sensors

MDPI

Article

Multimodal Behavioral Sensors for Lie Detection: Integrating Visual, Auditory, and Generative Reasoning Cues

Daniel Grabowski [1,*], Kamila Łuczaj [1] and Khalid Saeed [1,2]

[1] Faculty of Computer Science, Białystok University of Technology, 15–351 Białystok , Poland; kamila.luczaj.109503@student.pb.edu.pl (K.Ł.); k.saeed@pb.edu.pl (K.S.)
[2] Department of Computer Science and Electronics, Universidad de La Costa-CUC , Barranquilla 080002, Colombia
[*] Correspondence: daniel.grabowski@pb.edu.pl; Tel.: +48-668-390-793

Highlights

What are the main findings?

- A multimodal deception detection framework combining visual, audio, and language-based reasoning achieved high accuracy on a DOLOS dataset.
- The ViViT-based visual model reached 74.4% accuracy, while HuBERT audio classification showed strong performance on prosodic cues.

What is the implication of the main finding?

- Multimodal fusion enhances robustness and interpretability in behavioral biometrics for deception analysis.
- Language-guided models like GPT-5 prompt-level fusion provide explainable AI outputs, facilitating trust and real-world applicability.

Abstract

Advances in multimodal artificial intelligence enable new sensor-inspired approaches to lie detection by combining behavioral perception with generative reasoning. This study presents a deception detection framework that integrates deep video and audio processing with large language models guided by chain-of-thought (CoT) prompting. We interpret neural architectures such as ViViT (for video) and HuBERT (for speech) as digital behavioral sensors that extract implicit emotional and cognitive cues, including micro-expressions, vocal stress, and timing irregularities. We further incorporate a GPT-5-based prompt-level fusion approach for video–language–emotion alignment and zero-shot inference. This method jointly processes visual frames, textual transcripts, and emotion recognition outputs, enabling the system to generate interpretable deception hypotheses without any task-specific fine-tuning. Facial expressions are treated as high-resolution affective signals captured via visual sensors, while audio encodes prosodic markers of stress. Our experimental setup is based on the DOLOS dataset, which provides high-quality multimodal recordings of deceptive and truthful behavior. We also evaluate a continual learning setup that transfers emotional understanding to deception classification. Results indicate that multimodal fusion and CoT-based reasoning increase classification accuracy and interpretability. The proposed system bridges the gap between raw behavioral data and semantic inference, laying a foundation for AI-driven lie detection with interpretable sensor analogues.

Keywords: lie detection; multimodal AI; deception analysis; behavioral sensors; machine learning; transformers

check for updates

Academic Editors: Anastasios Doulamis and Stefanos Vrochidis

Received: 21 July 2025
Revised: 11 August 2025
Accepted: 18 September 2025
Published: 2 October 2025

Citation: Grabowski, D.; Łuczaj, K.; Saeed, K. Multimodal Behavioral Sensors for Lie Detection: Integrating Visual, Auditory, and Generative Reasoning Cues. *Sensors* **2025**, *25*, 6086. https://doi.org/10.3390/s25196086

1. Introduction

Detecting deception remains a persistent challenge across domains, such as forensic analysis, psychology, and artificial intelligence. Traditional lie detection methods, including polygraph tests and physiological sensors, often suffer from limited reliability due to susceptibility to stress, emotional variability, or intentional countermeasures. As a result, researchers have increasingly turned to behavioral indicators such as facial expressions, voice fluctuations, and linguistic patterns as non-invasive signals of deceptive behavior[1].

Recent advances in multimodal artificial intelligence (AI) and large language models (LLMs) have opened new avenues for developing sensor-inspired lie detection systems. Vision transformers, such as ViViT, and audio-based models, like HuBERT, act as high-level behavioral sensors, extracting implicit emotional cues from facial expressions, gestures, and prosodic features. These digital sensory modules can detect micro-expressions, vocal stress, or timing anomalies, traits often associated with deceptive intent.

In parallel, the emergence of multimodal generative models such as GPT-5 prompt-level fusion and GPT-4 has enabled the interpretation of complex behavioral patterns through natural language reasoning. Specifically, chain-of-thought (CoT) prompting allows these models to articulate step-by-step judgments about deceptive behavior, increasing system transparency and interpretability. This fusion of perceptual input and semantic inference reflects a shift toward explainable AI in behavioral analysis.

In this study, we propose a multimodal deception detection framework that integrates behavioral signals from visual and auditory modalities with CoT-enabled language reasoning. Our system leverages pretrained models to simulate perceptual sensing, while using prompting strategies to generate interpretable justifications for predictions. We validate our approach on the DOLOS dataset, which contains high-quality video segments with labeled deceptive and truthful speech. Additionally, we explore a continual learning setup, where emotion recognition serves as a pretext task for deception classification, testing the hypothesis that emotional cues can enhance model transferability.

The goal of this research is not only to improve deception detection performance but also to demonstrate how AI models can emulate sensor functions like observing, analyzing, and explaining, within a unified framework. By combining signal-level feature extraction with language-based reasoning, our approach represents a novel direction in intelligent lie detection systems.

2. State-of-the-Art

2.1. Traditional Approaches to Lie Detection

The detection of deception has long been a subject of interest in psychology and forensic science. The most well-known method, the polygraph test, is based on measuring physiological responses such as heart rate, respiration, and electrodermal activity while an individual responds to questions. One of the most widely used protocols is the comparison question test (CQT), which compares physiological reactions to control and relevant questions to infer deception. However, polygraph tests have been criticized for high false positives, susceptibility to countermeasures, and lack of a strong theoretical foundation.

To overcome these limitations, alternative techniques such as the guilty knowledge test (GKT) and cognitive load theory have been explored. GKT assumes that deceivers exhibit differential responses when exposed to crime-relevant stimuli. Cognitive load-based techniques measure the effort required to maintain a lie, often through linguistic or behavioral analysis.

2.2. Micro-Expression Analysis in Deception Detection

Micro-expressions, brief and involuntary facial expressions, have been extensively studied as indicators of deception. Ekman et al. [2] proposed the facial action coding system (FACS) [3], which identifies subtle facial muscle movements linked to emotional suppression. Research has shown that liars often attempt to control their facial expressions, but micro-expressions escape conscious suppression, making them useful for deception detection [4–6].

Recent approaches leverage computer vision and deep learning to automatically detect micro-expressions. The hierarchical transformer network (HTNet) [7] has been proposed to enhance micro-expression recognition by segmenting the face into localized regions (e.g., the eyes and lips) and applying attention-based feature extraction. Other methods, such as local binary patterns (LBP) and optical flow analysis, have also been used to identify subtle facial movements indicative of deception.

Despite advancements, existing micro-expression datasets are often small and lack diversity, making generalization to real-world deception detection challenging.

2.3. Multimodal Deception Detection Systems

Multimodal approaches integrate facial expressions, voice analysis, and physiological signals to improve lie detection performance. Previous systems have combined audio–visual embeddings using feature fusion or attention mechanisms. Models such as OpenFace (facial behavior) and eGeMAPS (prosodic features), as well as biometric sensors, have shown improvements in specific contexts; but, they suffer from limited interpretability and a high reliance on handcrafted features or expensive sensor setups.

Recent studies have introduced deep learning frameworks that combine speech (e.g., HuBERT) and video (e.g., ViViT or VideoMAE) to extract affective and behavioral embeddings. These models act as sensor analogues, capturing subtle deception cues without invasive instrumentation. However, their outputs often lack semantic interpretability, raising concerns about their use in sensitive or forensic scenarios[8].

2.4. CoT-Based Generative Analysis of Deception

Recent developments in multimodal large language models (MLLMs) have opened up new possibilities for interpreting human behavior in a more descriptive and transparent way. Instead of relying solely on discrete emotion labels or classification outputs, models like AffectGPT [9] demonstrate that deep learning systems can now generate natural-language explanations of emotional states, grounded in visual, auditory, and contextual data.

This shift from categorical prediction to generative reasoning represents a major step forward in human-centered AI. The AffectGPT framework introduces a large-scale dataset (MER-Caption) with detailed descriptions of affective behavior, and pairs it with a model capable of processing video, audio, and text jointly. These models are not only able to detect emotions such as anxiety, hesitation, or nervousness, which are closely linked to deceptive behavior, but also explain what they see and hear, using language that humans can understand.

In the context of deception detection, this descriptive capacity is especially valuable. Micro-expressions, tone of voice, and timing patterns often play a subtle but critical role in identifying dishonest behavior. Traditional classifiers may detect such patterns, but without interpretability, their outputs remain opaque. Multimodal LLMs, by contrast, can act as behavioral sensors that both perceive and articulate their observations. When guided by chain-of-thought (CoT) prompting, these models reason step by step, linking observed behaviors to likely emotional or cognitive states, and ultimately to truthful or deceptive intent.

This approach aligns with recent work using models like GPT-5 prompt-level fusion, which can perform zero-shot video-language reasoning. When prompted properly, such systems can describe behavioral cues in real time, for example, noting that a speaker is "avoiding eye contact while answering a direct question" and explaining how this might indicate dishonesty. This form of explanatory AI not only enhances trust in the system but also brings it closer to how humans interpret complex social behavior.

As the field progresses, these generative, multimodal models promise to bridge the gap between low-level perception and high-level reasoning, forming a core component of interpretable and reliable lie detection systems.

2.5. Emotion-Driven Cues in Deception Detection

Emotions play an important role in the expression of deceptive behavior. Psychological theories, such as Paul Ekman's concept of emotional leakage, suggest that individuals who lie often experience internal emotional states like guilt, fear, or cognitive dissonance. These emotions may involuntarily manifest through facial expressions, voice, or body movements, even when the speaker attempts to suppress them[10,11].

One prominent concept is that of micro-expressions, which are rapid and involuntary facial expressions lasting less than half a second. These brief expressions have been associated with concealed emotions and are often studied as potential indicators of deception. While detecting them can be difficult for human observers, deep learning models trained for facial emotion recognition may be capable of identifying such subtle cues.

In our experiments, we investigated the relationship between emotion and deception by applying a continual learning approach with the ViViT transformer model. The model was first trained on emotion classification using facial visual data. In the second phase, it was fine-tuned to classify deceptive versus truthful behavior. This approach was based on the hypothesis that emotional representations could provide meaningful priors for detecting deception-related signals, such as facial stress or inconsistencies in expression.

However, this method did not lead to a significant improvement in deception classification accuracy compared to training the model directly on the deception task. One possible reason is the domain shift between typical emotion datasets and the spontaneous behavior found in the DOLOS dataset, which includes more subtle and context-dependent expressions. Additionally, individuals who lie may deliberately attempt to suppress emotional cues, reducing the effectiveness of emotion-based models.

Despite these limitations, emotion remains a promising intermediate representation in multimodal deception detection. Future models may benefit from using emotion embeddings as auxiliary features or from applying multitask learning strategies to jointly train for both emotion recognition and deception detection objectives.

3. System Overview

The proposed system for multimodal deception detection integrates behavioral perception and generative reasoning using three complementary components: a visual encoder, an audio encoder, and a multimodal language model. Each module plays the role of a soft behavioral sensor, extracting and interpreting cues related to emotional state, cognitive load, and potential deception[12].

3.1. Visual Processing with ViViT

We use the Vision Transformer ViViT to process temporal sequences of facial and upper-body expressions. This model captures fine-grained motion patterns, such as micro-expressions, eye movement, and head gestures. These visual signals are essential in identifying subtle indicators of stress, discomfort, or concealment. The extracted visual features are passed downstream to inform the reasoning module. In some configurations, ViViT is pretrained on emotion recognition and then fine-tuned for deception detection using a continual learning setup.

3.2. Multimodal Prompt-Level Fusion in GPT-5

In the proposed framework, multimodal information is integrated using **prompt-level fusion** within a large multimodal model (GPT-5). Unlike classical late fusion approaches that combine classifier logits, or architectures such as GPT-5 prompt-level fusion, which natively encode both image and text modalities, our method embeds all available modality information into a structured prompt that is processed by GPT-5 in a zero-shot manner.

The pipeline consists of three main stages:

1. **Preprocessing of modalities:**
    - *Video frames*: Sixteen equally spaced frames are extracted from each video using OpenCV, preserving spatial, facial, and bodily cues while ensuring coverage across the entire utterance.
    - *Speech transcription*: Audio is extracted from the video and transcribed using the OpenAI Whisper-1 model, providing a verbatim text representation of the spoken content.
    - *Emotion metadata*: Paralinguistic emotional features are obtained from the audio signal using the SpeechBrain `emotion-recognition-wav2vec2-IEMOCAP` model. The predicted categorical label (e.g., *angry*, *neutral*, and *happy*) and its confidence score are recorded.
2. **Prompt construction:** Each modality is converted into a prompt component:
    - The 16 frames are attached to the user message as base64-encoded images.
    - The transcript text is inserted verbatim.
    - The detected emotion and its confidence score are provided in natural language.

    Depending on the ablation configuration, one or more modalities are omitted from the prompt to measure their individual impact.
3. **Reasoning and classification:** The system message instructs GPT-5 to act as a deception detection researcher, evaluating behavioral, verbal, and paralinguistic consistency. The model is required to output only a JSON object with three fields: label (“lie” or “truth”), confidence (0.0–1.0), and reasoning (a short explanation of observed cues). Decoding parameters are fixed (temperature = 0.2, top-p = 1.0) to ensure deterministic output.

This approach allows GPT-5 to jointly reason over multiple modalities without explicit feature concatenation or model fine-tuning. In the ablation study, we systematically evaluate unimodal (Video-only and Transcript-only), bimodal (Video + Transcript, Video + Emotion, and Emotion + Transcript), and trimodal (Video + Transcript + Emotion) configurations to quantify each modality’s contribution.

Our approach differs from prior multimodal lie detection pipelines in two important ways. First, instead of simply stacking pretrained encoders with a shallow fusion layer, we employ prompt-level fusion in GPT-5, where the outputs of each modality encoder (video frames, transcript text, and audio-based emotion) are jointly presented to the model in a unified reasoning prompt. This design allows the large language model to perform cross-modal reasoning implicitly, rather than relying solely on pre-computed feature concatenation.

Second, GPT-5 is a newly released large multimodal model and, to the best of our knowledge, this is among the first peer-reviewed studies to evaluate its performance in deception detection. The model is leveraged in a zero-shot setting, without task-specific fine-tuning, making it directly applicable to new domains and datasets without costly retraining. By integrating affective, visual, and lexical cues in a single reasoning step, the system can generate interpretable justifications alongside classification outputs.

3.3. DOLOS Dataset

The DOLOS dataset serves as the foundational benchmark for all experiments conducted in this study. It consists of high-quality video samples extracted from the British television show *Would I Lie to You?* (WILTY), in which participants engage in a structured deception game format. Each clip is annotated as either *truthful* or *deceptive*, providing a naturalistic and challenging setting for multimodal lie detection.

Unlike many synthetic or laboratory-collected datasets, DOLOS captures spontaneous human behavior in a semi-controlled yet socially dynamic environment. The clips feature a wide range of speakers, accents, and emotional expressions, recorded in studio-quality conditions with frontal camera views and clear audio. This diversity makes the dataset suitable for evaluating generalization across individuals and content types.

Each video sample is accompanied by aligned audio and, in some cases, automatic transcripts, enabling experiments across vision, audio, and language modalities. We selected a subset of DOLOS containing both lies and truths balanced across speakers to ensure fair evaluation. The dataset also provides sufficient temporal depth to analyze fine-grained behavioral cues such as facial micro-expressions, prosodic variations, and gestural inconsistencies. Figure 1 illustrates sample frames from the DOLOS dataset [13].

(a) (b) (c) (d)

Figure 1. Selected frames from the DOLOS dataset showing representative examples of truthful and deceptive responses. Facial cues and micro-expressions are used in downstream classification. (**a**) Truth sample: participant answering truthfully. (**b**) Truth sample: neutral facial expression. (**c**) Lie sample: subtle signs of discomfort. (**d**) Lie sample: raised eyebrows and eye contact aversion.

4. Preliminary Experiments

To evaluate the feasibility of multimodal deception detection using transformer-based architectures, we conducted a series of baseline experiments involving visual, auditory, and classical feature-based models. All evaluations were performed on the DOLOS dataset, which contains high-quality audiovisual recordings annotated for truthful and deceptive behavior.

4.1. Visual Stream: Video Transformers for Deception Classification

In the visual modality, we utilized ViViT-B/16 and Timesformer-B architectures as spatiotemporal feature extractors. Both models were trained using LoRA[14,15] adapters on short video clips extracted from DOLOS.

Two sampling strategies were explored: dense sampling and uniform sampling. The best results were achieved using ViViT-B/16 with uniform sampling and 32-frame clips, reaching an average accuracy of 74.4%. Timesformer-B achieved comparable performance (74.3%) on eight-frame sequences, although required early stopping due to overfitting. Results indicate that visual cues alone provide a solid foundation for deception detection, especially when fine-tuned on domain-specific data. Table 1 presents the performance of ViViT-B/16, Timesformer-B, and ResNet+LSTM models trained on the DOLOS dataset, comparing dense versus uniform sampling along with mean accuracy and key training settings. Figure 2 shows the training and validation accuracy for the HuBERT-based model, and Figure 3 shows the same for the VideoMAE-based model.

Table 1. Performance of deep learning models trained on DOLOS (video modality).

Model	Accuracy (Mean)	Training Details
ViViT-B-1 6 × 2-kinetics + LoRA (32 frames), dense sampling	0.740	30 epochs, Adam (0.00005), no overfitting
ViViT-B-1 6 × 2-kinetics + LoRA (32 frames), uniform sampling	0.744	30 epochs, Adam (0.00005), no overfitting
Timesformer-B-kinetics + LoRA (8 frames), uniform sampling	0.743	30 epochs, Adam (0.00005), early stopping used
ResNet (frozen) + LSTM	≈0.74	30 epochs, Adam StepLR (0.0005), overfitting noted

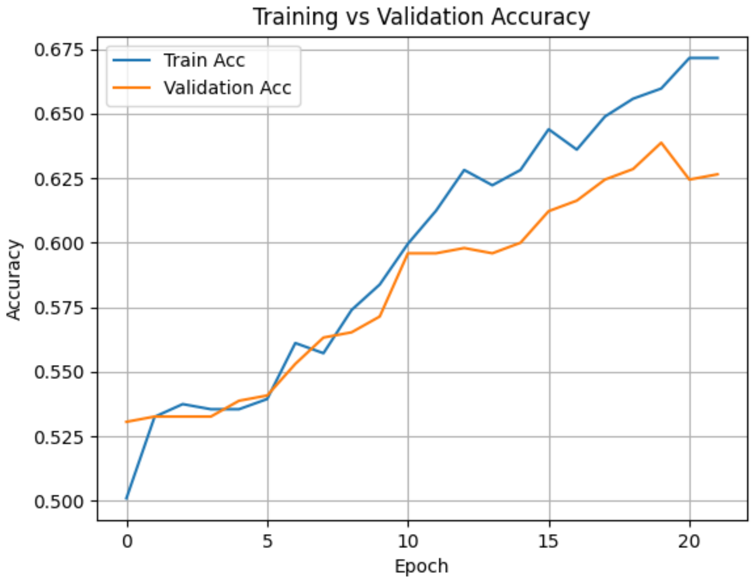

Figure 2. Training vs. validation accuracy for the HuBERT-based audio classification model. Performance improves steadily up to 67.5% with minimal overfitting.

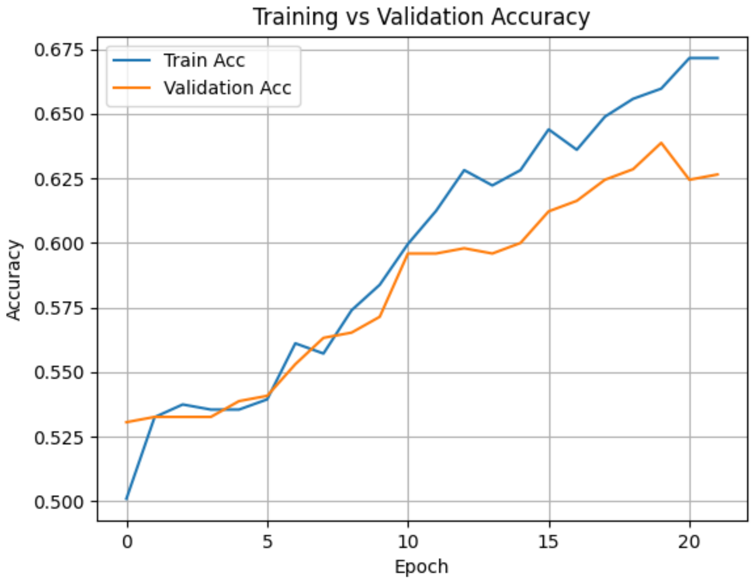

Figure 3. Training vs. validation accuracy for the VideoMAE-based visual classification model.

4.2. Facial Feature Streams: OpenFace and ResNet Backbones

We also experimented with facial feature-based models, using OpenFace landmarks and ResNet embeddings combined with LSTM layers. A frozen ResNet backbone achieved up to 74% accuracy but suffered from significant overfitting, even with regularization techniques such as dropout and weight decay. Fine-tuned ResNet variants did not yield conclusive results[16,17].

4.3. Classical Models on Multimodal Features

For comparison, we trained classical machine learning models (SVM, naive Bayes, random forest, and MLP) on MUMIN-style handcrafted features extracted from DOLOS clips. These models achieved accuracy in the 54–58% range, with MLP reaching the highest reported score (58%).

Although these results are notably lower than deep learning-based approaches, they establish a useful lower bound and highlight the complexity of deception classification from static features. Table 2 presents the performance of classical machine learning models on MUMIN features extracted from the DOLOS dataset, including accuracy, F1-score, and brief comments.

Table 2. Performance of classical machine learning models on MUMIN features (DOLOS dataset).

Model	**Accuracy**	**F1-Score**	**Comments**
Linear SVM	0.54	0.52	-
RBF SVM	0.55	0.54	-
Bernoulli Naive Bayes	0.56	0.55	Slightly better than other models
Random Forest	0.55	0.55	-
MLP	0.58	N/A	Best among classical models

4.4. Ablation Study

To quantify the contribution of each modality, we perform a systematic ablation study using the GPT-5 prompt-level fusion approach described in Section 3.2. In this setting, all predictions are generated in a zero-shot manner by conditioning GPT-5 on the available modalities embedded within a structured prompt. Ablation is implemented by selectively omitting modalities from the prompt while keeping all preprocessing, prompt formatting, and decoding parameters identical.

We evaluate:

- **Unimodal:** Video frames only (V) and transcript only (T).
- **Bimodal:** Video + Transcript (V + T), Video + Emotion (V + E), and Transcript + Emotion (T + E).
- **Trimodal:** Video + Transcript + Emotion (V + T + E), i.e., the full proposed configuration.

For each configuration, we report the precision, recall, F1-score, accuracy, and Matthews correlation coefficient (MCC) for the binary classification task (lie vs. truth). Confusion matrices are computed using the ground truth labels from the DOLOS dataset. All metrics are macro-averaged to account for class imbalance.

Table 3 summarizes the results. The trimodal configuration (V + T + E) achieves the highest F1-score, indicating that all three modalities contribute complementary information. Bold values denote the best score in each column.

Table 3. Ablation study results for different modality combinations using GPT-5 prompt-level fusion.

Configuration	Precision	Recall	F1-Score	Accuracy	MCC
Unimodal					
Video only (V)	0.639	0.883	0.741	0.692	0.415
Transcript only (T)	0.602	0.933	0.732	0.658	0.379
Bimodal					
Video + Transcript (V + T)	0.671	0.917	0.775	0.733	0.502
Video + Emotion (V + E)	0.558	**0.967**	0.707	0.600	0.294
Transcript + Emotion (T + E)	0.598	0.917	0.724	0.650	0.355
Trimodal					
Video + Transcript + Emotion (V + T + E)	**0.683**	0.933	**0.789**	**0.750**	**0.537**

Among the unimodal settings, **video-only** achieves the highest F1-score, suggesting that visual cues alone already encode significant deception-related patterns. Transcript-only performs slightly worse, indicating that verbal content alone may miss nonverbal leakage signals. We do not report an emotion-only configuration, as preliminary trials indicated insufficient stand-alone predictive value. Video cues, while less discriminative in isolation, provide valuable nonverbal signals, such as micro-expressions and gaze aversion. Emotion features derived from audio contribute subtle prosodic patterns that become more informative when combined with other modalities. In particular, adding emotion to the video and transcript improved the F1-score.

In the bimodal setting, the combination of video and emotion (V + E) yields the weakest performance among all tested configurations, with an F1-score of 0.71 and a Matthews correlation coefficient (MCC) of only 0.29. While the recall is exceptionally high (0.97), indicating that the model correctly identifies almost all truthful statements, the precision drops significantly (0.56), leading to a large number of false positives in deception detection. This suggests that emotion cues extracted from audio, when combined with visual frames alone, introduce noise rather than complementary information for the classification task.

Interestingly, when emotion features are incorporated into the trimodal configuration (V + T + E), performance improves notably across all metrics, including a substantial increase in MCC to 0.54 and a balanced F1-score of 0.79. This indicates that emotion features, while insufficient on their own or in the absence of textual context, can provide additional discriminative power when fused with both video and transcript data. Such a finding supports the hypothesis that paralinguistic cues are best interpreted in combination with verbal and visual content rather than in isolation.

Notably, GPT-5 shows a clear tendency to classify true statements correctly (high recall for the truth class) but is less sensitive to lies, a pattern observed across all modality configurations. In the bimodal group, combining visual and textual information (V + T) substantially improves performance over either modality alone, while the inclusion of emotion cues further boosts recall.

These results confirm that GPT-5 is able to leverage multimodal cues effectively when they are presented in a carefully structured prompt. While unimodal setups can provide meaningful predictions, combining modalities yields a consistent improvement in balanced accuracy and F1-score. The observed gains support the hypothesis that visual, verbal, and paralinguistic signals capture complementary aspects of deceptive behavior.

When comparing our zero-shot GPT-5–based system with AffectGPT, we observe competitive performance despite the fact that our approach does not involve any fine-tuning or task-specific training. AffectGPT, which is fully fine-tuned on multimodal sentiment and deception datasets, achieves a mean score of 74.17 across diverse benchmarks, including MER, MELD, IEMOCAP, MOSI, MOSEI, and the SIMS family of datasets. Notably, AffectGPT attains particularly high performance on SIMS v2 (88.99), indicating strong generalization in sentiment-rich conversational contexts. Our trimodal GPT-5 configuration (Video + Transcript + Emotion) attains an accuracy of 75.0% and an F1-score of 0.789 on the DOLOS deception detection benchmark, matching AffectGPT's mean performance while operating entirely in a zero-shot setting. This result is notable given the absence of gradient-based optimization in our pipeline and the reliance on prompt-level fusion, suggesting that modern large language models can leverage multimodal cues effectively without supervised adaptation.

The DOLOS study by Guo et al. reported, on a full dataset using a three-fold protocol, unimodal video performance of 61.44% ACC, an F1-score of 69.42%, audio performance of 59.19% ACC, and an F1-score of 73.46%. Their best fusion approach (PAVF + multitask) achieved 66.84% ACC and an F1-score of 73.35%. In our experiments on a balanced subset of DOLOS, unimodal video reached **74.4% ACC**, unimodal audio achieved ∼**67.5% ACC**, and prompt-level fusion of video, transcripts, and emotion cues using GPT-5 in a zero-shot setting achieved **75.0% ACC with an F1-score of 78.9%**. While the higher performance of our system is partly attributable to the additional transcript modality and a different evaluation split, it is notable that our zero-shot, prompt-based fusion surpasses the best supervised multimodal fusion results reported by Guo et al. These results highlight the potential of large language models to integrate multimodal cues effectively without task-specific fine-tuning.

4.5. Zero-Shot Inference and Testing Procedure

We adopt a **zero-shot inference** strategy, in which GPT-5 is not fine-tuned on the DOLOS dataset. Instead, the model receives modality-specific inputs (video frames, transcript text, and emotion label with confidence) encoded as natural language and images in a single prompt. The prompt template is fixed for all experiments and explicitly instructs the model to output a JSON object containing a predicted label (lie or truth), a confidence score between 0 and 1, and a brief reasoning.

Algorithm 1 outlines the zero-shot, prompt-level multimodal deception detection procedure using GPT-5.

Algorithm 1 Prompt -level multimodal deception detection (zero-shot GPT-5).

```
Require: Dataset $\mathcal{D}$ of videos with ground-truth labels $y \in \{\text{lie}, \text{truth}\}$
Require: Ablation flags: UseVideo, UseTranscript, UseEmotion
Initialize metrics containers
for each sample $x \in \mathcal{D}$ do
    $I \leftarrow \varnothing,\ T \leftarrow \varnothing,\ E \leftarrow \varnothing$
    if UseVideo then
        Extract 16 uniformly spaced frames $I = \{f_0, \ldots, f_{15}\}$
    end if
    if UseTranscript then
        Extract audio; obtain ASR transcript $T$ (Whisper-1)
    end if
    if UseEmotion then
        Extract audio; compute emotion label $e$ and confidence $c_e$ (SpeechBrain wav2vec2)
        $E \leftarrow$ “Detected emotion: $e$ ($c_e$)”
    end if
    Build user prompt: include $E$ (if any), $T$ (if any), instruction to return strictly JSON
    Attach frames $I$ (if any) as images to the same message
    System message: safety + research framing
    Query GPT-5 with deterministic decoding
    Parse first valid JSON object: {label, confidence, reasoning}
    $\hat{y} \leftarrow$ label                                  ▷ final class: lie or truth
    $q \leftarrow$ confidence $\in [0, 1]$           ▷ used only for analysis/threshold sweeps
    Store $(\hat{y}, q)$ and compare with ground truth $y$
end for
Compute Accuracy, Precision/Recall/F1 per class, Macro-F1, MCC, Cohen’s $\kappa$
```

4.6. Computational Environment

All experiments were conducted using high-performance GPU resources (NVIDIA Corporation, Santa Clara, USA). A100 40GB and 80GB cards available on institutional clusters. To accommodate large batch sizes and memory-intensive models such as ViViT[18], we employed DeepSpeed with ZeRO-3 optimization. This setup enabled efficient distributed training and rapid experimentation across multiple modalities. For smaller-scale tasks, RTX-class GPUs were also sufficient, particularly when using lower batch sizes.

4.7. Fusion Experiments and Modality Ablations

We also explored multimodal fusion strategies to combine visual and auditory information for enhanced deception detection. Both early fusion (feature-level concatenation) and late fusion (ensemble of modality-specific classifiers) were considered.

Preliminary findings suggest that each modality provides distinct and complementary cues: the visual stream captures facial expressions and microgestures, while the auditory stream encodes paralinguistic features such as hesitations and intonation changes. Ablation studies confirmed that removing either modality leads to a drop in performance, reinforcing the benefit of multimodal integration.

Experiments involving late fusion of HuBERT[19] and VideoMAE[20] models are currently ongoing. While early observations indicate potential synergy between these modalities, full evaluation on the DOLOS dataset has not yet been completed. These fusion pipelines are expected to improve generalization and robustness in real-world scenarios and constitute a promising direction for future research. Algorithm 2 summarizes the end-to-end multimodal deception detection experiment pipeline, from loading DOLOS data to model training, fusion, and evaluation. Figure 4 shows the block diagram of the multimodal deception detection pipeline: audio (HuBERT), video (VideoMAE), and transcripts are processed separately and then fused for final classification.

Algorithm 2 Multimodal deception detection experiment pipeline.

```
Input: DOLOS dataset with audiovisual clips and labels
Output: Model performance metrics
Load video and audio samples from DOLOS
for each sample in dataset do
    Extract video frames and audio waveform
    Generate transcript via speech-to-text model
end for
Split dataset into training, validation, and test sets
Train ViViT model:
  Preprocess video frames
  Train transformer on frame sequences
  Evaluate on validation set
Train HuBERT model:
  Extract speech embeddings
  Train classifier on HuBERT features
  Evaluate on validation set
Run GPT-5 prompt-level fusion (zero-shot):
for each test sample do
    Create Chain-of-Thought prompt using transcript and visual cues
    Generate textual reasoning and decision
end for
Fusion:
  Combine ViViT and HuBERT predictions (late fusion)
  Evaluate fused model accuracy
return accuracy and interpretability analysis
```

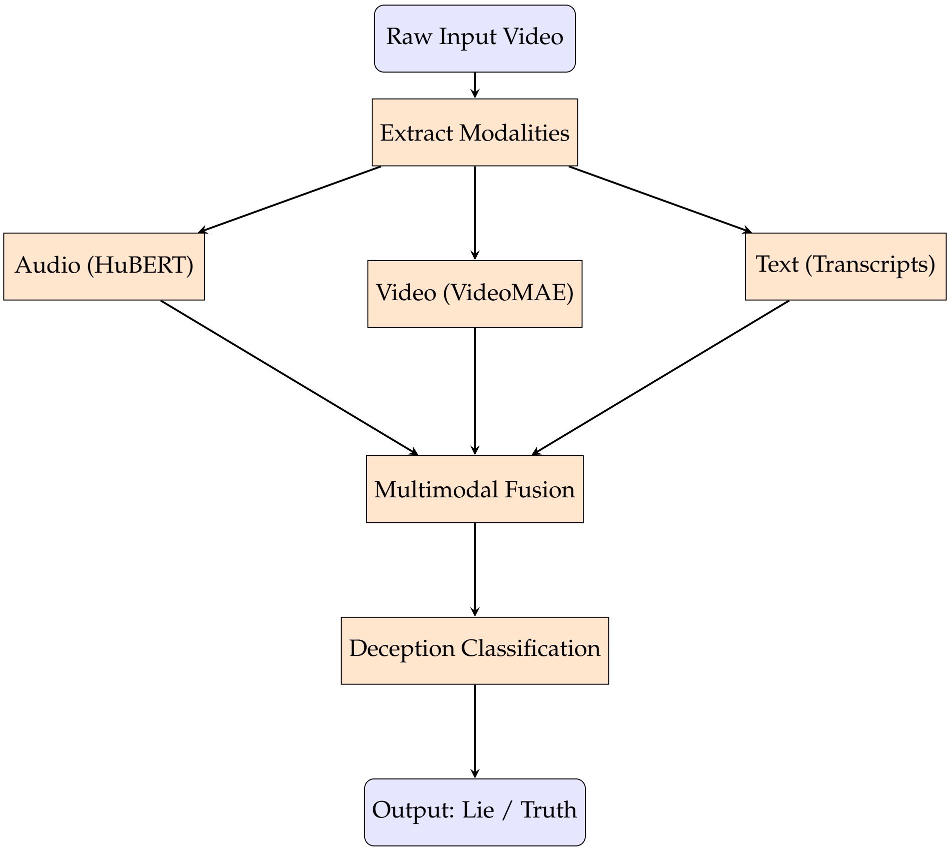

Figure 4. Block diagram of the multimodal deception detection pipeline. Each modality is processed individually before fusion and final classification.

5. Future Work

The preliminary results demonstrate the feasibility of integrating multimodal deep learning with generative reasoning for deception detection. However, several areas remain open for further exploration and improvement.

One of the key directions involves refining the interaction between perception and reasoning. While current results show promise with zero-shot prompting using GPT-5 prompt-level fusion, more structured prompt engineering and fine-tuning on task-specific reasoning paths could enhance consistency and robustness. In particular, chaining outputs from ViViT and HuBERT directly into CoT sequences via intermediate symbolic representations may offer greater interpretability.

Additionally, the integration of physiological and biometric signals represents a natural extension of the system. Signals such as heart rate variability (HRV), galvanic skin response (GSR), and pupillometry, collected via wearable sensors, can serve as additional behavioral modalities, enriching the model's perception of stress and deception. This aligns well with the thematic scope of Sensors, where multimodal sensing systems are central.

6. Conclusions

This study presents a novel multimodal system for deception detection that integrates visual, auditory, and generative reasoning components. By combining transformer-based models such as ViViT and HuBERT with a large multimodal language model (GPT-5 prompt-level fusion (replacing mPLUG-Owl for our final experiments)) guided by chain-of-thought prompting, we aim to bridge the gap between low-level behavioral signal processing and high-level, human-understandable inference.

Our preliminary experiments on the DOLOS dataset demonstrate that each modality contributes complementary information to the deception detection task. The ViViT model, trained with LoRA on 32-frame segments using uniform sampling, achieved a classification accuracy of 74.4%, marking the highest performance among the tested unimodal approaches. Audio-based classification with HuBERT also showed promising results, especially in detecting paralinguistic cues such as pitch variation and speech hesitation. Although early fusion and late fusion experiments are still ongoing, initial indications suggest that combining modalities may improve robustness and overall performance.

Notably, the generative reasoning approach using GPT-5 prompt-level fusion provided not only classification decisions but also transparent, textual justifications for its predictions. This aligns with growing interest in explainable AI and adds interpretive value beyond raw model outputs.

While the current system remains in a research prototype stage, the integration of descriptive behavioral modeling with multimodal sensing paves the way for future systems capable of real-world deployment in forensic, educational, or security contexts. Future work will focus on finalizing multimodal fusion pipelines, improving temporal modeling of behavior over longer sequences, and incorporating additional sensor modalities such as physiological signals to enhance detection accuracy and cross-domain generalization. We also plan to investigate domain adaptation and symbolic reasoning methods to further strengthen robustness and interpretability.

Overall, the findings underscore the potential of combining transformer-based multimodal perception with language-guided reasoning to build transparent and effective deception detection systems. Our approach illustrates how modern AI techniques can be grounded in psychological theory and applied toward complex behavioral understanding tasks.

Author Contributions: Methodology, D.G.; software, K.Ł.; writing—original draft preparation, D.G.; validation, K.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset used in this study is based on the DOLOS corpus, which was constructed from publicly available episodes of the television program *Would I Lie to You?* (BBC). Due to copyright restrictions and the processing involved in dataset construction, the dataset cannot be publicly shared. Researchers interested in replicating or extending this work may contact the authors for further guidance on dataset preparation. No new datasets were generated during this study.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Khare, S.K.; Blanes-Vidal, V.; Nadimi, E.S.; Acharya, U.R. Emotion recognition and artificial intelligence: A systematic review (2014–2023) and research recommendations. *Inf. Fusion* **2024**, *102*, 102019. https://doi.org/10.1016/j.inffus.2023.102019.
2. Ekman, P. *Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage*; W. W. Norton & Company: New York, NY, USA, 2009.
3. Oh, Y.H.; See, J.; Le Ngo, A.C.; Phan, R.C.W.; Baskaran, V. A Survey of Automatic Facial Micro-Expression Analysis: Databases, Methods, and Challenges. *Front. Psychol.* **2018**, *9*, 1128.
4. Rahman, H.A.; Shah, Q.; Qasim, F. Artificial Intelligence based Approach to Analyze the Lie Detection using Skin and Facial Expression: Artificial Intelligence using Skin and Facial Expression. *Pak. J. Emerg. Sci. Technol.* **2023**, *4*, 4.
5. Yan, W.J.; Wu, Q.; Liu, Y.J.; Wang, S.J.; Fu, X. CASME Database: A Dataset of Spontaneous Micro-Expressions Collected from Neutralized Faces. In Proceedings of the 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), Shanghai, China, 22–26 April 2013; pp. 1–7.
6. Chen, H.; Shi, H.; Liu, X.; Li, X.; Zhao, G. Smg: A micro-gesture dataset towards spontaneous body gestures for emotional stress state analysis. *Int. J. Comput. Vis.* **2023**, *131*, 1346–1366.
7. Wang, Z.; Zhang, K.; Luo, W.; Sankaranarayana, R. HTNet for Micro-Expression Recognition. *arXiv* **2023**, arXiv:2307.14637. Accessed on 03 September 2025.

8. Tsai, Y.-H.H.; Bai, S.; Liang, P.P.; Kolter, J.Z.; Morency, L.-P.; Salakhutdinov, R. Multimodal Transformer for Unaligned Multimodal Language Sequences. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL), Florence, Italy, 28 July–2 August 2019; pp. 6558–6569. Available online: https://aclanthology.org/P19-1656/. Accessed on 03 September 2025.
9. Lian, Z.; Chen, H.; Chen, L.; Sun, H.; Sun, L.; Ren, Y.; Cheng, Z.; Liu, B.; Liu, R.; Peng, X.; et al. AffectGPT: A New Dataset, Model, and Benchmark for Emotion Understanding with Multimodal Large Language Models. *arXiv* **2024**, arXiv:2501.16566. Accessed on 03 September 2025.
10. Kalchbrenner, N.; Grefenstette, E.; Blunsom, P. A convolutional neural network for modeling sentences. *arXiv* **2014**, arXiv:1404.2188. Accessed on 03 September 2025.
11. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Distributed representations of words and phrases and their compositionality. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Stateline, NV, USA, 5–10 December 2013; Volume 26, pp. 3111–3119.
12. Khalil, M.A.; Ramirez, M.; Can, J.; George, K. Implementation of Machine Learning in BCI Based Lie Detection. In Proceedings of the 2022 IEEE World AI IoT Congress (AIIoT), Seattle, WA, USA, 6–9 June 2022; pp. 213–217. https://doi.org/10.1109/AIIoT54504.2022.9817162. Accessed on 03 September 2025.
13. Guo, X.; Selvaraj, N.M.; Yu, Z.; Kong, A.; Shen, B.; Kot, A. Audio-Visual Deception Detection: DOLOS Dataset and Parameter-Efficient Crossmodal Learning. In Proceedings of the International Conference on Computer Vision (ICCV), Paris, France, 2–3 October 2023.
14. Ye, D.; Bai, S.; Han, Q.; Lu, H.; Wang, X.E.; Wang, S.; Lu, Y.; Bi, B.; Shou, L.; Zhou, M. GPT-5 prompt-level fusion (replacing mPLUG-Owl for our final experiments): Modularization Empowers Large Language Models with Multimodality. *arXiv* **2023**, arXiv:2304.10592. Available online: https://arxiv.org/abs/2304.10592. Accessed on 03 September 2025.
15. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv* **2021**, arXiv:2106.09685. Available online : https://arxiv.org/abs/2106.09685. Accessed on 03 September 2025.
16. Baltrušaitis, T.; Robinson, P.; Morency, L.-P. OpenFace: An Open Source Facial Behavior Analysis Toolkit. In Proceedings of the 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Placid, NY, USA, 7–10 March 2016
17. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. *arXiv* **2015**, arXiv:1512.03385. Available online: https://arxiv.org/abs/1512.03385. Accessed on 03 September 2025.
18. Arnab, A.; Dehghani, M.; Heigold, G.; Sun, C.; Lučić, M.; Schmid, C. ViViT: A Video Vision Transformer. *arXiv* **2021**, arXiv:2103.15691. Available online: https://arxiv.org/abs/2103.15691. Accessed on 03 September 2025.
19. Hsu, W.-N.; Bolte, B.; Tsai, Y.-H.H.; Lakhotia, K.; Salakhutdinov, R.; Mohamed, A. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *arXiv* **2021**, arXiv:2106.07447. Available online: https://arxiv.org/abs/2106.07447. Accessed on 03 September 2025.
20. Tong, Z.; Song, Y.; Wang, J.; Wang, L. VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training. *arXiv* **2022, arXiv:2203.12602**. Available online: https://arxiv.org/abs/2203.12602. Accessed on 03 September 2025.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.