MSTS Methods in Science and Technology Studies

futuristech PT. Teknologi Futuristik Indonesia

Received: July 23, 2026; Revised: November 5, 2026; Accepted: November 6, 2026; Published: November 7, 2026; Current version: November 7, 2026. e-ISSN: 3123-4232
Digital Object Identifier 10.64539/msts.v2i2.2026.555

Article

Controlled Evaluation of Gated Multimodal Fusion for Lie Detection Using Facial Motion and Speech Cues

Huy Phan Anh[1,*]

[1] Faculty of Information Technology, Saigon University, Ho Chi Minh City 72722, Vietnam;
e-mail: pahuy@sgu.edu.vn (A. H. Phan).
* Correspondence Author

This research received no external funding.

Abstract: Automatic lie detection from video requires combining weak behavioral cues under clip-level supervision. In gated multimodal models, the effects of modality-prior initialization and continuing regularization are difficult to distinguish without matched controls. This study evaluates a three-stream DOLOS pipeline combining static face crops, short-term optical flow, Wav2Vec2-base speech representations, and face-validity masks. Three gated logit-fusion conditions are compared: uniform initialization without regularization, directed initialization without regularization, and directed initialization with Prior-KL. Each condition is evaluated across three predefined splits and three random seeds. The directed-initialized Prior-KL model achieves 64.16 ± 4.56% AUC, 58.96 ± 4.16% balanced accuracy at the validation-selected threshold, 56.74 ± 10.67% lie-class F1, and 57.45 ± 5.15% macro-F1. Relative to the uniform-initialized control, the combined initialization-plus-KL configuration gains 2.26 percentage points in AUC and 2.29 points in macro-F1. With directed initialization held constant, adding KL yields higher lie-class F1 but no improvement in AUC or balanced accuracy. A matched seed-42 ablation finds that static facial appearance alone achieves higher AUC and balanced accuracy than the fused model, while fusion yields higher lie-class F1. These findings associate the observed mean AUC gain with directed initialization and show a class-specific trade-off from regularization. They also motivate identity-disjoint evaluation to assess the role of appearance-related shortcuts.

Keywords: Deception detection; Multimodal learning; Gated fusion; Optical flow; Speech representation; DOLOS.

Copyright: © 2026 by the authors. This is an open-access article under the CC-BY-SA license.

1. Introduction

Automatic deception detection is difficult because deception rarely produces a single observable marker. In video, possible cues appear as subtle facial actions, short-term head or mouth motion, prosodic variation, pauses, or mismatch between speech and facial behavior. Most of these cues are weak, brief, and context dependent. Identity, recording conditions, emotion, cognitive load, and the social setting of the interaction can easily become confounders [1]-[6].

Micro-expressions are often discussed in this context because they are brief, low-intensity facial expressions that may reveal affective leakage under controlled or high-stakes conditions [7]-[10]. Prior recognition studies also show how apex-frame cues and dataset-level constraints shape micro-expression analysis [11], [12]. Here, facial appearance and short-term motion are represented by face crops, optical flow and temporal modeling under clip-level supervision. The pipeline does not include explicit micro-expression detection, apex-frame identification or action-unit extraction.

Recent multimodal deception datasets and benchmarks have shifted the field from hand-coded behavioral cues toward audio-visual learning. DOLOS is useful for this study because it provides game-show video clips labeled as truth or lie and supports controlled evaluation of multimodal models [1], [2]. The original DOLOS work reports gains from parameter-efficient cross-modal learning and auxiliary supervision [1], a broader direction related to adapter-style and low-rank adaptation methods [13], [14]. With only clip-level labels and limited compute, a simpler three-stream system provides a setting for ex-

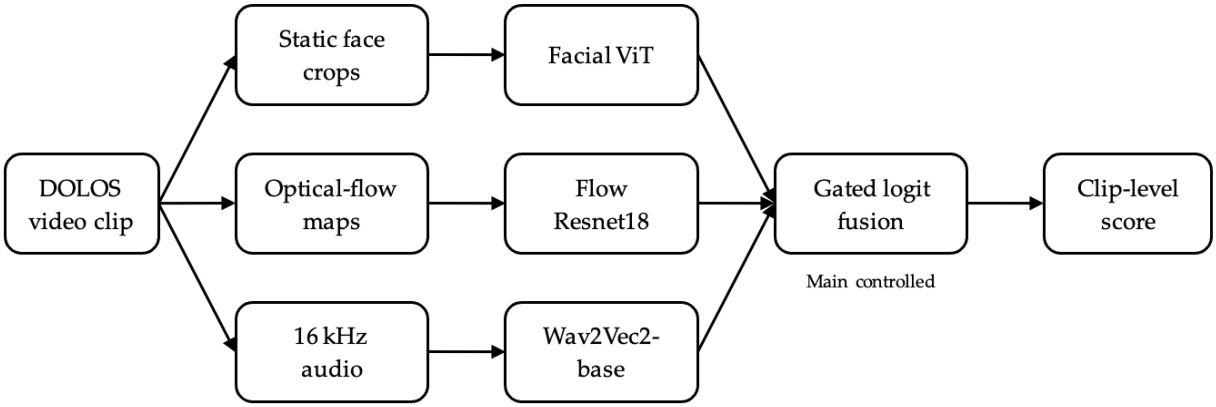


Figure 1. Audio-visual pipeline. Static face, optical-flow, and audio inputs are encoded separately and fused through the gated logit model used in the controlled study.

Table 1. DOLOS three-fold split used in the experiments.

Fold	Split	Clips (% of 1,429)	Truth / Lie
fold1	train	857 (59.97%)	411 / 446
fold1	validation	88 (6.16%)	36 / 52
fold1	test	484 (33.87%)	227 / 257
fold2	train	832 (58.22%)	400 / 432
fold2	validation	115 (8.05%)	45 / 70
fold2	test	482 (33.73%)	229 / 253
fold3	train	835 (58.43%)	403 / 432
fold3	validation	130 (9.10%)	52 / 78
fold3	test	464 (32.47%)	219 / 245

amining the contribution of directly extractable audio-visual cues without fine-grained auxiliary supervision.

We evaluate gated audio-visual fusion through three complementary analyses. First, a DOLOS pipeline combines static face crops, short-term optical flow, Wav2Vec2-base speech representations, and face-validity masks. Second, a three-fold, three-seed comparison separates neutral initialization, directed prior initialization, and Prior-KL regularization. Third, single-stream ablations examine modality-specific performance, with vote-score-sum fusion considered separately as an exploratory analysis.

Section 2 describes the dataset, preprocessing, architecture and training procedure. Section 3 presents the controlled results, diagnostic ablation and limitations. Section 4 concludes the paper. Figure 1 summarizes the evaluated pipeline.

2. Method

2.1. Dataset and evaluation protocol

All experiments use DOLOS (dataset source: https://rose1.ntu.edu.sg/dataset/DOLOS/) [1], [2]. The clips come from the television program Would I Lie To You? and carry clip-level truth or lie labels. All reported configurations use the same three cache-filtered protocol splits. Within each fold, training data are used for parameter estimation, validation data for checkpoint and decision-threshold selection, and test data for performance reporting. The main gated comparison comprises three folds and three random seeds (42, 123, and 2025), giving nine fold-seed runs per condition. Single-stream ablations use seed 42 on all three folds and a seed-matched gated reference. The published DOLOS release comprises 1,675 clips, whereas the protocol files used here contain 1,646 unique clip IDs. Video-integrity checks removed 215 clips, and missing or invalid face/flow caches excluded two more, leaving 1,429 clips per fold. The exclusion procedure was based only on protocol membership, video integrity, timestamp validity, and multimodal-cache availability, and was applied independently of class labels and test predictions. Validation data were sampled from the corresponding training partition.

The video-integrity exclusions comprise 215 unique clips: 200 unavailable videos (87 Truth, 113 Lie) and 15 duration/timestamp failures (7 Truth, 8 Lie). The two cache exclusions are both Lie. Overall, filtering removes 94/768 Truth clips (12.24%) and 123/878 Lie clips (14.01%), leaving 674 Truth and 755 Lie clips. Figure 2 presents stage-wise counts and class-specific removal rates.

The three folds are predefined experimental splits with overlapping data across folds. Within each split, validation sampling uses episode grouping, duplicate IDs are removed, and the training, validation and test sets have no clip-ID overlap. One clip appearing in both the original training and test files of splits 1 and 3 is assigned to test only; the processed test sets of these two splits consequently share that clip. The nine fold-seed evaluations are therefore dependent, and their standard deviation summarizes variation across runs. Table 1 reports sample counts and percentages.

2.2. Preprocessing

Preprocessing produces synchronized audio, face, motion, and quality-mask inputs. Audio is extracted from each video and resampled to a mono 16 kHz waveform. Each clip is segmented into two-second windows.

For clips of duration D greater than two seconds, training samples three independent two-second windows

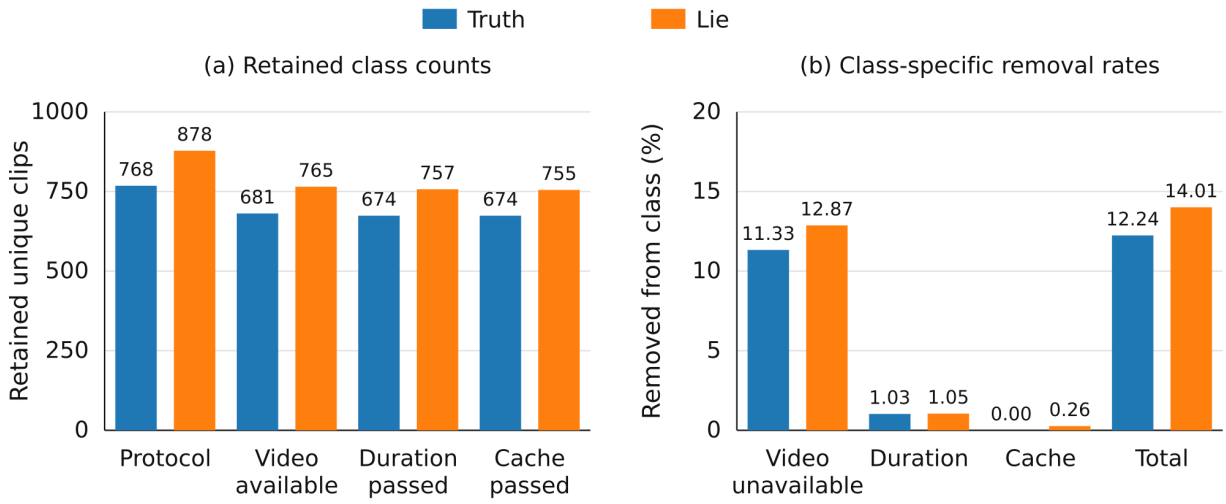


Figure 2. Truth/Lie clips retained after each filtering stage (left) and class-specific removal rates (right). Stage rates use the class counts entering each stage; total rates use the protocol counts of 768 Truth and 878 Lie clips.

Table 1. Input configuration after preprocessing.

Component	Setting	Role
Audio	16 kHz, 2 s window	Speech and prosody
Face crop	224 x 224 RGB	Static facial appearance
Optical flow	2 channels	Short-term motion
face_valid	Frame mask; window ratio r; 0.75 flag is not hard rejection	Visual reliability mask
Windows	Train: 3 random windows; eval: stride 1 s, at most 16	Clip-level aggregation

with start times drawn uniformly from [0, D − 2]; shorter clips contribute one window. Validation and testing use a one-second stride, include the final boundary window, and subsample approximately evenly when there are more than 16 windows. Each window contains 16 visual frames. A clip-level score is the arithmetic mean of its window-level softmax Lie probabilities. Table 2 summarizes the input configuration.

Training-only augmentation comprises RGB random erasing (probability 0.25, shared region across window frames) and independent audio speed perturbation (uniform rate 0.9–1.1, followed by padding/truncation to two seconds). RGB and flow share sampled frame indices; flow is not augmented.

MediaPipe [15] locates faces and produces 224 by 224 face crops. Optical flow is computed on face-aligned frames as cached dense flow [16] and represented as a two-channel motion map. Missing flow is replaced by zeros rather than by discarding the entire sample.

Stored validity values are binarized at 0.5 for each sampled frame. Their mean within window i gives the valid-frame ratio r_i. The loader uses this frame-level mask in preference to the separate $r_i \geq 0.75$ flag, so the flag does not reject windows in the main pipeline. Masking is applied after the visual LSTMs to temporal pooling and to the visual feature means supplied to the gate. The visual gate logits receive an additive $\log(r_i + 10^{-6})$ adjustment, and windows with no valid visual frames use the audio stream. The same ratio weights visual auxiliary losses and the KL reduction. Missing or unreadable validity files are assigned all-valid masks.

The flow-extraction implementation supports TV-L1 with a Farneback fallback. Flow values are clipped to ±20 and stored in float16. The Farneback parameters (pyr_scale, levels, winsize, iterations, poly_n, poly_sigma, flags) are (0.5, 3, 15, 3, 5, 1.2, 0). The MediaPipe Face Mesh implementation defaults to a maximum of four faces, detection and tracking confidence thresholds of 0.5, and a target sampling rate of 25 fps.

In dominant-track mode, IoU association uses a 0.3 threshold and a maximum gap of 12 sampled frames; tracks are ranked by detections after the first second, total detections and mean face area.

2.3. Three-stream architecture

The model uses three modality-specific encoders. The static-face stream uses a facial-expression ViT encoder and projects its token representation to 256 dimensions, following the transformer-based visual-encoding line of work [17], [18]. The motion stream uses a modified ResNet18 [19] whose first convolution accepts two-channel optical flow, followed by a projection to 256 dimensions. The audio stream uses Wav2Vec2-base [20] to extract speech representations and temporal attention pooling to obtain 16 audio tokens. The architecture keeps the usual multimodal division of labor: each encoder captures evidence in its own modality before any fusion step is applied [21]-[23].

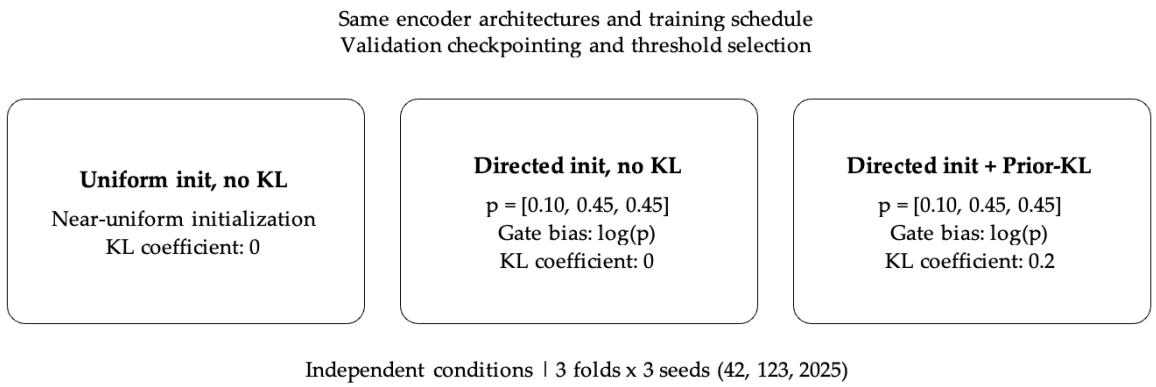


Figure 1. Controlled gate configurations shown as three independent experimental conditions. Encoder architectures and training settings are held fixed; gate initialization and the KL penalty vary between columns.

Table 2. Model streams and their functions.

Stream	Backbone	Output	Function
Static face	Facial ViT: LaurenGurgiolo/vit-micro-facial-expressions (frozen)	256	Appearance and expression
Motion	ResNet18_Weights.DEFAULT; 2-channel initialization (trainable)	256	Mouth, head, facial motion
Audio	facebook/wav2vec2-base (frozen)	256	Prosody and speech rhythm

Table 3 lists pretrained sources and backbone trainability. The flow input convolution copies the channel-averaged RGB kernels into two channels. Projections, temporal/classification heads, the gate and audio pooling remain trainable. Audio pooling uses 16 learned queries, fixed sinusoidal positions and four attention heads, followed by projection to 256 dimensions. Figure 3 summarizes the gate configurations.

Each stream uses LayerNorm, a two-layer bidirectional LSTM (128 hidden units per direction; inter-layer dropout 0.1), temporal mean pooling and a two-class linear head. Visual pooling uses the validity mask. The gate maps concatenated normalized encoder-token means (768 dimensions) through a 128-unit hidden layer, LayerNorm, GELU and dropout 0.1 to three logits. Stream dropout (0.15) retains at least one modality during training.

2.4. Fusion strategies

An exploratory audio-to-visual cross-attention module uses audio queries, visual keys/values and a soft temporal bias [23], [24]. The module is excluded from the gated models evaluated in Tables 5 and 6.

In gated logit fusion, each stream produces class logits and pooled features. A gate network maps the pooled multimodal representation to non-negative weights that sum to one, and the fused window logits are the weighted sum of stream logits. The main Gated Prior-KL condition uses a directed target prior p = [0.10, 0.45, 0.45] for static face, optical flow, and audio. For each sample, the gate module outputs a modality distribution g. The Prior-KL term is computed manually as:

$$D_{KL}(g \parallel p)=\sum_{m} g_m \log\left(\frac{g_m}{p_m}\right) \quad (1)$$

where the divergence is taken from the learned gate distribution to the fixed prior. The total training objective is:

$$L=L_{cls}+L_{aux}+\lambda_{KL}L_{KL} \quad (2)$$

with λ_{KL} fixed to 0.2 for all folds and random seeds. The same directed prior initializes the gate: the final gate weight is set to zero and the final gate bias to log(p), so the raw initial softmax gate matches p before validity adjustment and stream dropout. The lower static-face prior is intended to reduce reliance on identity or production cues.

Three conditions isolate the initialization and regularization effects. The uniform-initialized, unregularized gate uses a neutral initialization [0.333333, 0.333333, 0.333334] and disables the KL term. The directed-initialized, unregularized gate initializes the final gate bias as log([0.10, 0.45, 0.45]) but removes the KL term. The directed-initialized Prior-KL gate uses the same directed initialization and adds 0.2 L_{KL}, where p = [0.10, 0.45, 0.45]. The main analysis compares these gated models without ensembling.

All fused and auxiliary CE terms use the clip's Truth/Lie label for each sampled window. Let B denote the number of windows in a microbatch, r_i their valid-frame ratios and z_i,m the stream logits. Visual auxiliary CE and KL use r_i-weighted means; fused and audio CE use ordinary means. Each auxiliary coefficient is 0.2; gate entropy regularization is disabled. KL uses gate probabilities after validity adjustment and training-time stream dropout. Zero-weight terms evaluate to zero; denominators and logarithm inputs are clamped at 10^{-8}.

$$L_{cls}=\frac{\sum_{i=1}^{B} CE(z_{i,fused}, y_i)}{B} \quad (3)$$

Table 4. Training and evaluation configuration.

Item	Value
Optimizer	AdamW
Learning rate for all trainable components	1e-3
Weight decay	0.01
Scheduler	CosineAnnealingWarmRestarts (T_0=10)
Max epochs	50
Checkpoint	Validation AUC-ROC
Threshold	Balanced accuracy on validation
Main gated protocol	3 folds × 3 seeds (42, 123, 2025)
Single-stream protocol	3 folds × seed 42
Reporting	mean ± standard deviation across runs
Hardware	RTX 3060 12GB
Prior target distribution	p = [0.10, 0.45, 0.45]
Prior-KL coefficient (lambda_KL)	lambda_KL = 0.2; fixed for all folds and seeds
KL regularizer	D_KL(g||p) = sum_m g_m log(g_m/p_m); g = learned gate, p = fixed prior
Gate-prior initialization	final gate weight = 0; final gate bias = log([0.10, 0.45, 0.45])

$$L_m = \frac{\sum_{i=1}^{B} r_i CE(z_{i,m}, y_i)}{max(\sum_{i=1}^{B} r_i, 10^{-8})} \quad (4)$$

$$L_{audio} = \frac{\sum_{i=1}^{B} CE(z_{i,audio}, y_i)}{B} \quad (5)$$

$$L_{aux} = 0.2(L_{static} + L_{flow} + L_{audio}) \quad (6)$$

$$L_{KL} = \frac{\sum_{i=1}^{B} r_i D_{KL}(g_i \| p)}{max(\sum_{i=1}^{B} r_i, 10^{-8})} \quad (7)$$

The auxiliary losses update the trainable projections, temporal and classification heads, and flow encoder. The facial ViT and Wav2Vec2-base backbones remain frozen.

2.5. Training and metrics

Training was implemented in PyTorch [25] and uses AdamW with decoupled weight decay [26], CosineAnnealingWarmRestarts with T_0=10 [27], mixed precision, batch size 4, gradient accumulation 4, a maximum of 50 epochs, and early stopping patience of 10 epochs. Checkpoints are selected by validation AUC-ROC. For each fold-seed run, the decision threshold is selected on the validation split to maximize balanced accuracy and then applied unchanged to the test split. We report AUC-ROC, balanced accuracy at the validation-selected threshold, lie-class F1, and macro-F1 [28]-[30]. Results are reported as mean ± sample standard deviation across nine fold-seed runs for the main gated comparison and across three folds at seed 42 for the diagnostic ablation.

A single AdamW parameter group uses an initial learning rate of 1e -3 for the flow backbone, projections, audio pooling, temporal and classification heads, and gate. Frozen backbones are excluded from optimization. Ties in validation balanced accuracy during threshold selection are resolved by macro-F1. The reported thresholded metrics use validation-selected thresholds without probability calibration. Table 4 summarizes the training and evaluation settings.

Each training batch contains four clips (up to twelve windows); four-batch accumulation normally covers sixteen clips per optimizer step. Losses are reduced within each window microbatch and divided by four before backpropagation. Short clips and partial batches reduce the window count.

Random seeds are applied to Python, NumPy, PyTorch CPU and all CUDA devices, and the DataLoader generator. cuDNN uses benchmark=True and deterministic=False. The experiments used Python 3.10.20, PyTorch 2.5.1+cu121, torchvision 0.20.1+cu121, torchaudio 2.5.1+cu121, Transformers 5.5.3, MediaPipe 0.10.14, OpenCV 4.13.0 and scikit-learn 1.7.2 on an NVIDIA GeForce RTX 3060 with 12 GB memory.

3. Results and Discussion

3.1. Controlled evaluation of gated Prior-KL

Table 5 compares the three gated conditions. Relative to the uniform-initialized, unregularized gate, directed initialization with Prior-KL increases mean AUC by 2.26 percentage points, balanced accuracy at the validation-selected threshold by 1.62 points, lie-class F1 by 4.96 points, and macro-F1 by 2.29 points. This combined configuration changes both initialization and regularization.

Directed initialization alone increases mean AUC by 2.30 percentage points over uniform initialization. With directed initialization held fixed, adding KL changes mean AUC by -0.04 points and balanced accuracy by -0.31 points. At validation-selected thresholds, mean Lie F1 rises by 3.99 points while Truth F1 falls by 3.45 points. This class-specific trade-off does not indicate a uniform classification improvement or establish its mechanism. Figure 4 shows the descriptive comparisons and variation across runs.

Table 5. Main controlled gated experiments on DOLOS (mean ± SD across 3 folds × 3 seeds).

Method	Runs	AUC	BA (val. thr.)	F1 Lie (val. thr.)	Macro-F1 (val. thr.)
Uniform init, no KL	9	61.90 ± 3.31	57.34 ± 2.85	51.78 ± 12.67	55.16 ± 5.10
Directed init, no KL	9	64.20 ± 4.20	59.27 ± 3.62	52.75 ± 11.45	57.18 ± 4.93
Directed init + Prior-KL	9	64.16 ± 4.56	58.96 ± 4.16	56.74 ± 10.67	57.45 ± 5.15

Table 6. Single-stream ablation over the three folds at seed 42; all models, including Gated Prior-KL, use seed 42 only.

Experiment	Runs	AUC	BA (val. thr.)	F1 Lie (val. thr.)	Macro-F1 (val. thr.)
Static face only	3	65.33 ± 1.71	60.91 ± 1.86	54.39 ± 7.38	59.23 ± 3.18
Optical flow only	3	54.46 ± 7.03	53.21 ± 5.85	50.39 ± 9.79	52.49 ± 6.28
Audio only	3	58.03 ± 2.50	56.04 ± 2.38	53.09 ± 6.40	55.22 ± 2.84
Directed init + Prior-KL*	3	64.72 ± 6.55	59.10 ± 4.82	62.76 ± 4.77	58.85 ± 4.88

**All models in this diagnostic use the same three folds and seed 42; the separate main gate comparison in Table 5 retains nine runs.*

Table 7. Exploratory vote-score-sum results across three folds and three seeds.

Method	Runs	AUC	BA (val. thr.)	F1 Lie (val. thr.)	Macro-F1 (val. thr.)
Vote-score-sum (gamma=0.5; floor=0.05)	9	66.06 ± 3.82	61.01 ± 3.08	58.52 ± 10.05	59.87 ± 4.43

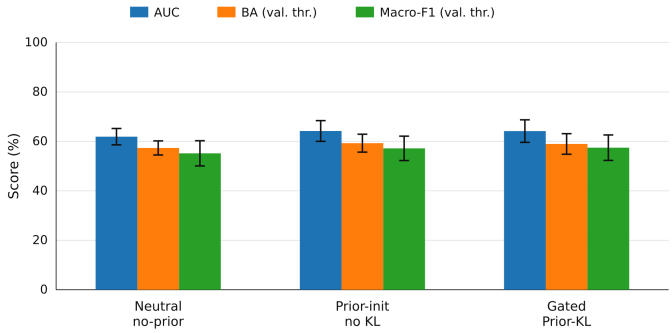


Figure 4. Controlled gated results. Error bars show standard deviation across the nine fold-seed runs.

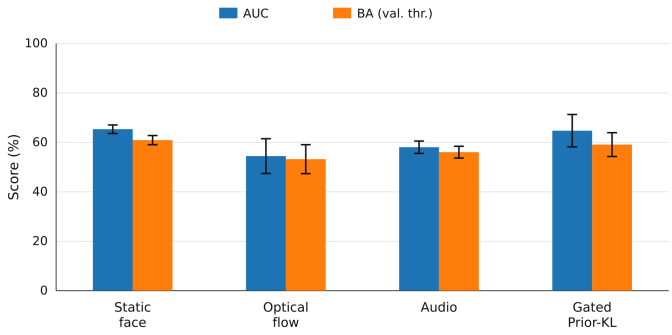


Figure 5. Single-stream results and Gated Prior-KL reference. Error bars show standard deviation; all models use three seed-42 fold evaluations.

3.2. Single-stream ablation and potential shortcut learning

In the seed-42 comparison across three folds (Table 6), static face achieves 65.33 ± 1.71% AUC and 60.91 ± 1.86% balanced accuracy at the validation-selected threshold, compared with 64.72% and 59.10% for Gated Prior-KL. Optical flow and audio have lower mean performance and greater AUC variability than static face. The strong appearance-only result is consistent with possible speaker, episode or production-condition cues, although this experiment does not identify the features responsible.

Gated Prior-KL achieves higher lie-class F1 than static face alone (62.76% versus 54.39%), indicating a trade-off between class-specific decisions and overall ranking. Because the ablation uses one seed per fold, the comparison is diagnostic and does not establish a general multimodal advantage. Figure 5 summarizes AUC and balanced accuracy.

3.3. Exploratory vote-score experiment

The exploratory vote-score-sum variant uses gamma = 0.5, floor = 0.05 and consensus_bonus = 0.0, fixed across three folds and three seeds (Table 7). It obtains higher mean AUC and balanced accuracy at the validation-selected threshold than Gated Prior-KL. The scoring rule and parameter values were developed through preliminary experiments after inspection of the same test folds, rather than selected by a validation-only search. These results therefore characterize an exploratory agreement-aware fusion method; confirmation requires evaluation with the scoring rule fixed before testing on independent data.

3.4. Interpretation relative to the DOLOS reference

The original DOLOS paper reports 64.58% AUC for PAVF plus multi-task supervision [1]. Gated Prior-KL reaches 64.16 ± 4.56% mean AUC in our setting. The similar AUC values provide contextual reference rather than a matched comparison: the present pipeline differs in preprocessing and architecture, uses only clip-level Truth/Lie supervision, and aggregates performance across fold-seed runs. The result places this simpler gated system near the reported reference AUC, while its thresholded classification performance remains modest.

3.5. Limitations and implications

Lie-class F1 varies substantially across folds and random seeds, and the overall performance remains modest. Weak clip-level supervision, variable modality quality and limited adaptation of the frozen backbones may contribute to this behavior. Clip-level labels also do not identify which facial movements or speech events are associated with deception. The current results support research on multimodal fusion, not high-stakes judgments about individuals.

The predefined splits do not establish identity- or episode-disjoint generalization; further evaluation is needed to assess potential shortcuts. Label-independent filtering may still alter the usable subset.

4. Conclusions

This study evaluates a three-stream audio-visual pipeline on DOLOS using controls that distinguish directed gate initialization from Prior-KL regularization. Across three folds and three random seeds, the directed-initialized Prior-KL model achieves 64.16 ± 4.56% AUC, 58.96 ± 4.16% balanced accuracy at the validation-selected threshold, 56.74 ± 10.67% lie-class F1, and 57.45 ± 5.15% macro-F1. Directed initialization is associated with higher mean AUC than the uniform control; additional KL regularization yields higher lie-class F1 without improving AUC or balanced accuracy.

The seed-matched ablation shows complementary metric behavior: static facial appearance has higher AUC and balanced accuracy, whereas fusion has higher lie-class F1. Further evaluation should extend single-stream ablations across seeds and test identity-disjoint generalization. The exploratory vote-score-sum findings warrant independent confirmation. Parameter-efficient adaptation through LoRA or adapters is another direction for improving the frozen backbones under the same evaluation protocol [13], [14].

5. Declarations

5.1. Author Contributions

Huy Phan Anh: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Visualization, Writing - original draft, and Writing - review and editing.

5.2. Institutional Review Board Statement

This study is a secondary analysis of the existing DOLOS dataset accessed with approval from its provider. No new participants were recruited and no new human recordings were collected.

5.3. Informed Consent Statement

This secondary analysis used existing gameshow material in DOLOS under provider-approved access. The present study did not recruit participants or collect new recordings.

5.4. Data Availability Statement

The DOLOS dataset must be obtained from its original provider under the stated academic, non-commercial access conditions (https://rose1.ntu.edu.sg/dataset/DOLOS/). Code, experiment configurations and aggregate results are available at https://github.com/Pah2704/lie_detection_paper_code (public code snapshot: commit 77730a6ade851f3d54932cbdda5192edd6065023). This release excludes DOLOS annotations, timestamp and split-membership tables, original videos, derived media/caches, and clip/window prediction tables.

5.5. Acknowledgment

The author would like to thank the ROSE Lab at Nanyang Technological University (NTU) for granting access to the DOLOS dataset for this research.

5.6. Conflicts of Interest

The author declare no conflicts of interest.

5.7. Generative AI Use Disclosure

During the preparation of this work, the authors used OpenAI ChatGPT to assist with code debugging, literature summarization, revision planning, and language editing. All research concepts, schematics, and data visualizations were independently developed by the authors from verified datasets. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

6. References

[1] X. Guo, N. M. Selvaraj, Z. Yu, A. W.-K. Kong, B. Shen, and A. Kot, "Audio-visual deception detection: Dolos dataset and parameter-efficient cross-modal learning," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 22135–22145. https://doi.org/10.1109/ICCV51070.2023.02023.

[2] ROSE Lab, "DOLOS Audio-Visual Deception Dataset," Nanyang Technological University. Accessed: Jul. 23, 2026. [Online]. Available: https://rose1.ntu.edu.sg/dataset/DOLOS/.

[3] E. Braverman *et al.*, "DecepBench: Benchmarking Multimodal Deception Detection," in *Proceedings of the Third Workshop on Social Influence in Conversations (SICon 2025)*, 2025, pp. 50–64. https://doi.org/10.18653/v1/2025.sicon-1.3.

[4] A. D'Ulizia, A. D'Andrea, P. Grifoni, and F. Ferri, "Analysis, evaluation, and future directions on multimodal deception detection," *Technologies*, vol. 12, no. 5, p. 71, 2024. https://doi.org/10.3390/technologies12050071.

[5] S. A. Prome, N. A. Ragavan, M. R. Islam, D. Asirvatham, and A. J. Jegathesan, "Deception detection using machine learning (ML) and deep learning (DL) techniques: A systematic review," *Natural Language Processing Journal*, vol. 6, p. 100057, 2024. https://doi.org/10.1016/j.nlp.2024.100057.

[6] P. Ekman and W. V. Friesen, "Nonverbal leakage and clues to deception," *Psychiatry*, vol. 32, no. 1, pp. 88–106, 1969. https://doi.org/10.1080/00332747.1969.11023575.

[7] A. Vrij, Detecting lies and deceit: Pitfalls and opportunities. John Wiley & Sons, 2008. https://books.google.co.id/books?id=20pg76wmAucC.

[8] V. Pérez-Rosas, M. Abouelenien, R. Mihalcea, and M. Burzo, "Deception detection using real-life trial data," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, 2015, pp. 59–66. https://doi.org/10.1145/2818346.2820758.

[9] T. Pfister, X. Li, G. Zhao, and M. Pietikäinen, "Recognising spontaneous facial micro-expressions," in *2011 international conference on computer vision*, IEEE, 2011, pp. 1449–1456. https://doi.org/10.1109/ICCV.2011.6126401.

[10] Y. Li, J. Wei, Y. Liu, J. Kauttonen, and G. Zhao, "Deep learning for micro-expression recognition: A survey," *IEEE Transactions on Affective Computing*, vol. 13, no. 4, pp. 2028–2046, 2022. https://doi.org/10.1109/TAFFC.2022.3205170.

[11] S.-T. Liong, J. See, K. Wong, and R. C.-W. Phan, "Less is more: Micro-expression recognition from video using apex frame," *Signal Processing: Image Communication*, vol. 62, pp. 82–92, 2018. https://doi.org/10.1016/j.image.2017.11.006.

[12] X. Ben *et al.*, "Video-based facial micro-expression analysis: A survey of datasets, features and algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 9, pp. 5826–5846, 2021. https://doi.org/10.1109/TPAMI.2021.3067464.

[13] E. J. Hu *et al.*, "Lora: Low-rank adaptation of large language models.," *arXiv preprint arXiv:2106.09685*, 2021. https://doi.org/10.48550/arXiv.2106.09685.

[14] N. Houlsby *et al.*, "Parameter-efficient transfer learning for NLP," in *Proceedings of the 36th International Conference on Machine Learning*, PMLR, 2019, pp. 2790–2799. https://proceedings.mlr.press/v97/houlsby19a/houlsby19a.pdf.

[15] C. Lugaresi *et al.*, "Mediapipe: A framework for building perception pipelines," *ArXiv preprint arXiv:1906.08172*, 2019. https://doi.org/10.48550/arXiv.1906.08172.

[16] C. Zach, T. Pock, and H. Bischof, "A duality based approach for realtime tv-l 1 optical flow," in *Joint pattern recognition symposium*, Springer, 2007, pp. 214–223. https://doi.org/10.1007/978-3-540-74936-3_22.

[17] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020. https://doi.org/10.48550/arXiv.2010.11929.

[18] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM Computing Surveys*, vol. 54, no. 10s, pp. 1–41, 2022. https://doi.org/10.1145/3505244.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016*

IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778. https://doi.org/10.1109/CVPR.2016.90.

[20] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pp. 12449–12460, 2020. https://dl.acm.org/doi/abs/10.5555/3495724.3496768.

[21] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2018. https://doi.org/10.1109/TPAMI.2018.2798607.

[22] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Information Fusion*, vol. 37, pp. 98–125, 2017. https://doi.org/10.1016/j.inffus.2017.02.003.

[23] P. Xu, X. Zhu, and D. A. Clifton, "Multimodal learning with transformers: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 10, pp. 12113–12132, 2023. https://doi.org/10.1109/TPAMI.2023.3275156.

[24] A. Vaswani *et al.*, "Attention is all you need," *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, vol. 30, 2017. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[25] A. Paszke *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, vol. 32, 2019. https://proceedings.neurips.cc/paper_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf.

[26] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," , *arXiv preprint arXiv:1711.05101*, 2017. https://doi.org/10.48550/arXiv.1711.05101.

[27] I. Loshchilov and F. Hutter, "SGDR: Stochastic Gradient Descent with Warm Restarts," *arXiv preprint arXiv:1608.03983*, 2017. https://doi.org/10.48550/arXiv.1608.03983.

[28] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006. https://doi.org/10.1016/j.patrec.2005.10.010.

[29] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, "The balanced accuracy and its posterior distribution," in *2010 20th International Conference on Pattern Recognition*, IEEE, 2010, pp. 3121–3124. https://doi.org/10.1109/ICPR.2010.764.

[30] D. M. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," *arXiv preprint arXiv:2010.16061*, 2020. https://doi.org/10.48550/arXiv.2010.16061.