution of cue correctness and reasoning quality categories; (b) text length distribution of reasoning.

GT descriptions cannot exhaustively cover all valid audio-visual observations, case-by-case expert analysis against the original videos is necessary to avoid penalizing correct but unannotated cues. We randomly sample 100 model outputs for human auditing and further adopt a cascaded taxonomy, where each output is assigned to exactly one category within each axis: cue existence is verified before attribute-level errors for audio-visual cues, and reasoning chains are checked sequentially for multimodal cue usage, logical coherence, and factual consistency.

Quantitative distributions and answer length statistics (Fig. 7) demonstrate that our framework enhances behavioral understanding. To highlight how the generated schema-constrained reasoning chains reveal the evidentiary basis for each prediction, we also perform qualitative analysis in Appendix 5.

6 Conclusion

We present DecepGPT, an auditable MDD framework targeting evidence-based and generalizable deception detection in high-stakes applications. Through structured cue-reasoning supervision and T4-Deception, DecepGPT supports human-checkable reports and cross-cultural evaluation; through proposed SICS and DMC, it improves robust multimodal learning under small-data conditions. Quantitative experiments and qualitative audits demonstrate that our framework improves detection performance while making final judgments easier to verify.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (Grant No. 62576076 and No. 62441619), Ningbo Science and Technology Innovation 2025 Major Project (2025Z027), Guangdong Basic and Applied Basic Research Foundation (Grant No. 2023A1515140037), CCF-Tencent Rhino-Bird Open Research Fund, Guangdong Research Team for Communication and Sensing Integrated with Intelligent Computing (Project No. 2024KCXTD047), and SongShan Lake HPC Center (SSL-HPC) in Great Bay University.

References

1. Arevalo, J., Solorio, T., Montes-y Gómez, M., González, F.A.: Gated multimodal units for information fusion. arXiv preprint arXiv:1702.01992 pp. 1 – 17 (2017)
2. Bond, C.F.J., DePaulo, B.: Accuracy of deception judgments. Personality and Social Psychology Review **10**, 214 – 234 (2006)
3. Cheng, Z., Leng, S., Zhang, H., Xin, Y., Li, X., Chen, G., Zhu, Y., Zhang, W., Luo, Z., Zhao, D., Bing, L.: Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. ArXiv **abs/2406.07476**, 1–29 (2024)
4. Ding, M., Zhao, A., Lu, Z., Xiang, T., Wen, J.R.: Face-focused cross-stream network for deception detection in videos. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7794–7803 (2019)
5. Geirhos, R., Jacobsen, J.H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A.: Shortcut learning in deep neural networks. Nature Machine Intelligence **2**, 665 – 673 (2020)
6. 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. 22078–22088 (2023)
7. Guo, X., Yu, Z., Selvaraj, N.M., Shen, B., Kong, A.W.K., Kot, A.C.: Benchmarking cross-domain audio-visual deception detection. IEEE Transactions on Affective Computing pp. 1–18 (2026)
8. Gupta, V., Agarwal, M., Arora, M., Chakraborty, T., Singh, R., Vatsa, M.: Bag-of-lies: A multimodal dataset for deception detection. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 83–90 (2019)
9. Hazarika, D., Zimmermann, R., Poria, S.: Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In: Proceedings of the 28th ACM international conference on multimedia. pp. 1122–1131 (2020)
10. 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)
11. Huang, C., Wei, Y., Yang, Z., Hu, D.: Adaptive unimodal regulation for balanced multimodal information acquisition. In: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 25854–25863 (2025)
12. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 pp. 1–33 (2024)
13. Ji, Z., Tian, X., Liu, Y.: Affakt: A hierarchical optimal transport based method for affective facial knowledge transfer in video deception detection. In: AAAI Conference on Artificial Intelligence. pp. 1336–1344 (2025)
14. 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 (2022)
15. Lian, Z., Sun, H., Sun, L., Yi, J., Liu, B., Tao, J.: Affectgpt: Dataset and framework for explainable multimodal emotion recognition. arXiv preprint arXiv:2407.07653 pp. 1 – 11 (2024)
16. 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. Mach. Intell. Res. **23**, 484–500 (2026)

17. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: Advances in Neural Information Processing Systems. pp. 34892–34916 (2023)
18. 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**, 429 – 439 (2019)
19. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations. pp. 1–10 (2019)
20. Maaz, M., Rasheed, H., Khan, S., Khan, F.S.: Video-chatgpt: Towards detailed video understanding via large vision and language models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. pp. 12585–12602 (2024)
21. Miah, M.M.M., Anika, A., Shi, X., Huang, R.: Hidden in plain sight: Evaluation of the deception detection capabilities of LLMs in multimodal settings. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 31013–31034 (2025)
22. 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. Applied Intelligence **53**, 27413–27427 (2023)
23. Neverova, N., Wolf, C., Taylor, G., Nebout, F.: Moddrop: Adaptive multi-modal gesture recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **38**(8), 1692–1706 (2016)
24. 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)
25. 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. p. 59–66 (2015)
26. Pérez-Rosas, V., Abouelenien, M., Mihalcea, R., Xiao, Y., Linton, C., Burzo, M.: Verbal and nonverbal clues for real-life deception detection. In: Proceedings of the 2015 conference on empirical methods in natural language processing. pp. 2336–2346 (2015)
27. Pérez-Rosas, V., Mihalcea, R.: Cross-cultural deception detection. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 440–445 (2014)
28. Soldner, F., Pérez-Rosas, V., Mihalcea, R.: Box of lies: Multimodal deception detection in dialogues. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 1768–1777 (2019)
29. Tang, C., Li, Y., Yang, Y., Zhuang, J., Sun, G., Li, W., Ma, Z., Zhang, C.: video-salmonn 2: Caption-enhanced audio-visual large language models. ArXiv **abs/2506.15220**, 1–18 (2025)
30. Tang, Z., Xiao, Q., Zhou, X., Li, Y., Chen, C., Li, K.: Learning discriminative multi-relation representations for multimodal sentiment analysis. Information Sciences **641**, 119125 (2023)
31. Tolstikhin, I., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., Yung, J., Steiner, A., Keysers, D., Uszkoreit, J., Lucic, M., Dosovitskiy, A.: Mlp-mixer: an all-mlp architecture for vision. In: Proceedings of the 35th International Conference on Neural Information Processing Systems. pp. 24261 – 24272 (2021)
32. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: 31st Conference on Neural Information Processing Systems (NIPS 2017). pp. 1–11 (2017)

33. 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**, 308–330 (2026)
34. Wu, Z., Singh, B., Davis, L.S., Subrahmanian, V.S.: Deception detection in videos. In: AAAI Conference on Artificial Intelligence. pp. 1695 – 1702 (2018)
35. Xu, J., Guo, Z., Hu, H., Chu, Y., Wang, X., He, J., Wang, Y., Shi, X., He, T., Zhu, X., et al.: Qwen3-omni technical report. arXiv preprint arXiv:2509.17765 pp. 1–25 (2025)
36. 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)
37. Zhang, J., Lin, X., Huang, J., Ye, S., Guo, X., Zhu, D., Hu, R., Guo, D., Liang, Y., Yu, Z., Cao, X.: Multimodal deception detection: A survey. Machine Intelligence Research **23**, 284–307 (2026)
38. Zhang, Z., Zhang, A., Li, M., Zhao, H., Karypis, G., Smola, A.: Multimodal chain-of-thought reasoning in language models. arXiv preprint arXiv:2302.00923 pp. 1–25 (2023)
39. 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)
40. Zhuo, Y., Baskaran, V.M., Wang Yee Kiaw, L., 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 (2024)
41. Şen, M.U., Pérez-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 (2022)