CAEIR: CROSS-ARCHITECTURE EXPERTS WITH INTERACTION-AWARE ROUTING FOR DEEPFAKE DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Generalizable deepfake detection requires identifying weak and heterogeneous forensic cues that vary across manipulation methods and data distributions. Existing mixture-of-experts detectors, however, typically specialize experts within a single architectural family, while expert-selection decisions are often based solely on the input representation, without explicitly comparing the forensic evidence produced by heterogeneous experts. Our key insight is that routing should depend not only on the input, but also on how heterogeneous experts respond to each sample. We introduce CAEIR, a cross-architecture framework that places convolutional, Transformer, and state-space experts in a unified competitive pool, summarizes their responses as evidence nodes, and models interactions among these nodes to produce expert-specific routing corrections that refine an input-conditioned Top-1 ranking. A separate bounded modulator scales the selected response without altering the routing decision. All candidate responses are computed to support interaction-aware expert selection, while only the selected response is injected into the detection residual. Trained on FF++ real videos with online synthetic fake supervision, CAEIR achieves video-level AUCs of 96.61%, 95.71%, 82.19%, and 89.76% on Celeb-DF v2, DFD, DFDC, and FFIW, respectively. Controlled ablations further support the benefits of cross-architecture diversity and response-conditioned routing beyond full-candidate evaluation alone.

1 INTRODUCTION

Generalizing deepfake detectors to unseen manipulation methods remains challenging. Learned representations can capture identity-related or method-specific patterns, causing performance to degrade when the manipulation process or data distribution changes (Dong et al., 2023; Yan et al., 2023a;b). These observations motivate extracting transferable forensic evidence across multiple spatial scales and dependency patterns rather than relying on a single shortcut.

Forgery evidence is heterogeneous: local textures, boundary transitions, and facial-component inconsistencies can reveal low-level synthesis or blending artifacts (Shiohara & Yamasaki, 2022; Nguyen et al., 2024; Han et al., 2025), whereas long-range relationships between facial parts and ordered dependencies across a feature grid require broader context. Convolution, self-attention, and visual state-space models implement different information-propagation mechanisms: convolutions aggregate local neighborhoods, self-attention establishes input-adaptive global interactions, and state-space models propagate context along an ordered scan (Dai et al., 2021; Peng et al., 2021; Liu et al., 2024b). These complementary mechanisms suggest that the appropriate computational perspective may depend on the evidence present in each sample.

Mixture-of-experts (MoE) provides a natural mechanism for sample-adaptive specialization. A router can select an expert path for each input while keeping the final representation compact. However, existing MoE designs for forgery detection and related tasks generally specialize experts within one architectural family, for example by changing convolutional adapters, low-rank factors, facial regions, abstraction levels, or training domains (Miao et al., 2025; Kong et al., 2026; Dutta et al., 2026; Fang et al., 2025). Such specialization increases capacity but does not expose the fundamentally different inductive biases of local convolution, global attention, and directional state propagation in one

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

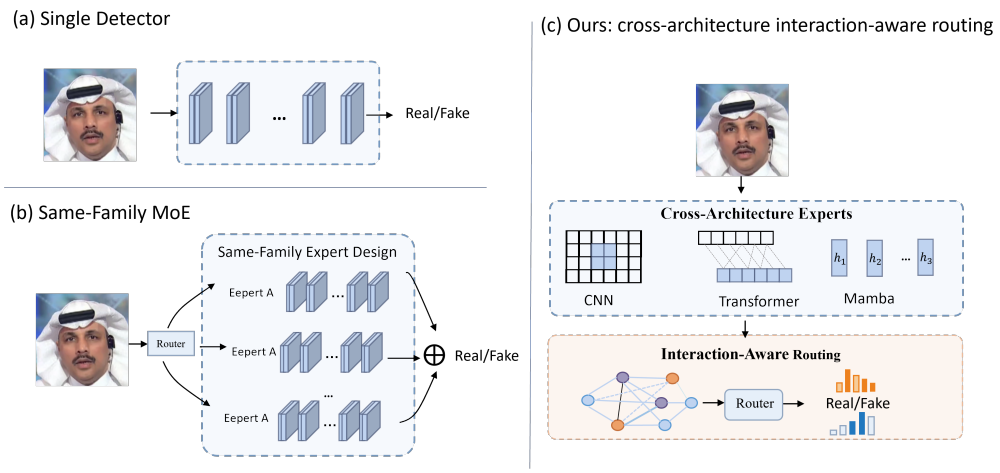


Figure 1: Conceptual comparison. Within-family MoE designs select among candidates with similar information propagation, whereas CAEIR combines architecture-level heterogeneous experts and uses their candidate responses to condition the Top-1 selection.

decision space. Moreover, when experts from different architectures process the same sample, their responses can agree, complement one another, or support different choices. A conventional input-conditioned gate based only on a CLS summary cannot directly use these already formed candidate responses.

This paper asks how to make routing aware of cross-architecture evidence while preserving a single dominant response in the final residual. Direct Top-k routing and soft mixtures allow several outputs to contribute to the prediction, whereas our objective is to model relations among candidate responses before dispatch. The resulting route is conditioned on both the global input prior and the evidence produced by all candidate architectures. Figure 1 contrasts single-detector, within-family MoE, and cross-architecture routing designs.

We introduce CAEIR (Cross-Architecture Experts with Interaction-Aware Routing for Deepfake Detection). The method combines CNN, Transformer, and Mamba-based state-space experts in a unified competitive pool. For an input feature map, all candidate responses are first computed and summarized as expert evidence nodes. A lightweight expert-interaction module then relates these nodes and generates interaction-conditioned routing corrections. The corrections are added to the CLS-based router logits to update expert preferences, and the selected expert response is dispatched to the detection residual.

The contributions are:

- We formulate an architecture-level heterogeneous MoE for deepfake detection that puts CNN, Transformer, and Mamba experts in one competitive routing pool, exposing distinct local, global, and state-space propagation mechanisms.
- We propose candidate-response-conditioned expert interaction. Candidate responses are summarized as evidence nodes, attention constructs a cross-expert relation context, and interaction-conditioned calibration updates the Top-1 ranking before output dispatch.
- We provide controlled cross-dataset and structural analyses that examine architecture diversity, candidate computation, and response-conditioned interaction. We also include a qualitative Detector-to-LLM analysis that verbalizes prediction and routing traces.

2 Related Work

2.1 Generalizable Deepfake Detection

Cross-dataset generalization is a central challenge in face forgery detection. Online synthetic blending constructs training examples that reduce dependence on a specific manipulation method

(Shiohara & Yamasaki, 2022); FSBI extends self-blending with frequency-domain feature extraction (Abul Hasanaath et al., 2025). Other approaches include alternating optimization of spatial and temporal weights (Wang et al., 2023), localized artifact attention (Nguyen et al., 2024), vulnerability-aware spatio-temporal learning (Nguyen et al., 2025), facial-component adaptation (Han et al., 2025), and temporal forgery cue unraveling (Guo et al., 2025b). Unified benchmarks further show that data construction, backbone choice, preprocessing, and video aggregation can materially change reported cross-domain conclusions (Yan et al., 2023b; Zhao et al., 2025). Our experiments follow a single-source protocol in which FF++ real videos are the only real training source and online synthetic blending provides the auxiliary fake supervision.

2.2 Vision-Language Models for Explainable Forgery Detection

Recent work uses vision-language models for direct classification and explanation, detector-guided generation, or joint vision-language optimization (Guo et al., 2025a; Xu et al., 2026; Peng et al., 2025; He et al., 2025). LLaVA provides a general visual-token alignment paradigm (Liu et al., 2023). Our Detector-to-LLM interface focuses on fine-grained forensic evidence by exposing prediction probabilities and routing traces as structured non-text tokens.

2.3 Mixture of Experts and Cross-Architecture Routing

Sparse-gated MoE routes samples or tokens among multiple experts (Shazeer et al., 2017), and systematic studies of vision routers show that routing rules influence utilization and generalization (Liu et al., 2024a). In forgery detection, experts have been differentiated by noise modeling, facial regions, domains, abstraction levels, or synthesis sources (Miao et al., 2025; Luo et al., 2026; Sun et al., 2026; Dutta et al., 2026; Fang et al., 2025). Wang and Liu (2025), for example, define heterogeneity through expert capacity and scale and fuse the outputs of an input-selected Top-k subset. Related work on convolution-attention hybrids and visual state-space models motivates a broader candidate space spanning CNN, Transformer, and Mamba architectures (Dai et al., 2021; Peng et al., 2021; Liu et al., 2024b; Li et al., 2024).

Positioning of CAEIR. The distinguishing design choice in CAEIR is how architectural heterogeneity participates in routing. Selecting heterogeneous experts from the input representation leaves their sample-specific responses outside the selection decision. CAEIR treats architecture-specific candidate responses as routing evidence: all candidates respond first, their relations refine the input-conditioned ranking, and only the selected response is injected into the residual. Heterogeneous experts therefore provide both alternative feature transformations and evidence for deciding which transformation should dominate each sample. This combination of architecture-level diversity, candidate-response-conditioned selection, and Top-1 output dispatch defines the distinction from the input-selected expert pools discussed above.

3 Method

3.1 Problem Formulation and Overview

Given an input image I, let the l-th ViT block output $\widehat{X}^l = [\widehat{x}^l_{\text{cls}}, \widehat{X}^l_{\text{sp}}]$, where $\widehat{x}^l_{\text{cls}} \in \mathbb{R}^C$ is the CLS token and $\widehat{X}^l_{\text{sp}} \in \mathbb{R}^{N\times C}$ contains spatial tokens. We augment the last nine blocks of the 12-block backbone, indexed as $\mathcal{M} = \{4, 5, \ldots, 12\}$. For $l \in \mathcal{M}$, the MoE branch operates after the ViT block and injects an output into the spatial residual; other blocks remain unchanged. Parameters are independent across augmented blocks.

A conventional input-conditioned router scores experts from the CLS token before observing their responses. CAEIR instead decomposes routing into an input prior and a candidate-response correction. Each augmented block computes all candidate responses, models their relations, calibrates the route scores, and dispatches one selected response to the spatial residual. Figure 2 summarizes the resulting architecture.

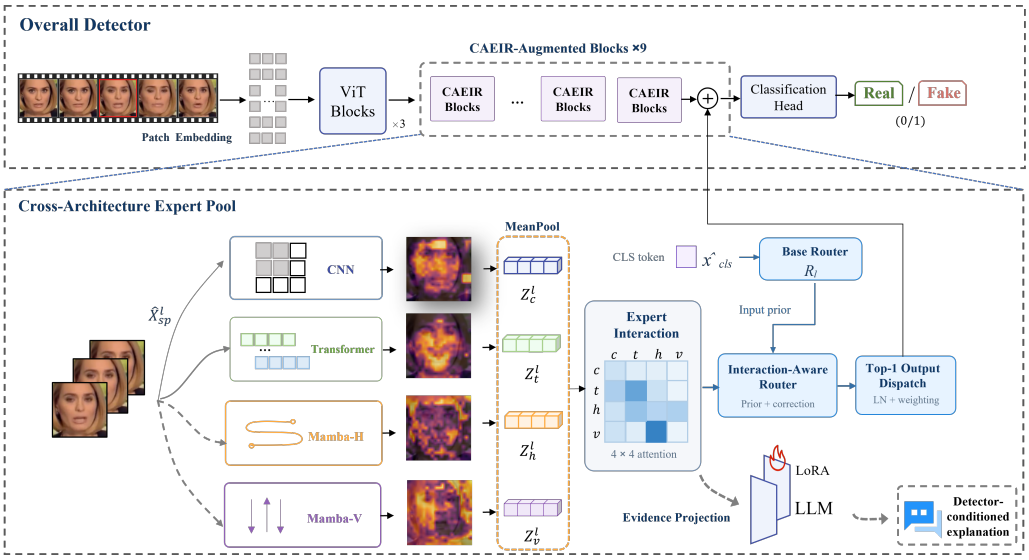


Figure 2: Overview of CAEIR. Nine ViT blocks are augmented with a cross-architecture expert pool. Each block pools four candidate responses into evidence nodes; expert interaction calibrates the CLS-based routing prior before Top-1 output dispatch. An auxiliary evidence projection supports Detector-to-LLM explanations without changing the detection path.

3.2 Architecture-Level Heterogeneous Expert Pool

Architecture-level heterogeneity places CNN, Transformer, and state-space experts in one competitive pool. Let $\mathcal{E} = \{\text{CNN}, \text{Tr}, \text{Mamba-H}, \text{Mamba-V}\}$ denote its four candidates. Given post-ViT spatial tokens, each candidate produces a response that is pooled into an evidence node:

$$F_i^l = E_{i,l}(\widehat{X}_{\text{sp}}^l) \in \mathbb{R}^{N\times C}, \qquad z_i^l = \text{MeanPool}(F_i^l) \in \mathbb{R}^C, \qquad Z^l = [z_i^l]_{i\in\mathcal{E}} \in \mathbb{R}^{4\times C}. \tag{1}$$

The pooled node provides compact routing evidence and does not replace the spatial expert response. The full response F_i^l is retained and, if expert i is selected, enters the residual through Eq. (4). This separation lets cross-expert comparison operate on a fixed-size representation while preserving spatial forensic information in the selected response for the final detector update.

The candidates share the same input and output shape but implement local convolution, global self-attention, and directional state propagation. The two state-space experts are independent single-direction selective-scan modules based on Mamba (Gu & Dao, 2024). Mamba-H scans spatial tokens in row-major order. Mamba-V transposes the patch grid before flattening, scans the resulting sequence, and restores the original patch order afterward.

3.3 Candidate-Response-Conditioned Expert Interaction

The input summary alone does not reveal what evidence each architecture extracted for the current sample. We therefore pass the evidence nodes through a layer-specific interaction module over the expert dimension:

$$H^l = \mathcal{I}_l(Z^l) \in \mathbb{R}^{4\times C}. \tag{2}$$

Specifically, $\mathcal{I}_l$ is a pre-norm multi-head self-attention–MLP block applied across the four experts. Let h_i^l denote row i of H^l; it is conditioned on all candidate nodes. The layer-specific projections are therefore learned under the detection objective to translate cross-expert relations into routing corrections.

3.4 Interaction-Conditioned Route Calibration and Top-1 Output Routing

$$\begin{aligned} \mathbf{d}^l &= H^l w_{s,l}, \\ \mathbf{s}^l &= R_l(\widehat{x}_{\text{cls}}^l) + \alpha \mathbf{d}^l. \end{aligned} \tag{3}$$

The first term provides an input-conditioned prior, while the second introduces a candidate-response-conditioned correction derived from cross-expert interactions. For experts i and j, their calibrated logit difference is

$$s_i^l - s_j^l = \left[R_{l,i}(\widehat{x}_{\text{cls}}^l) - R_{l,j}(\widehat{x}_{\text{cls}}^l)\right] + \alpha(d_i^l - d_j^l),$$

where $R_{l,i}$ denotes the i-th router output. Thus, candidate-response interaction changes relative routing preferences through expert-specific correction differences. A shared logit shift would leave both the softmax distribution and the Top-1 ranking unchanged.

$$\begin{aligned} \mathbf{p}^l &= \text{Softmax}(\mathbf{s}^l), \qquad k_l^* = \arg\max_{i \in \mathcal{E}} p_i^l, \\ m_{k_l^*}^l &= 1 + \beta \tanh(w_{f,l}^\top h_{k_l^*}^l + b_{f,l}), \\ X^{l+1} &= \widehat{X}^l + \gamma_l \left[\mathbf{0},\ p_{k_l^*}^l m_{k_l^*}^l \,\text{LN}_{\text{sp},l}\Big(F_{k_l^*}^l\Big)\right], \qquad l \in \mathcal{M}. \end{aligned} \tag{4}$$

The bounded modulation is applied only after expert selection, so it cannot alter k_l^*. Here $\mathbf{0} \in \mathbb{R}^C$ leaves the CLS position unchanged and γ_l is initialized to zero. All four candidates participate in pre-selection relation modeling, but only the selected response enters the residual. Because all candidates are evaluated to form Z^l, CAEIR is output-sparse and retains the full candidate-expert computation cost.

3.5 Detector-to-LLM Interface

An auxiliary Detector-to-LLM interface exposes visual semantics and detector evidence as non-text tokens. A frozen visual encoder and a trainable projection produce $Z_v = P_v(F_v(I))$. The frozen detector provides the prediction $q = [P(\text{real}), P(\text{fake})] \in \mathbb{R}^2$, route distributions $p^{(l)} \in \mathbb{R}^4$ from early, middle, and late MoE layers, and their mean $\bar{p}$. We use the first, fifth, and ninth augmented MoE layers (ViT blocks $l = 4, 8, 12$), denoting their route distributions by $p^{(\text{early})} = p^{(4)}$, $p^{(\text{mid})} = p^{(8)}$, and $p^{(\text{late})} = p^{(12)}$. We form five six-dimensional evidence slots and project the resulting sequence:

$$Z_d = P_d\Big(\Big[[q; \mathbf{0}_4],\ [\mathbf{0}_2; p^{(\text{early})}],\ [\mathbf{0}_2; p^{(\text{mid})}],\ [\mathbf{0}_2; p^{(\text{late})}],\ [\mathbf{0}_2; \bar{p}]\Big]\Big). \tag{5}$$

The `<image>` and `<deepfake>` placeholders are replaced by Z_v and Z_d, respectively, and the resulting embeddings are concatenated with the text embeddings. The LLM generates explanations from this context. We use a frozen CLIP ViT-L/14@336 visual encoder (Radford et al., 2021) and a LLaMA-7B-based LLaVA language model (Liu et al., 2023), with trainable LoRA adapters (Hu et al., 2022) applied to the language model's attention and MLP projection layers during explanation training. The detector is trained first; it remains frozen during subsequent evidence-projection and explanation training. No explanation loss is back-propagated to the detector, so the interface is post-decision and does not modify the route or real/fake prediction.

4 Experiments

4.1 Datasets and Evaluation Protocol

We use the c23-compressed version of FaceForensics++ (FF++; Rossler et al., 2019) as the only training domain. Real examples are taken from the official FF++ real-video split, while auxiliary fake supervision is generated online using the self-blending procedure of Shiohara and Yamasaki (2022). The training set excludes the pre-generated DeepFakes, Face2Face, FaceSwap, and NeuralTextures videos from FF++, as well as samples from any external evaluation datasets.

For training and validation, eight temporal positions are uniformly sampled over the full duration of each FF++ video and fixed across epochs. Training applies stochastic spatial, photometric, compression, and downsampling augmentations; validation omits these stochastic augmentations.

The main evaluation uses Celeb-DF v2 (CDF; Li et al., 2020), Google DeepFakeDetection (DFD; Dufour & Gully, 2019), the Deepfake Detection Challenge (DFDC; Dolhansky et al., 2020), and FFIW (Zhou et al., 2021). CDF uses its complete official test set. For external evaluation, we uniformly sample 32 temporal positions across each video and detect faces using RetinaFace. At

each position, we retain boxes with at least 50% of the largest detected face's bounding-box area, expand them by one eighth of their width and height on each side, and obtain 336×336 crops through CLIP preprocessing (Appendix B). Detection failures and multiple faces can change the crop count. We report frame- and video-level AUC, averaging the highest 20% of valid-crop fake probabilities for each video score.

We transcribe external results from original papers or their benchmark tables. These include the FF++ Xception baseline (Rossler et al., 2019), with its cross-dataset scores taken from the ResProto-FD benchmark table (Jing et al., 2026b); EffiNet-B4 and F3-Net as reported by SRI-Net (Fei et al., 2026); and LocalRL (Chen et al., 2021) and FTCN (Zheng et al., 2021). The 2022 sources are SBI (Shiohara & Yamasaki, 2022), SLADD (Chen et al., 2022), CORE (Ni et al., 2022), DCL (Sun et al., 2022), and UIA-ViT (Zhuang et al., 2022). The SBI CDF, DFD, and DFDC scores in Table 1 follow the video-level benchmark table of ResProto-FD (Jing et al., 2026b); its FFIW score comes from the SBI paper (Shiohara & Yamasaki, 2022). The 2023–2024 sources are UCF (Yan et al., 2023a), SeeABLE (Larue et al., 2023), CFM (Luo et al., 2024), and LSDA (Yan et al., 2024). The 2025–2026 sources include DFMoE (Wang & Liu, 2025), DFD-FCG (Han et al., 2025), MoE-FFD (Kong et al., 2026), LGF-Net (Pandey et al., 2026), CAST (Thakre et al., 2026), ACMD (Jing et al., 2026a), ResProto-FD (Jing et al., 2026b), and WAFF (Pan et al., 2026).

Because training data and video aggregation differ across studies, Table 1 provides contextual rather than protocol-matched comparisons. All CAEIR results use frozen dataset manifests, a unified evaluation script, and fixed random seeds.

4.2 Main Cross-Dataset Comparison

Table 1: Published cross-dataset video-level AUC (%). Avg. is reported only when all four dataset scores are available. ‡ indicates that video aggregation is not documented in the source table. A dash indicates that the result was not reported.

Method	CDF	DFD	DFDC	FFIW	Avg.
Xception	81.60	89.60	73.20	–	–
EffiNet-B4	80.80	86.20	–	–	–
F3-Net	78.90	84.40	–	–	–
LocalRL‡	78.26	89.24	76.53	–	–
FTCN	86.90	94.40	71.00	74.47	81.69
UIA-ViT	82.40	94.70	–	–	–
SBI	90.60	87.70	75.20	84.83	84.58
SLADD	79.70	–	77.20	68.30	–
CORE	80.90	88.20	72.10	71.00	78.05
DCL	88.20	92.10	75.00	–	–
UCF	83.70	86.70	77.00	69.70	79.28
SeeABLE	87.30	–	75.90	–	–
CFM	89.65	95.21	80.22	–	–
LSDA	89.80	95.60	73.50	72.40	82.83
DFMoE	79.32	81.92	73.95	–	–
DFD-FCG	95.00	–	81.80	–	–
MoE-FFD	91.28	93.57	–	–	–
LGF-Net	85.72	–	76.34	–	–
CAST	76.98	93.31	81.25	–	–
ACMD	89.89	–	76.27	–	–
ResProto-FD	91.50	95.80	81.90	–	–
WAFF	87.43	**96.02**	79.63	74.30	84.35
CAEIR (Ours)	**96.61**	95.71	**82.19**	**89.76**	**91.07**

Table 1 places CAEIR alongside published cross-dataset results. CAEIR obtains 96.61%, 95.71%, 82.19%, and 89.76% video AUC on CDF, DFD, DFDC, and FFIW, respectively, with an average of 91.07% across these four datasets.

Within this contextual comparison, CAEIR achieves the highest reported CDF, DFDC, and FFIW scores among the listed methods with available results, and the highest four-dataset average among methods reporting all four scores. Its DFD score is 0.31 percentage points below WAFF at 96.02%.

This pattern supports transfer across multiple external datasets under our single-source training setup. The ranking compares published scores under the protocol differences described in Section 4.1.

4.3 Implementation Details

We use a ViT-S/16 backbone initialized from ImageNet-pretrained timm weights and operate it at an input resolution of 336×336. With 16×16 patches, this yields a 21×21 patch grid plus one CLS token, with hidden dimension $C = 384$. CAEIR augments the last nine of the 12 Transformer blocks (ViT blocks 4–12) with the four-expert pool described in Section 3.2.

The CNN expert uses depthwise 3×3 convolution, BatchNorm, SiLU, and pointwise 1×1 convolution. The Transformer expert uses eight attention heads and an MLP ratio of four. The two independent Mamba experts use row-major and transposed-grid scan orders, respectively, with $d_{\text{state}} = 16$, $d_{\text{conv}} = 3$, expansion ratio 2, and internal dimension $C/2$. The interaction module uses four heads, dropout 0.1, and MLP hidden dimension $\max(C/2, 64)$. We set $\alpha = 0.1$ and $\beta = 0.075$. Models are optimized with Adam at a fixed learning rate of 10^{-4}, with zero weight decay and no schedule or warmup. Training uses a batch size of 24 with five-step gradient accumulation for at most 60 epochs.

The training objective is $\mathcal{L} = \mathcal{L}_{\text{CE}} + 0.005\mathcal{L}_{\text{lb}}$. Load balancing combines batch-average routing probabilities and hard Top-1 selection frequencies, averaged over the nine MoE layers (Appendix B).

4.4 Controlled Routing and Structural Ablations

Table 2 compares routing configurations that modify routing regularization, expert sparsity, load regulation, shared processing, or conflict-aware calibration under the unified nine-layer protocol while keeping the remaining training and evaluation settings unchanged. Detailed definitions are provided in Appendix B.

Table 2: Routing ablation on CDF under the unified nine-layer protocol. Values are frame-level and video-level AUC (%).

Routing strategy	CDF Frame AUC	CDF Video AUC
Input-conditioned Top-1	88.40	94.82
Orthogonality-regularized	89.17	95.51
Top-2 cooperative	88.97	94.92
Load-biased	89.75	96.03
Shared-CNN	88.60	95.23
Soft expert mixture	89.16	96.00
Conflict-entropy calibration	88.31	94.75
CAEIR (interaction-aware Top-1)	**90.36**	**96.61**

The input-conditioned Top-1 baseline obtains 88.40% frame AUC and 94.82% video AUC, whereas CAEIR reaches 90.36% and 96.61%, respectively: gains of 1.96 and 1.79 percentage points.

Among the alternative routing strategies, load-biased routing and the soft expert mixture achieve the strongest video AUCs at 96.03% and 96.00%, respectively, below CAEIR at 96.61%. Top-2 cooperative routing reaches 94.92%, only 0.10 percentage points above the Top-1 baseline. In these evaluated configurations, allowing more expert outputs to contribute or changing routing regularization and load regulation does not match the accuracy obtained with candidate-response-conditioned interaction. These results support the benefit of candidate-response-conditioned interaction beyond increased candidate participation alone, with the associated parameter and latency differences reported in Table 3.

In Table 3, evaluating all candidates without interaction changes frame/video AUC from 88.40/94.82% to 89.37/94.88%, profiled computation from 31.14 to 44.37 GFLOPs, and median latency from 10.46 to 19.21 ms. Adding candidate-response interaction reaches 90.36/96.61% AUC (gains of 0.99/1.73 percentage points), adds approximately 6.7M parameters, and increases median latency to 23.74 ms. Measurements cover the detector encoder and output projection, excluding

face preprocessing and the Detector-to-LLM interface. The gain over the full-candidate baseline is therefore not explained by full-candidate evaluation alone.

Table 3: Accuracy–efficiency trade-off under the unified nine-layer protocol. Measurements use $1 \times 3 \times 336 \times 336$ inputs, batch size 1, and a single RTX 5090. † Profiled FLOPs are lower bounds because selective-scan operations may not be fully counted.

Routing configuration	Params (M)	GFLOPs per frame†	Memory (GB)	Latency p50/p95 (ms)	Frame AUC	Video AUC
Selected expert only, no interaction	43.86	31.14	0.202	10.46/11.85	88.40	94.82
All candidates, no interaction	43.86	44.37	0.202	19.21/22.02	89.37	94.88
CAEIR	50.53	44.42	0.229	23.74/25.30	**90.36**	**96.61**

Table 4: Architecture-pool ablations under the unified nine-layer protocol. Panel (a) compares heterogeneous and within-family expert pools; panel (b) compares the four-expert configuration with three-expert single-scan variants. All rows use seed 3407 and report CDF AUC (%).

Expert pool / variant	Composition	Frame AUC	Video AUC
(a) Architecture-family composition			
CAEIR (Ours)	CNN + Transformer + Mamba-H + Mamba-V	**90.36**	**96.61**
CNN-only expert pool	Four independent CNN experts	87.97	94.56
Transformer-only expert pool	Four independent Transformer experts	90.00	95.94
State-space-only expert pool	Two Mamba-H + two Mamba-V experts	89.29	95.96
(b) State-space scan composition			
CAEIR (single horizontal state-space branch)	CNN + Transformer + one horizontal scan	88.94	94.93
CAEIR (single vertical state-space branch)	CNN + Transformer + one vertical scan	89.93	96.36

In Table 4(a), CAEIR achieves the highest frame and video AUC among the evaluated expert pools. Its video AUC exceeds the best within-family result by 0.65 percentage points. Both three-expert single-scan variants yield lower AUCs than the full four-expert CAEIR configuration (Table 4(b)); this comparison changes both scan composition and pool size. Together with Table 3, these results support the effectiveness of combining a cross-architecture expert pool with candidate-response interaction.

Table 5: MoE insertion-depth ablation.

MoE depth	Frame AUC	Video AUC
3	89.89	95.52
6	90.17	96.39
9 (main)	**90.36**	**96.61**
12	89.36	95.46

Table 6: Response-modulation sensitivity.

Scale β	Frame AUC	Video AUC
0.050	89.95	96.21
0.075 (main)	**90.36**	**96.61**
0.095	87.71	93.79

Both tables report CDF frame-level and video-level AUC (%) under the unified protocol. Table 6 uses nine MoE layers and fixed routing-correction scale $\alpha = 0.1$; $\beta = 0.075$ is the CAEIR (interaction-aware Top-1) configuration.

For depth d in Table 5, CAEIR augments the last d ViT blocks. Among the evaluated insertion depths, nine MoE blocks achieve the highest frame and video AUC; performance decreases when all twelve blocks are augmented.

Increasing β from 0.050 to 0.075 improves both CDF AUCs, but 0.095 causes a decline, indicating non-monotonic sensitivity. We use $\beta = 0.075$, yielding 90.36% frame AUC and 96.61% video AUC.

4.5 Detector-to-LLM Analysis

The Detector-to-LLM interface verbalizes the detector decision alongside its route profile and facial evidence. Figure 3 shows four representative cases. For the three correctly classified examples, the generated text associates real predictions with smooth skin or regular facial structure, and fake predictions with low-detail skin, blurred or misaligned facial parts, or unusual proportions. In the

final row, the detector misclassifies a real sample as fake, and the generated explanation follows this error. We therefore treat the interface as a qualitative verbalization of detector traces; assessing explanation fidelity requires a dedicated benchmark and annotation protocol.

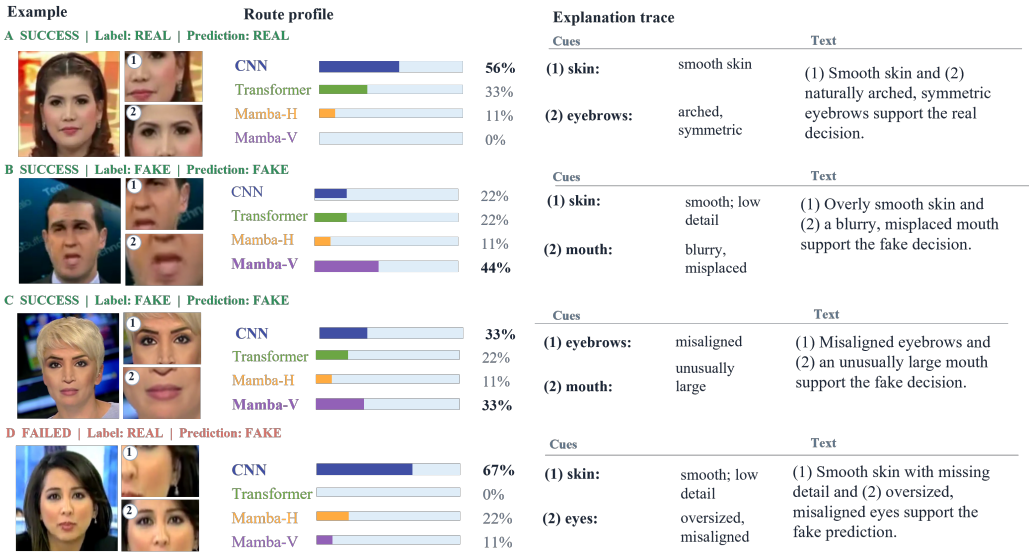


Figure 3: Qualitative Detector-to-LLM examples. Each row shows the input face and selected crops, the route profile, visual cues, and generated explanation. Rows A–C are correctly classified; row D is a real sample misclassified as fake.

5 Discussion and Limitations

CAEIR is output-sparse but evaluates all candidates for routing evidence. All experts inform route formation, while single-expert dispatch preserves an architecture-specific residual update. A dense soft mixture instead combines all candidate outputs in the residual. The main practical limitation is the cost of computing candidate responses and their interactions. Table 3 quantifies this accuracy–efficiency trade-off, including the additional parameters and latency. Future work could reduce this cost through lightweight evidence predictors, hierarchical routing, or shared computations while preserving candidate-response-informed selection.

6 Conclusion

We presented CAEIR, a cross-architecture deepfake detector that uses interactions among CNN, Transformer, and Mamba response summaries to calibrate input-conditioned Top-1 routing. Cross-dataset results and controlled ablations support combining heterogeneous experts with response-conditioned routing. The Detector-to-LLM interface qualitatively illustrates how visual evidence and routing traces can be verbalized after prediction.

AI use statement. Generative AI tools assisted with language editing, LaTeX conversion, and format checks. The authors verified claims, equations, results, citations, and figures against project files and source papers and take responsibility for all manuscript content.

Reproducibility statement. The release will provide manifests, configurations, seeds, evaluation scripts, and table-generation code.

Ethics statement. This work uses public face-manipulation datasets under their respective licenses and does not release identities or new face data. Deployment should account for distribution shift, false positives, and the consequences of automated decisions.

REFERENCES

Ahmed Abul Hasanaath, Hamzah Luqman, Raed Katib, and Saeed Anwar. FSBI: Deepfake detection with frequency enhanced self-blended images. *Image and Vision Computing*, 154:105418, 2025. doi: 10.1016/j.imavis.2025.105418.

Liang Chen, Yong Zhang, Yibing Song, Lingqiao Liu, and Jue Wang. Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Siyi Chen, Taiping Yao, Yandan Chen, Shouhong Ding, Jilin Li, and Rongrong Ji. Local relation learning for face forgery detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 1081–1088, 2021.

Damai Dai, Chengqi Deng, Chenggang Zhao, R.X. Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Y. Wu, Zhenda Xie, Y.K. Li, Panpan Huang, Fuli Luo, Chong Ruan, Zhifang Sui, and Wenfeng Liang. DeepSeekMoE: Towards ultimate expert specialization in mixture-of-experts language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1280–1297, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.70. URL https://aclanthology.org/2024.acl-long.70/.

Zihang Dai, Hanxiao Liu, Quoc V. Le, and Mingxing Tan. CoAtNet: Marrying convolution and attention for all data sizes. In *Advances in Neural Information Processing Systems*, 2021.

Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The DeepFake detection challenge (DFDC) dataset. *arXiv preprint arXiv:2006.07397*, 2020.

Shichao Dong, Jin Wang, Renhe Ji, Jiajun Liang, Haoqiang Fan, and Zheng Ge. Implicit identity leakage: The stumbling block to improving deepfake detection generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3994–4004, 2023.

Nick Dufour and Andrew Gully. Contributing data to deepfake detection research. Google Research Blog, 2019. URL https://research.google/blog/contributing-data-to-deepfake-detection-research/.

Anurag Dutta, Ruchira Naskar, and Rajat Subhra Chakraborty. TRIMOE: Three-domain mixture-of-experts framework for deepfake identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2026.

Ming Fang, Zhen Li, Lian Yu, Qi Yang, Hongtao Xie, and Yutao Zhang. Forensic-MoE: Exploring comprehensive synthetic image detection traces with mixture of experts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17772–17782, 2025.

Hongyan Fei, Zexi Jia, Chuanwei Huang, Jinchao Zhang, and Jie Zhou. Exploring specular reflection inconsistency for generalizable face forgery detection, 2026. URL https://arxiv.org/abs/2602.06452.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *Conference on Language Modeling*, 2024.

Xiao Guo, Xiufeng Song, Yue Zhang, Xiaohong Liu, and Xiaoming Liu. Rethinking vision-language model in face forensics: Multi-modal interpretable forged face detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 105–116, 2025a. doi: 10.1109/CVPR52734.2025.00019.

Zonghui Guo, Yingjie Liu, Jie Zhang, Haiyong Zheng, and Shiguang Shan. Face forgery video detection via temporal forgery cue unraveling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025b.

Yue-Hua Han, Tai-Ming Huang, Kai-Lung Hua, and Jun-Cheng Chen. Towards more general video-based deepfake detection through facial component guided adaptation for foundation model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22995–23005, 2025.

Xinan He, Yue Zhou, Bing Fan, Bin Li, Guopu Zhu, and Feng Ding. VLForgery face triad: Detection, localization and attribution via multimodal large language models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 180893–180927, 2025. doi: 10.52202/085713-5442.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Jiuyao Jing, Feiyang Sun, Chunlei Peng, Decheng Liu, Nannan Wang, and Xinbo Gao. Adaptive consensus multi-teacher distillation for generalizable face forgery detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 36(7):9336–9348, 2026a. doi: 10.1109/TCSVT.2026.3668432.

Jiuyao Jing, Yu Zheng, and Chunlei Peng. ResProto-FD: Visual-language residual prototype sets for generalized face forgery detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 5548–5556, 2026b.

Chenqi Kong, Anwei Luo, Peijun Bao, Yi Yu, Haoliang Li, Zengwei Zheng, Shiqi Wang, and Alex C. Kot. MoE-FFD: Mixture of experts for generalized and parameter-efficient face forgery detection. *IEEE Transactions on Dependable and Secure Computing*, 23(1):82–96, 2026. doi: 10.1109/TDSC.2025.3604443.

Nicolas Larue, Nhu-Son Vu, Vitomir Struc, Peter Peer, and Vassilis Christophides. SeeABLE: Soft discrepancies and bounded contrastive learning for exposing deepfakes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21011–21021, 2023.

Shufan Li, Harkanwar Singh, and Aditya Grover. Mamba-ND: Selective state space modeling for multi-dimensional data. In *European Conference on Computer Vision*, 2024.

Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-DF: A large-scale challenging dataset for DeepFake forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3207–3216, 2020.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, 2023.

Tianlin Liu, Mathieu Blondel, Carlos Riquelme, and Joan Puigcerver. Routers in vision mixture of experts: An empirical study. *Transactions on Machine Learning Research*, 2024a.

Yue Liu, Yunjie Tian, Yuzhong Zhao, et al. VMamba: Visual state space model. In *Advances in Neural Information Processing Systems*, 2024b.

Anwei Luo, Chenqi Kong, Jiwu Huang, Yongjian Hu, Xiangui Kang, and Alex C. Kot. Beyond the prior forgery knowledge: Mining critical clues for general face forgery detection. *IEEE Transactions on Information Forensics and Security*, 19:1168–1182, 2024. doi: 10.1109/TIFS.2023.3332218.

Yuhan Luo, Tao Chen, and Decheng Liu. Rethinking cross-domain evaluation for face forgery detection with semantic fine-grained alignment and mixture-of-experts. *arXiv preprint arXiv:2604.21478*, 2026.

Changtao Miao, Qi Chu, Tao Gong, Zhentao Tan, Zhenchao Jin, Wanyi Zhuang, Man Luo, Honggang Hu, and Nenghai Yu. Mixture-of-noises enhanced forgery-aware predictor for multi-face manipulation detection and localization. In *Proceedings of the ACM International Conference on Multimedia*, 2025.

Dat Nguyen, Nesryne Mejri, Inder Pal Singh, Polina Kuleshova, Marcella Astrid, Anis Kacem, Enjie Ghorbel, and Djamila Aouada. LAA-Net: Localized artifact attention network for quality-agnostic and generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Dat Nguyen, Marcella Astrid, Anis Kacem, Enjie Ghorbel, and Djamila Aouada. Vulnerability-aware spatio-temporal learning for generalizable deepfake video detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10786–10796, 2025. doi: 10.1109/ICCV51701.2025.01004.

Yunsheng Ni, Depu Meng, Changqian Yu, Chengbin Quan, Dongchun Ren, and Youjian Zhao. CORE: Consistent representation learning for face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 12–21, 2022.

Nabil Omi, Siddhartha Sen, and Ali Farhadi. Load balancing mixture of experts with similarity preserving routers, 2025. URL https://arxiv.org/abs/2506.14038.

Zhengzhuo Pan, Bohan Chen, Longxiang Ma, Dawei Jin, Yu Zhou, and Yudi Huang. WAFF: A synergetic face forgery video detection method via weakly supervised EfficientNet. *Journal of Imaging*, 12(6):240, 2026. doi: 10.3390/jimaging12060240.

Mukesh Pandey, Deepika Koundal, and Sanjeev Kumar. LGF-Net: A spatial-frequency deepfake detection network with learnable gabor filters and gated cross-modal fusion. *Discover Computing*, 29(1), 2026. doi: 10.1007/s10791-026-10548-5.

Siran Peng, Zipei Wang, Li Gao, Xiangyu Zhu, Tianshuo Zhang, Ajian Liu, Haoyuan Zhang, and Zhen Lei. MLLM-enhanced face forgery detection: A vision-language fusion solution. *arXiv preprint arXiv:2505.02013*, 2025.

Zhiliang Peng, Wei Huang, Shanzhi Gu, Lingxi Xie, Yaowei Wang, Jianbin Jiao, and Qixiang Ye. Conformer: Local features coupling global representations for visual recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 367–376, 2021.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. FaceForensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017.

Kaede Shiohara and Toshihiko Yamasaki. Detecting deepfakes with self-blended images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18720–18729, 2022.

Jiamu Sun, Zhiyuan Yan, Ke-Yue Zhang, Taiping Yao, and Shouhong Ding. DFD-HR: Generalizable deepfake detection via hierarchical routing learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13984–13995, 2026.

Ke Sun, Taiping Yao, Shen Chen, Shouhong Ding, Jilin Li, and Rongrong Ji. Dual contrastive learning for general face forgery detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 2316–2324, 2022.

Hao Tan, Jun Lan, Zichang Tan, Senyuan Shi, Ajian Liu, Chuanbiao Song, Huijia Zhu, Weiqiang Wang, Jun Wan, and Zhen Lei. Veritas: Generalizable deepfake detection via pattern-aware reasoning. In *International Conference on Learning Representations*, 2026.

Aryan Thakre, Omkar Nagwekar, Vedang Talekar, and Aparna Santra Biswas. CAST: Cross-attentive spatio-temporal feature fusion for deepfake detection. *Knowledge-Based Systems*, 338: 115560, 2026. doi: 10.1016/j.knosys.2026.115560.

Lean Wang, Huazuo Gao, Chenggang Zhao, Xu Sun, and Damai Dai. Auxiliary-loss-free load balancing strategy for mixture-of-experts, 2024. URL https://arxiv.org/abs/2408.15664.

Qiang Wang and Ru Liu. A new heterogeneous mixture of experts model for deepfake detection. In *Computer Vision and Machine Intelligence*, pp. 428–449, 2025.

Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, and Houqiang Li. AltFreezing for more general video face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Junhao Xu, Jingjing Chen, Yang Jiao, Jiacheng Zhang, Zhiyu Tan, Hao Li, and Yu-Gang Jiang. Identity-aware vision-language model for explainable face forgery detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 11278–11286, 2026. doi: 10.1609/aaai.v40i13.38108.

Zhiyuan Yan, Yong Zhang, Yanbo Fan, and Baoyuan Wu. UCF: Uncovering common features for generalizable deepfake detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22412–22423, 2023a.

Zhiyuan Yan, Yuhao Luo, Siwei Lyu, Qingshan Liu, and Baoyuan Wu. Transcending forgery specificity with latent space augmentation for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Zhiyuan Yan et al. DeepfakeBench: A comprehensive benchmark of deepfake detection. In *Advances in Neural Information Processing Systems*, 2023b.

Kangran Zhao, Yupeng Chen, Xiaoyu Zhang, Yize Chen, Weinan Guan, Baicheng Chen, Chengzhe Sun, Soumyya Kanti Datta, Qingshan Liu, Siwei Lyu, and Baoyuan Wu. DeepfakeBench-MM: A comprehensive benchmark for multimodal deepfake detection. *arXiv preprint arXiv:2510.22622*, 2025.

Yinglin Zheng, Jianmin Bao, Dong Chen, Ming Zeng, and Fang Wen. Exploring temporal coherence for more general video face forgery detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

Tianfei Zhou, Wenguan Wang, Zhiyuan Liang, and Jianbing Shen. Face forensics in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5778–5788, 2021.

Wanyi Zhuang, Qi Chu, Zhentao Tan, Qiankun Liu, Haojie Yuan, Changtao Miao, Zixiang Luo, and Nenghai Yu. UIA-ViT: Unsupervised inconsistency-aware method based on vision transformer for face forgery detection. In *European Conference on Computer Vision*, 2022.

Barret Zoph, Irwan Bello, Sameer Kumar, Nan Du, Yanping Huang, Jeff Dean, Noam Shazeer, and William Fedus. ST-MoE: Designing stable and transferable sparse expert models, 2022. URL https://arxiv.org/abs/2202.08906.

APPENDIX

A ANALYSIS OF CROSS-ARCHITECTURE EXPERT ROUTING

Analysis setup. We further examine how the heterogeneous expert pool is utilized across different fake sources. The analysis uses the official HydraFake test data released with Veritas (Tan et al., 2026) and the four experts in CAEIR: CNN, Transformer, Mamba-H, and Mamba-V. HydraFake is used only for this post-hoc analysis and does not participate in training, validation, checkpoint selection, or the cross-dataset AUC results. We first analyze routing probabilities at the original source or generator level and then aggregate the sources into six manipulation processes: entire-face generation, face swapping, face reenactment, face editing, face relighting, and face restoration. We additionally compare routed expert responses between fake and real samples to characterize how the router emphasizes different experts across sources.

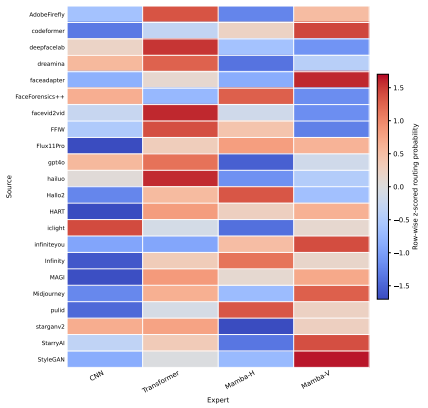


Figure A1: Expert routing across HydraFake sources. Rows follow HydraFake source labels (Tan et al., 2026); colors show row-wise z-scores of mean routing probabilities.

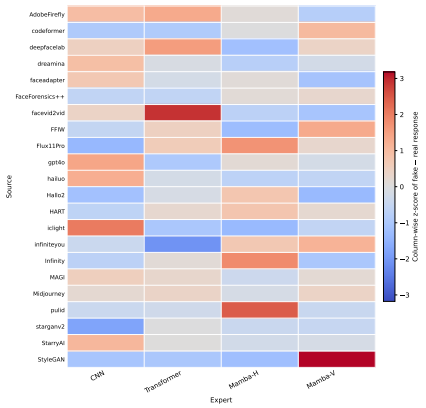


Figure A2: Routed expert-response contrast across HydraFake sources. Colors show column-wise z-scores of the fake–real difference.

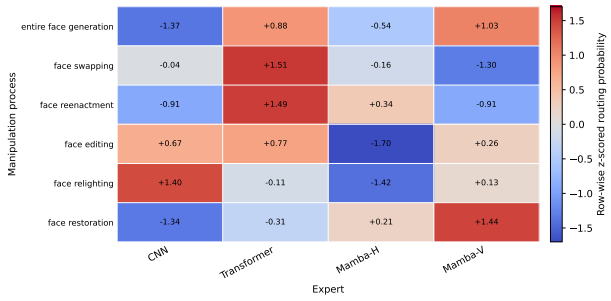


Figure A3: Expert routing by manipulation process. Aggregated routing probabilities are standardized across the four experts within each process (row-wise z-scores).

Routing statistics. For each evaluated frame, routing probabilities are first averaged over the nine MoE layers. We then average these probabilities over all evaluated frames from the same source to obtain frame-weighted source means. Process-level statistics average the same layer-averaged probabilities over all evaluated frames belonging to each manipulation process, so sources contribute in proportion to their frame counts. For visualization, raw source and process means are standardized across the four experts within each row (row-wise z-scores).

Routed-response statistics. For each expert response, we compute its mean absolute feature magnitude over spatial tokens and channels. We weight this magnitude by its routing probability when the expert is selected by Top-1 routing and assign zero to unselected experts. After averaging over the nine MoE layers and evaluated frames, routed-response contrast is the fake mean minus the real mean for each source and expert. For each fake source, the real mean is computed from the corresponding protocol-matched real subset. Contrasts are standardized independently for each expert across sources (column-wise z-scores).

Source-level routing. Figure A1 shows that routing preferences vary across fake sources. FFIW, FaceForensics++, and several conventional manipulation sources show stronger relative preference for CNN or Transformer, whereas several recent generation sources produce stronger preferences for the Mamba experts. This behavior does not reduce to a simple conventional-versus-generative partition: sources such as Dreamina and GPT-4o still exhibit strong CNN preference. The learned router therefore displays source-dependent patterns across the heterogeneous expert pool.

Routed-response profiles. Figure A2 complements the routing-probability analysis with selection-weighted fake–real response contrasts. CNN shows comparatively strong routed-response contrasts for facevid2vid and CodeFormer; Transformer shows strong contrasts for CodeFormer and several conventional manipulation sources. Mamba-H shows relatively strong contrasts for sources including Flux11Pro, Adobe Firefly, and PuLID. Mamba-V shows relatively strong contrasts for Adobe Firefly, StarGAN v2, and GPT-4o. These patterns show that the responses emphasized by the routing mechanism vary across both experts and fake sources.

Manipulation-level routing. We finally aggregate frame-level routing profiles by manipulation process to examine whether the observed differences persist at a coarser semantic level (Figure A3). Entire-face generation shows the strongest relative preference for Mamba-V (1.03), while face swapping favors Transformer (1.51). Face reenactment also favors Transformer (1.49). Face relighting shows the strongest preference for CNN (1.40), whereas face restoration favors Mamba-V (1.44). Face editing favors Transformer (0.77) and CNN (0.67) over Mamba-H (-1.70). These patterns are consistent with the complementary information-propagation mechanisms represented by convolution, self-attention, and directional state-space modeling.

Overall observation. The three views provide complementary evidence about the behavior of the heterogeneous expert pool. Routing varies across individual fake sources, routed fake–real response contrasts differ across experts, and related routing structure remains visible after aggregation by manipulation process. Together, these observations suggest that CAEIR learns structured, sample-dependent use of cross-architecture experts rather than collapsing to a uniform routing pattern.

B Training Objective and Evaluation Details

Load balancing. For a batch of B samples and $E = 4$ experts, let $p^l_{b,i}$ denote the routing probability of expert i for sample b at layer l. We compute the mean soft routing probability and the hard Top-1 selection frequency as

$$P^l_i = \frac{1}{B}\sum_{b=1}^{B} p^l_{b,i}, \qquad f^l_i = \frac{1}{B}\sum_{b=1}^{B} \mathbf{1}\left[i = \arg\max_j p^l_{b,j}\right]. \tag{6}$$

The load-balancing term averages the per-layer losses over the nine augmented blocks $\mathcal{M} = \{4, \ldots, 12\}$:

$$\mathcal{L}_{\text{lb}} = \frac{1}{|\mathcal{M}|}\sum_{l\in\mathcal{M}} E\sum_{i=1}^{E} f^l_i P^l_i. \tag{7}$$

Routing-ablation configurations. Unless otherwise specified, all routing baselines use the same expert pool, router backbone, training schedule, and evaluation protocol as the main model and differ only in the routing configuration. **Input-conditioned Top-1** uses the CLS-based router with $k = 1$ and no candidate-response interaction. **Orthogonality-regularized**, inspired by similarity-preserving router regularization (Omi et al., 2025), retains Top-1 routing and adds an orthogonality

regularizer to the router weights, encouraging decorrelated routing directions. **Top-2 cooperative** is our Top-k routing ablation, motivated by sparse multi-expert routing (Zoph et al., 2022); it dispatches each input to the two experts with the highest routing probabilities and aggregates their outputs using normalized Top-2 scores. **Load-biased**, adapted from auxiliary-loss-free load balancing (Wang et al., 2024), removes the gradient-based load-balancing loss and instead adjusts expert-selection preferences using a loss-free routing bias derived from expert utilization. **Shared-CNN**, motivated by the shared-expert design in DeepSeekMoE (Dai et al., 2024), retains the routed heterogeneous expert pool and additionally introduces a shared CNN expert whose output is added to every routed output with a scaling factor of 0.1. **Soft expert mixture** is our dense-mixture baseline; it replaces sparse Top-1 dispatch with dense probability-weighted fusion of all candidate expert outputs. **Conflict-entropy calibration** retains Top-1 dispatch and adds a conflict-aware routing objective based on disagreement among candidate expert responses, encouraging lower routing entropy when inter-expert conflict is stronger.

Face preprocessing and external evaluation. The detector uses ImageNet-pretrained `vit_small_patch16_224` weights from timm. For video-level external evaluation, we uniformly sample 32 temporal positions over the full duration of each video and detect faces using RetinaFace with the `resnet50_2020-07-20` model. RetinaFace is configured with a maximum image size of 2048, a confidence threshold of 0.7, and an NMS threshold of 0.4. At each sampled position, we retain detected faces whose bounding-box area is at least 50% of that of the largest detected face in the frame. Each retained bounding box is expanded by one eighth of its width and height on each side and clipped to the image boundaries. We do not apply landmark-based face alignment. The resulting crop is first resized to 380×380, after which the CLIP image processor resizes and center-crops it to 336×336 and applies CLIP normalization. Video-level scores average the highest 20% of fake probabilities among all retained crops from the video.

C Additional Detector-to-LLM Examples

To complement the four cases shown in Figure 3, we provide eight additional qualitative Detector-to-LLM examples in Figures A4 and A5. These cases show correct predictions and classification errors together with their route profiles, visual cues, and generated explanations.

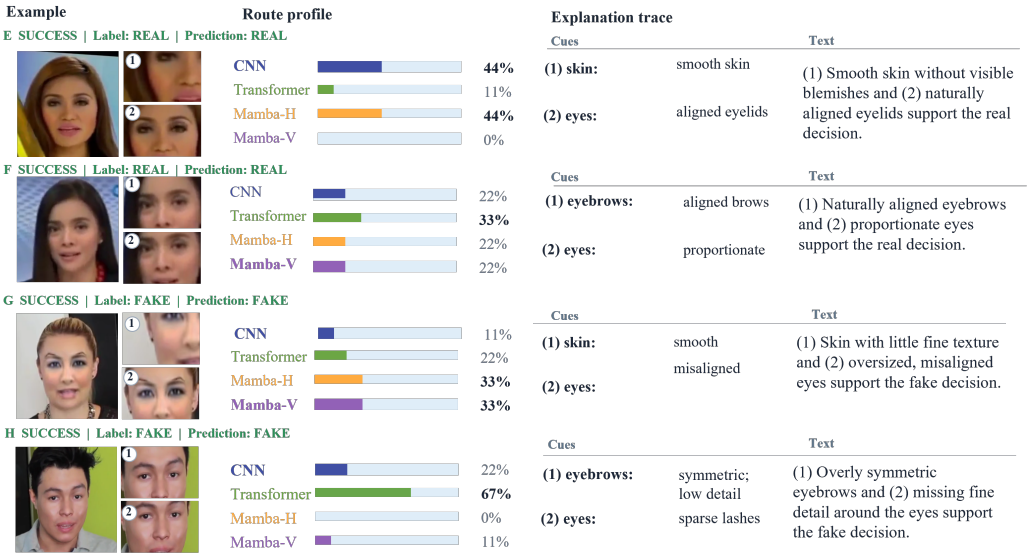


Figure A4: Additional qualitative Detector-to-LLM examples (E–H). Each row shows the input face and selected crops, the route profile, visual cues, and generated explanation.

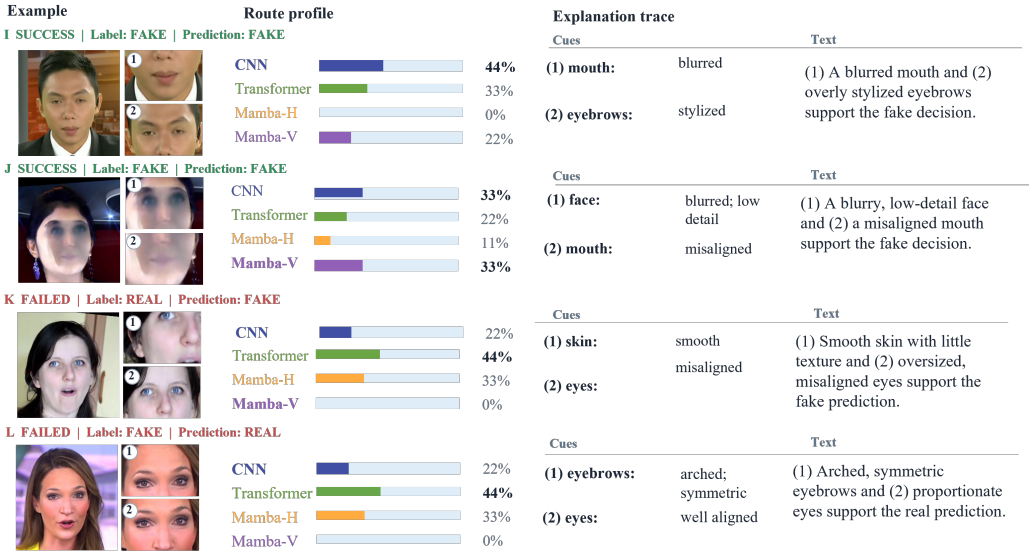


Figure A5: Additional qualitative Detector-to-LLM examples (I–L). Each row shows the input face and selected crops, the route profile, visual cues, and generated explanation.