PARALLEL HYPOTHESIS DECISIONS FOR VISION-LANGUAGE DEEPFAKE FACE DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models (VLMs) are increasingly asked to judge whether a face image is real or forged, and the default recipe is to let them reason step by step before answering. We argue that authenticity judgment is a perceptual decision that is poorly served by slow, generative reasoning. With the same 4-billion-parameter VLM, the same forensic evidence, and the same images, reasoning for about 998 tokens before answering is not only slow but also inaccurate: on all six external benchmarks it scores below the forensic evidence it was given. We introduce PDM-Face, a detector that answers a panel of eight forensic hypotheses (diffusion, GAN, face swap, real photograph, abnormal spectral energy, up-sampling peaks, and two expert-evidence checks) in parallel and without generating a single token. The image and a compact set of evidence tokens from a gated forensic expert layer are encoded once; the key-value and recurrent states of this shared prefix are then copied into eight branches that are answered in one batched forward pass, and a learned fusion turns the eight three-way answers into a calibrated fake probability. Across six external face benchmarks, PDM-Face outperforms generative reasoning in all 12 comparisons (paired 95% confidence lower bound at least +0.051 AUROC) while being 21 times faster per image. It also significantly outperforms four recent detectors in 19 of 23 benchmark comparisons, improves over its own expert layer on the hardest video benchmarks (+0.034 on FaceForensics++ and +0.021 on DFDC, full test sets), and its parallel hypotheses beat a single-question decision layer on GenImage and 3rdBench for every seed and for a second VLM backbone.

1 INTRODUCTION

Face forgeries are now produced by a wide range of generators: text-to-image diffusion models, GANs, face-swapping pipelines, and talking-head reenactment. A face-forgery detector deployed for identity verification or content moderation must therefore generalize to generators it has not seen during training (Wang et al., 2020; Yan et al., 2023b; 2024b). Specialized forensic detectors capture strong low-level cues, such as blending boundaries (Shiohara & Yamasaki, 2022), spectral fingerprints (Frank et al., 2020; Dzanic et al., 2020), and up-sampling residuals (Tan et al., 2024). Foundation-model detectors instead read semantic features of frozen vision-language backbones (Ojha et al., 2023; Liu et al., 2024; Koutlis & Papadopoulos, 2024). Both families remain brittle once the manipulation type changes.

Multimodal large language models appear to offer a way out. They see the image, can be told what the forensic tools report, and can explain their verdict (Jia et al., 2024; Xu et al., 2025; Huang et al., 2025). The prevailing recipe asks them to reason in natural language before answering, following the success of chain-of-thought prompting on mathematical and symbolic tasks (Wei et al., 2022; Kojima et al., 2022). This paper asks whether that recipe is appropriate for authenticity judgment.

Our answer is no, and the evidence is direct. We give Qwen3.5-4B a face image together with the scores of a strong forensic expert layer and let it think in its native reasoning mode before answering. It generates about 998 tokens per image. Its accuracy is then lower than that of the expert scores it was handed, on every one of six external benchmarks (Figure 2). On Celeb-DF, for instance, the expert layer alone reaches 0.834 AUROC and the reasoning model given those scores reaches 0.688. Deliberate verbal reasoning dilutes precise numerical evidence into text, and the decision becomes

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.

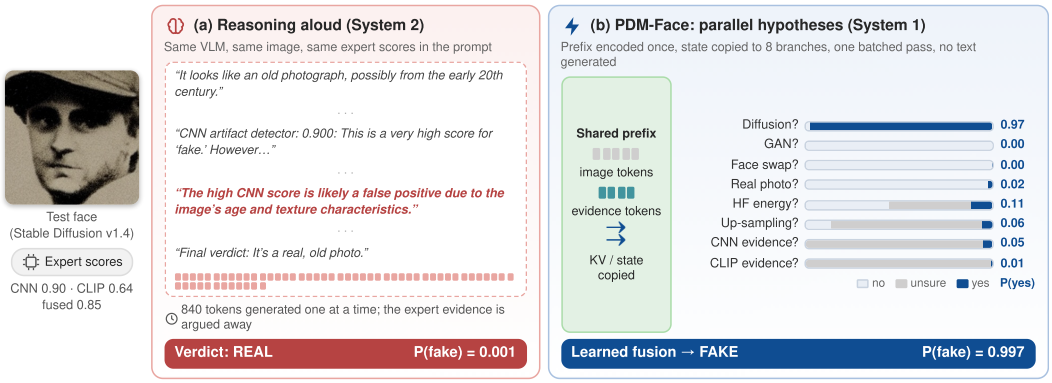

Figure 1: Two ways to let a VLM judge the same forged face (Stable Diffusion v1.4, GenImage). (a) Given the image and the expert scores, the VLM reasons for 840 tokens and argues the confident CNN evidence away (quotes are verbatim excerpts of its trace). (b) PDM-Face encodes the image and evidence tokens once, copies the prefix state into eight hypothesis branches, and reads a no, unsure, or yes answer from each branch in one batched forward pass without generating text; the bars show its answers for this face.

worse. This matches a growing body of evidence that chain-of-thought helps mainly on symbolic problems and can hurt on tasks that people solve by fast perception (Sprague et al., 2025; Liu et al., 2025). In the terms of dual-process theory (Kahneman, 2011), detecting a forged face is a System 1 task.

We therefore design a VLM detector that decides fast and in parallel instead of reasoning aloud. PDM-Face (Parallel Decision-Making for faces) poses a fixed panel of eight forensic hypotheses to the VLM at once (Figure 1). Four hypotheses ask about the generation mechanism: diffusion, GAN, face swap, and real photograph. Two ask about spectral signatures: abnormal high-frequency energy and periodic up-sampling peaks. Two ask whether the forensic CNN and CLIP evidence point to synthesis. The face image and four soft evidence tokens, computed from a gated forensic expert layer, form a shared prefix that is encoded once. Its key-value cache and recurrent states are copied into eight branches, and each branch appends one hypothesis question. All eight branches are answered in a single batched forward pass, and each answer is read from a three-way head (no, uncertain, yes) without generating text. A learned fusion maps the eight answer distributions and the expert scores to a calibrated fake probability. Hypothesis labels are free: they come from dataset metadata and from spectral statistics computed on the training reals, so no textual rationale or teacher model is required.

Contributions.

- We show, on identical images and with identical evidence, that generative reasoning is the wrong decision mode for VLM face-forgery detection: it scores below its own input evidence on all six external benchmarks and costs 2.89 seconds per image (Section 5.1).
- We introduce PDM-Face, a parallel hypothesis decision layer that turns a VLM into a System 1 detector. Branching the prefix state instead of recomputing it is exact up to numerical precision and 11.9 times faster than answering the eight hypotheses one after another (Section 3).
- We show that the decision layer adds measurable value on top of a strong forensic expert layer: significant gains on the full FaceForensics++ and DFDC test sets, and consistent gains of parallel hypotheses over a single question on GenImage and 3rdBench across seeds and across two VLM backbones (Section 5.2).
- Against four recent detectors evaluated on the same images, PDM-Face is significantly better in 19 of 23 comparisons, with margins up to +0.40 AUROC on Celeb-DF (Section 5.3).

2 RELATED WORK

Generalizable forgery detection. Early detectors exploit up-sampling and spectral artifacts of CNN generators (Wang et al., 2020; Frank et al., 2020; Dzanic et al., 2020), and face-specific methods add frequency-aware streams (Qian et al., 2020; Luo et al., 2021). Self-blended images (Shiohara & Yamasaki, 2022) synthesize blending boundaries from real faces, and feature-level augmentation targets forgery-agnostic cues (Yan et al., 2023a; 2024a). Reconstruction-based methods score how well an autoencoder or diffusion model reproduces an image (Wang et al., 2023; Ricker et al., 2024), and NPR models up-sampling residuals directly (Tan et al., 2024). Foundation-feature detectors probe or adapt frozen CLIP features (Radford et al., 2021; Ojha et al., 2023; Cozzolino et al., 2024; Liu et al., 2024; Koutlis & Papadopoulos, 2024), and recent work constrains the adapted subspace (Yan et al., 2025b; Cui et al., 2025). AIDE combines low-level patch statistics with semantic features (Yan et al., 2025a). Benchmarks for this problem include FaceForensics++ (Rössler et al., 2019), Celeb-DF (Li et al., 2020), DFDC (Dolhansky et al., 2020), WildDeepfake (Zi et al., 2020), GenImage (Zhu et al., 2023), DeepfakeBench (Yan et al., 2023b), and DF40 (Yan et al., 2024b). Our work is complementary: we keep a forensic expert layer and change how a VLM turns its evidence into a decision.

Multimodal language models for forensics. General VLMs (Liu et al., 2023; Qwen Team, 2025) can be prompted for authenticity with modest zero-shot accuracy (Jia et al., 2024). FakeShield (Xu et al., 2025) and SIDA (Huang et al., 2025) fine-tune multimodal models to detect, localize, and explain manipulations, producing textual rationales. These systems pay the cost of autoregressive generation for every image. PDM-Face keeps the VLM's visual grounding but makes the decision without generation, and it obtains its supervision from labels that already exist rather than from written explanations.

Reasoning, fast and slow. Chain-of-thought prompting improves multi-step arithmetic and symbolic reasoning (Wei et al., 2022; Kojima et al., 2022), but a meta-analysis finds little benefit outside those domains (Sprague et al., 2025), and deliberate reasoning can reduce accuracy on tasks where verbal thinking also hurts people (Liu et al., 2025). Dual-process theory (Kahneman, 2011) separates fast, automatic judgment from slow, deliberate reasoning. We give the first controlled measurement of this trade-off for VLM face-forgery detection and design the decision layer on the fast side.

Shared-prefix inference and efficient adaptation. Serving systems reuse the key-value cache of a shared prompt across many continuations (Kwon et al., 2023; Zheng et al., 2024; Juravsky et al., 2024). We use the same principle inside the model, as part of the decision rule. Hybrid VLMs that interleave full attention with gated delta-rule layers (Yang et al., 2025) carry a recurrent state as well as a key-value cache; we show how to branch both exactly. We adapt the language model with LoRA (Hu et al., 2022). The expert layer is a small mixture of experts (Jacobs et al., 1991; Shazeer et al., 2017) over a forensic CNN ensemble and a CLIP probe.

3 METHOD

PDM-Face has two stages (Figure 1b). A forensic expert layer produces numerical evidence for each face. A parallel decision layer, built on a VLM, reads the face and the evidence and answers a panel of hypotheses at once.

3.1 FORENSIC EXPERT LAYER

The expert layer combines two frozen detectors with complementary strengths. The first is an ensemble of $K = 5$ forensic CNNs, each a four-branch network with an EfficientNetV2-S backbone (Tan & Le, 2021) that fuses RGB, spectral, residual, and self-blending streams (Shiohara & Yamasaki, 2022). With horizontal-flip test-time augmentation and a per-model temperature T_k, it gives

$$p_{\text{cnn}}(x) = \frac{1}{2K} \sum_{v \in \{x, \text{flip}(x)\}} \sum_{k=1}^{K} \text{softmax}\big(f_k(v)/T_k\big)_{\text{fake}}. \tag{1}$$

The second is a logistic probe on the pooled CLIP ViT-L/14 image feature $\phi(x) \in \mathbb{R}^{1024}$ (Radford et al., 2021; Ojha et al., 2023), $p_{\text{clip}}(x) = \sigma\big(\boldsymbol{w}^\top(\phi(x) - \boldsymbol{\mu})/\boldsymbol{s} + b\big)$. A small gate reads the same CLIP feature and decides, per image, how much to trust each detector (Jacobs et al., 1991):

$$g(x) = c + (1-c)\,\sigma\big(\mathrm{MLP}(\phi(x))\big), \qquad p_{\text{exp}}(x) = g(x)\,p_{\text{cnn}}(x) + \big(1 - g(x)\big)\,p_{\text{clip}}(x), \tag{2}$$

with anchor $c = 0.5$ so that the forensic CNN always keeps at least half the weight. We use p_{exp} as the "expert layer only" reference throughout the paper.

Evidence vector. For each face the expert layer also exposes its internal evidence: the mean penultimate CNN feature (256 dimensions), the CLIP feature (1,024), a 30-bin radial log-power spectrum, and five scalars (p_{cnn}, p_{clip}, g, a high-frequency energy ratio h, and a spectral peak score q). Here h is the mean log power in the high band minus that in the middle band, and q is the largest second difference of the radial spectrum in the high band, which responds to periodic up-sampling artifacts. The 1,315-dimensional vector $\boldsymbol{e}(x)$ is computed once per image and cached.

3.2 Parallel decision layer

Evidence tokens. A two-layer adapter A maps $\boldsymbol{e}(x)$ to $k = 4$ soft tokens in the language model's embedding space, $[\boldsymbol{t}_1, \ldots, \boldsymbol{t}_k] = A(\boldsymbol{e}(x)) \in \mathbb{R}^{k \times d}$. The VLM prompt is a shared prefix

$$P(x) = \big[\,\text{header},\ \text{image tokens of } x,\ \text{"Forensic evidence:"},\ \boldsymbol{t}_1, \ldots, \boldsymbol{t}_k,\ \text{instruction}\,\big], \tag{3}$$

where the instruction reads "Answer each question about the face above."

Hypothesis panel. We pose $M = 8$ yes-or-no questions $q_1, \ldots, q_M$ (the exact wording is in Appendix C): whether the face was produced by a diffusion model (H_1), by a GAN (H_2), by face swapping (H_3), whether it is an unedited photograph (H_4), whether its high-frequency energy is abnormal (H_5), whether its spectrum shows up-sampling peaks (H_6), and whether the CNN (H_7) and CLIP (H_8) evidence indicates synthesis. Each branch is $[P(x), q_m + \boldsymbol{u}_m, \text{"Answer:"}]$, where $\boldsymbol{u}_m$ is a learned embedding added to the question tokens to separate the branches. A shared head reads the hidden state $\boldsymbol{h}_m$ at the final token and outputs a distribution over {no, uncertain, yes}, $\pi_m = \mathrm{softmax}(W\,\mathrm{LN}(\boldsymbol{h}_m)) \in \Delta^3$.

Branching the prefix state. Answering the hypotheses one after another would recompute the prefix, which contains all image tokens, M times. Instead we run the prefix once with a cache and copy its state M times along the batch dimension. The eight question tails are then processed as one batch of continuations. Qwen3.5 (Yang et al., 2025) interleaves full-attention layers with gated delta-rule layers, so its prefix state has two parts: the key-value cache of the attention layers, and, for each delta-rule layer, the convolution buffer and the recurrent state. We copy both. We also resume the short causal convolution from the cached buffer for multi-token continuations. Without this step the continuation would silently restart from a zero state, and packing branches into one sequence with a block mask would leak information across branches through the recurrence. The branched computation is exact: in float32 the three-way probabilities of branched, fully recomputed, and one-hypothesis-at-a-time forward passes agree to within 2.7×10^{-4} (Appendix D). In bfloat16 with the flash-linear-attention kernels, one image costs 11.7 ms with branching, 25.6 ms with full recomputation, and 140.1 ms when the eight questions are asked sequentially, a 11.9-fold speedup at batch size 16.

Fusion. A small multilayer perceptron maps the concatenated answer distributions and the three expert scalars to a fake logit, $z(x) = F\big([\pi_1, \ldots, \pi_M,\ p_{\text{cnn}},\ p_{\text{clip}},\ g]\big)$, and the output probability is $\sigma(z/\tau)$ with a temperature τ fitted on the validation set (Guo et al., 2017).

3.3 Supervision without rationales

Every hypothesis is supervised by labels that already exist. H_1 to H_4 follow from each training image's generator family: diffusion, GAN, face swap, or real. H_5 to H_8 use statistics computed on the training reals only, so no test statistic is involved. For H_6 to H_8, a value above the 90th percentile of the training distribution is "yes", below the 10th percentile is "no", and anything in between is "uncertain". H_5 is "no" inside the central 30th to 70th percentile band of real-image high-frequency

energy, "yes" outside the 10th to 90th band, and "uncertain" otherwise. The uncertain class lets the model abstain on ambiguous evidence instead of forcing a guess. The training loss is

$$\mathcal{L} = \mathrm{BCE}\big(z(x), y\big) + \lambda \frac{1}{M} \sum_{m=1}^{M} \mathrm{CE}\big(\pi_m,\, \ell_m(x)\big), \qquad \lambda = 0.5. \tag{4}$$

The expert layer is extremely confident on its own training distribution, with an in-domain AUROC of 0.998. A model that sees its scores would therefore learn to copy them. To prevent this, we zero the evidence vector and the expert scalars for each training image with probability 0.5, which forces the VLM to use the image as well as the evidence.

4 Experimental setup

Data. All learned components train on a quality-controlled corpus of aligned 256×256 face crops covering real photographs and diffusion, GAN, face-swap, and reenactment forgeries. Real and fake faces pass identical quality gates on detection confidence, native resolution, sharpness, pose, and CLIP faceness, so a detector cannot separate them by image quality. Splits are group-aware, so the same identity never appears in two splits. The decision layer trains on a balanced subset of 60,000 faces with fakes spread evenly over generator families, and it is validated on 6,000 faces. We test in-domain on 12,000 held-out faces and on six external benchmarks from sources that are disjoint from training. The benchmarks are Celeb-DF (Li et al., 2020) (1,109 faces), 3rdBench (33,085 faces from four third-party sources: StyleGAN faces, two face-swap collections, and a text-to-image diffuser), GenImage faces (Zhu et al., 2023) (20,157), DFDC (Dolhansky et al., 2020) (30,460), FaceForensics++ (Rössler et al., 2019) (21,144), and WildDeepfake (Zi et al., 2020) (7,746). Each face is cropped with the same detector and crop rule as the training data.

Models and training. The default backbone is Qwen3.5-4B, a hybrid attention and gated delta-rule VLM. We apply LoRA (Hu et al., 2022) with rank 16 to all attention, delta-rule, and MLP projections of the language model. Together with the adapter, question embeddings, answer head, and fusion, this gives 42.4 million trainable parameters. The vision encoder and the expert layer stay frozen. Images are resized to 224×224. We train for 1,250 steps with batch size 16, AdamW, learning rate 10^{-4} for LoRA and 10^{-3} for the new modules, a 3% warm-up, and cosine decay. We report three seeds for the main models.

Baselines. We compare against (i) the forensic expert layer alone; (ii) the same VLM zero-shot, and directly fine-tuned with one authenticity question and no evidence; (iii) a single-query decision layer, identical to PDM-Face in backbone, evidence tokens, evidence dropout, data, and schedule, but asking only "Is this face AI-generated or manipulated?"; (iv) the same VLM reasoning in thinking mode (Section 5.1); and (v) four recent detectors: NPR (Tan et al., 2024) and AIDE (Yan et al., 2025a), retrained on our face corpus, and FatFormer (Liu et al., 2024) and RINE (Koutlis & Papadopoulos, 2024), with their released weights. Every method is scored on the same image files, matched by path.

Metrics. We report AUROC. Differences between two methods are tested with a paired, class-stratified bootstrap over images (1,000 resamples) (Efron & Tibshirani, 1993), and we call a difference significant when its 95% interval excludes zero. For the large external sets we score a class-stratified sample of 4,000 faces with a fixed seed. Celeb-DF is scored in full. On the full DFDC, FaceForensics++, and WildDeepfake test sets, AUROC differs from the 4,000-face estimate by at most 0.007 (Appendix G).

5 Results

5.1 Reasoning aloud is the wrong decision mode

We let Qwen3.5-4B reason in its native thinking mode for up to 1,536 tokens under a forensic-analyst instruction. The reasoning is then followed by "Answer:", and we read the probability of "Yes" against "No" as the score. We test two variants. In the first the model sees only the image. In

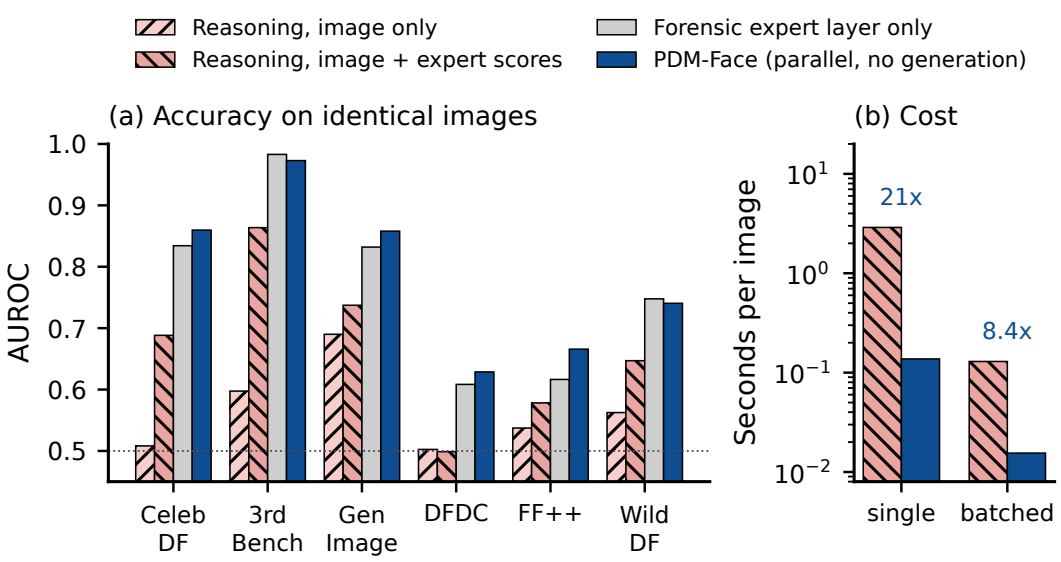

Figure 2: Generative reasoning against parallel hypothesis decisions with the same Qwen3.5-4B backbone. (a) AUROC on identical images, 1,000 per benchmark. Reasoning over the image alone is near chance. Given the expert scores, reasoning still falls below the expert layer it was handed, on all six benchmarks. PDM-Face is above reasoning everywhere. (b) Seconds per image on one H100 GPU, one image at a time and batched.

the second it additionally receives the expert layer's CNN score, CLIP score, and gate weight as text, the same scalars PDM-Face uses. The reasoning runs in vLLM (Kwon et al., 2023) with continuous batching and prefix caching. We evaluate 1,000 class-balanced faces per benchmark, drawn from the images PDM-Face was scored on (949 for Celeb-DF, whose fake class is smaller).

Accuracy. Reasoning over the image alone is close to chance, at 0.502 to 0.690 AUROC (Figure 2a). Adding the expert scores lifts it to 0.499 to 0.864. It still falls below the expert layer it was given on all six benchmarks. For example, 3rdBench drops from 0.983 for the expert layer to 0.864 after reasoning, and Celeb-DF drops from 0.834 to 0.688. PDM-Face reads the same image and the same evidence and reaches 0.860 on Celeb-DF. It is significantly better than reasoning in all 12 comparisons (six benchmarks, two variants), and the smallest lower confidence bound on its margin is +0.051 AUROC (Appendix F). The failure is not a lack of effort: the model writes 998 tokens per image on average before answering.

Cost. Processing one image at a time, reasoning takes 2.89 seconds per face and PDM-Face takes 0.137 seconds, which is 21 times faster (Figure 2b). With batching, the vLLM reasoning server reaches 0.130 seconds per face, while PDM-Face in a plain PyTorch implementation reaches 0.0155 seconds at batch size 16, a 8.4-fold advantage. PDM-Face generates no tokens, so its cost does not grow with the length of an explanation.

5.2 The decision layer adds value on top of the experts

Table 1 compares three ways of turning the same evidence into a decision. Adding a VLM decision layer helps most where the forensic experts are weakest. On the full test sets, PDM-Face improves over the expert layer by +0.034 AUROC on FaceForensics++ (21,144 faces) and by +0.021 on DFDC (30,460 faces), and both paired confidence intervals exclude zero (Appendix G). On the evaluation samples it also improves Celeb-DF by +0.025 (95% interval [+0.014, +0.034]) and GenImage by +0.026 ([+0.020, +0.031]). These are the benchmarks dominated by face swapping, reenactment, and heavy video compression, where visual context helps interpret ambiguous low-level evidence.

Table 1: Decision layer study, Qwen3.5-4B, mean and standard deviation over three seeds (AUROC). The last row is the seed-paired gain of PDM-Face over the single-query decision layer; bold marks gains that are positive for all three seeds. External sets use the fixed 4,000-face samples; Celeb-DF is complete.

Decision layer	In-domain	Celeb-DF	3rdBench	GenImage	DFDC	FF++	WildDeepfake
Expert layer only	0.998	0.840	0.977	0.825	0.622	0.622	0.749
Single query, with evidence	0.997 $\pm$ 0.000	0.855 $\pm$ 0.026	0.948 $\pm$ 0.005	0.801 $\pm$ 0.011	0.613 $\pm$ 0.023	0.650 $\pm$ 0.015	0.724 $\pm$ 0.009
PDM-Face, 8 parallel hypotheses	0.998 $\pm$ 0.000	0.855 $\pm$ 0.021	0.965 $\pm$ 0.001	0.841 $\pm$ 0.012	0.626 $\pm$ 0.009	0.651 $\pm$ 0.011	0.738 $\pm$ 0.013
Paired gain of PDM-Face	+0.000	-0.000	**+0.017**	**+0.040**	**+0.013**	+0.001	+0.014

Table 2: AUROC against recent detectors on identical images. NPR and AIDE are retrained on our face corpus; FatFormer and RINE use their released weights. PDM-Face reports mean and standard deviation over three seeds. Bold marks the best detector per benchmark; the expert-layer row is a component of PDM-Face and is shown for reference. †AIDE on GenImage covers the 454 images in the intersection with its released evaluation. RINE has no GenImage evaluation.

Method	Celeb-DF	3rdBench	GenImage	DFDC	FF++	WildDeepfake
NPR (Tan et al., 2024)	0.462	0.588	0.616	0.465	0.517	0.593
AIDE (Yan et al., 2025a)	0.560	0.670	0.824†	0.537	0.533	0.629
FatFormer (Liu et al., 2024)	0.667	0.687	0.771	0.630	0.796	0.662
RINE (Koutlis & Papadopoulos, 2024)	0.582	0.821	–	**0.671**	**0.808**	0.666
Forensic expert layer only	0.840	0.977	0.825	0.622	0.622	0.749
PDM-Face (ours)	**0.855** $\pm$ 0.021	**0.965** $\pm$ 0.001	**0.841** $\pm$ 0.012	0.626 $\pm$ 0.009	0.651 $\pm$ 0.011	**0.738** $\pm$ 0.013

Parallel hypotheses matter beyond simply adding a VLM. Relative to the single-query decision layer with the same evidence, PDM-Face gains +0.040 AUROC on GenImage, +0.017 on 3rdBench, and +0.013 on DFDC, and each gain is positive for all three seeds. On the remaining benchmarks the two decision layers are statistically tied. The single-query layer answers one holistic question. PDM-Face's panel asks about the generation mechanism and the spectral signature separately, which is most useful on GenImage and 3rdBench because their fakes come from several distinct generator families.

The benefit carries over to another backbone. We repeat the comparison with Qwen3-VL-4B (Qwen Team, 2025), a full-attention VLM with multi-level visual features, using the same recipe. PDM-Face's accuracy changes by at most 0.017 AUROC on any benchmark. Its advantage over the single-query layer reappears on GenImage (+0.024, 95% interval [+0.019, +0.030]) and 3rdBench (+0.006, [+0.003, +0.009]), as detailed in Appendix H. The parallel decision layer therefore does not depend on the linear-attention structure of Qwen3.5.

5.3 Comparison with recent detectors

Table 2 compares PDM-Face with four recent detectors on the same images. PDM-Face is significantly better in 19 of 23 benchmark and method pairs (paired bootstrap, Appendix J). The margins are largest on the face benchmarks, where general AI-image detectors transfer poorly. On Celeb-DF, PDM-Face exceeds FatFormer by +0.197 and NPR by +0.402. On WildDeepfake, whose videos come from the internet, it exceeds RINE by +0.078 and FatFormer by +0.082. On GenImage, the home domain of general AI-image detectors, PDM-Face still leads FatFormer by +0.080. The remaining pairs all involve the two CLIP-based detectors on the video benchmarks: on DFDC PDM-Face is statistically tied with FatFormer and RINE leads by 0.034, and on FaceForensics++ both lead, mostly on face swaps (Appendix I). These are the two benchmarks where our expert layer is weakest (0.622 on each) and where the decision layer adds the most over it (Section 5.2). The evidence PDM-Face uses is complementary to that of CLIP-based detectors. A parameter-free rank average of PDM-Face with either detector significantly exceeds both of its inputs on DFDC (0.686 with RINE and 0.663 with FatFormer) and, with FatFormer, on GenImage (0.899); on FaceForensics++ it raises PDM-Face by +0.127 (Appendix K). Because the expert layer enters only through evidence tokens, such a detector can also be connected as an additional expert.

Table 3: Ablations, Qwen3.5-4B, seed 0 (AUROC). The first block contains VLM baselines without evidence. The second block varies one component of PDM-Face at a time.

Variant	In-domain	Celeb-DF	3rdBench	GenImage	DFDC	FF++	WildDeepfake
Qwen3.5-4B zero-shot, no training	0.499	0.351	0.545	0.494	0.564	0.531	0.500
Direct LoRA fine-tuning, no evidence	0.942	0.496	0.615	0.745	0.528	0.550	0.588
PDM-Face without evidence tokens	0.943	0.487	0.603	0.728	0.539	0.559	0.564
PDM-Face with mean fusion	0.996	0.879	0.951	0.829	0.618	0.662	0.732
PDM-Face with 4 hypotheses	0.997	0.879	0.965	0.857	0.647	0.638	0.750
PDM-Face without the uncertain answer	0.998	0.885	0.960	0.848	0.628	0.662	0.765
PDM-Face with frozen language model	0.998	0.850	0.971	0.839	0.625	0.652	0.745
PDM-Face on Qwen3.5-2B	0.997	0.879	0.959	0.844	0.648	0.671	0.732
PDM-Face (full, seed 0)	0.997	0.864	0.966	0.851	0.637	0.659	0.743

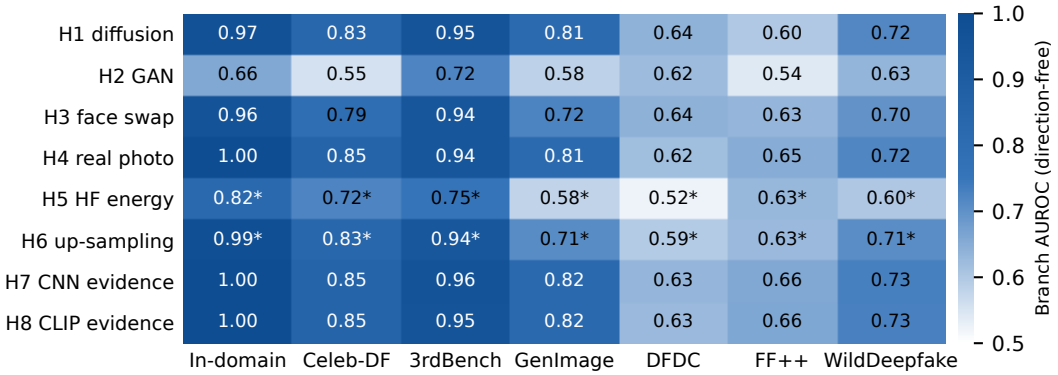

Figure 3: How well each hypothesis branch alone separates real from forged faces (seed 0, identical evaluation images). A branch's score is its "yes" minus "no" probability; values are direction-free AUROC, and * marks branches whose "yes" answer is more frequent on real faces. The fusion learns the direction and weight of each branch.

5.4 What each component contributes

Table 3 isolates the components. A VLM without forensic evidence does not detect forgeries. Zero-shot, Qwen3.5-4B scores 0.351 to 0.564 on all seven test sets. Direct fine-tuning reaches 0.942 in-domain but falls to 0.496 on Celeb-DF, because it memorizes the training generators. Removing the evidence tokens from PDM-Face has the same effect: Celeb-DF drops from 0.864 to 0.487, and 3rdBench drops from 0.966 to 0.603. The evidence tokens are thus the bridge through which forensic knowledge reaches the VLM, and the decision layer is what uses it. Replacing the learned fusion with an average of the hypothesis answers costs -0.015 on 3rdBench and -0.022 on GenImage, so the hypotheses carry different information and should be weighted. The panel is robust to its exact configuration. Using four hypotheses, removing the uncertain answer, freezing the language model, or switching to the 2-billion-parameter Qwen3.5 changes accuracy by amounts comparable to the seed-to-seed variation reported in Table 1. The design therefore does not rely on fine-tuned choices, and a smaller backbone can be used when cost matters.

5.5 What the hypothesis branches read

Each branch is an interpretable detector in its own right (Figure 3). The mechanism branches transfer beyond the families they were named after. On Celeb-DF, which contains only face swaps, the diffusion branch reaches 0.83 AUROC and the face-swap branch reaches 0.79. On GenImage the diffusion branch reaches 0.81 and the GAN branch only 0.58, in line with GenImage being dominated by diffusion generators. The two spectral branches carry a different kind of signal. On in-domain faces they separate real from forged almost perfectly (0.99 for the up-sampling branch), but with the opposite sign to their wording: their "yes" answer is more frequent on real faces. The fusion reads them accordingly. Because the branches disagree in informative ways, a learned fusion

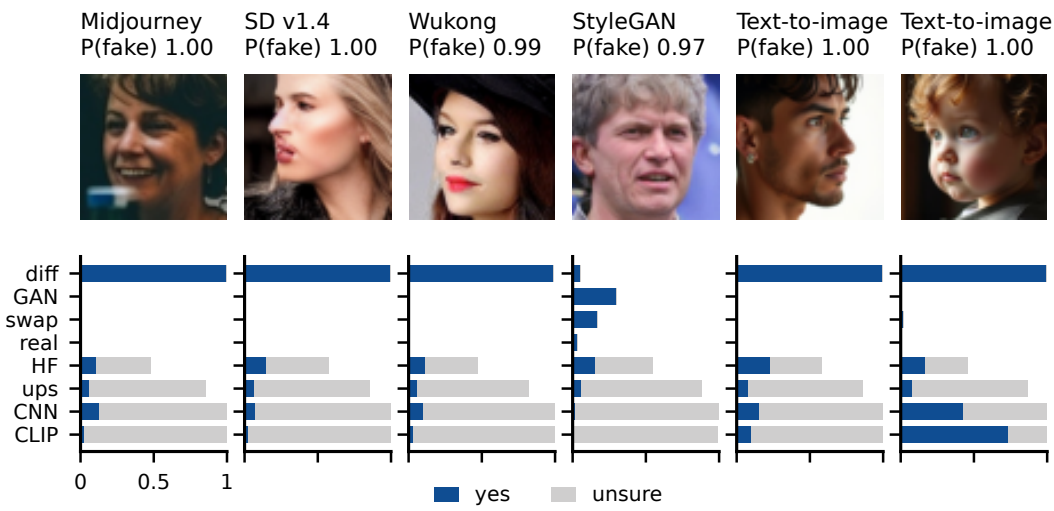

Figure 4: Forged faces from five generator families: PDM-Face's fake probability and each branch's answer ("yes" blue, "unsure" gray). The diffusion branch fires on every diffusion face; on the StyleGAN face the GAN branch gives the highest yes probability.

outperforms any single branch; on GenImage the best branch reaches 0.82 and the fused decision 0.851. The same pattern shows up per generator. On all 8 GenImage generators, from ADM to Midjourney, PDM-Face improves over its expert layer, by +0.010 to +0.034 AUROC (Appendix I); Figure 4 shows the branch answers on single faces.

6 Discussion

Why fast decisions win. Given the expert scores as text, the reasoning model still ends below those scores on every benchmark (Section 5.1), so the verbal chain loses information that was present in its input. PDM-Face never paraphrases. Its evidence enters as continuous tokens, each hypothesis is read from a hidden state, and the fusion is trained end to end. This is the regime where fast, learned judgment is more reliable than deliberation (Kahneman, 2011; Liu et al., 2025).

Interpretable without generation. Although PDM-Face writes no text, its eight answer distributions are an explanation. A single forward pass reports whether the face looks diffusion-generated, GAN-generated, or swapped, whether its spectrum is abnormal, and whether each forensic expert fired. The uncertain answers report where the model abstains, at no extra cost.

Scope. PDM-Face is designed to strengthen, not replace, a forensic expert layer; on FaceForensics++ and DFDC that layer is the weakest link, and CLIP-based detectors carry complementary evidence there (Appendix K). We evaluate still face crops; extending the hypotheses to temporal evidence for video is a natural next step.

7 Conclusion

For VLM face-forgery detection, thinking aloud is both slower and less accurate than deciding fast. PDM-Face turns a VLM into a parallel System 1 detector. It encodes the face and forensic evidence once, branches the prefix state into a panel of eight forensic hypotheses, and reads their answers in a single batched pass without generating text. On six external benchmarks it outperforms generative reasoning in every comparison while being 21 times faster. It improves its own expert layer on the hardest video benchmarks, and it outperforms recent detectors in 19 of 23 comparisons. Its parallel hypotheses also improve on a single-question decision layer across seeds and backbones on the multi-generator benchmarks.

ETHICS STATEMENT

This work aims to reduce harm from forged face images, for example in identity fraud and misinformation. All face data come from established public research datasets and were used under their licenses and terms of use. No new images of people were collected, and no human-subject study was conducted. We release no generator. Because forgery detectors can be misused to certify manipulated content as authentic, or can fail unevenly across demographic groups, we recommend that PDM-Face be used as one signal alongside human review, and that its operating threshold be audited on data representative of the deployment population before use.

REPRODUCIBILITY STATEMENT

Section 3 and Section 4 define the model, the hypothesis labels, and the training schedule, and Appendix B lists every hyperparameter. Appendix C gives the exact prompts, questions, and reasoning instruction, and Appendix D gives the equivalence test for prefix-state branching. The supplementary material contains the source code of the decision layer, the prefix-state branching for hybrid gated delta-rule models, the training and evaluation scripts, the reasoning baseline, and the scripts that regenerate every table and figure from per-image predictions. All numbers in the paper are generated from these result files, and per-seed results are given in Appendix E.

AI USE STATEMENT

We used generative AI tools for two purposes. First, to aid and polish the writing: large language model tools helped improve the wording, grammar, and readability of the text. Second, the study relies on synthetic datasets generated by AI models: the forged faces used for training and evaluation are synthetic images produced by generative models (diffusion text-to-image models, GANs, face-swapping, and reenactment systems), as released with the public datasets listed in Section 4 and Appendix A. The authors take full responsibility for the content of the paper.

REFERENCES

Davide Cozzolino, Giovanni Poggi, Riccardo Corvi, Matthias Nießner, and Luisa Verdoliva. Raising the bar of AI-generated image detection with CLIP. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024.

Xinjie Cui, Yuezun Li, Ao Luo, Jiaran Zhou, and Junyu Dong. Forensics adapter: Adapting CLIP for generalizable face forgery detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

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.

Tarik Dzanic, Karan Shah, and Freddie Witherden. Fourier spectrum discrepancies in deep network generated images. In *Advances in Neural Information Processing Systems*, 2020.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall/CRC, 1993.

Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *International Conference on Machine Learning*, 2020.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, 2017.

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.

Zhenglin Huang, Jinwei Hu, Xiangtai Li, Yiwei He, Xingyu Zhao, Bei Peng, Baoyuan Wu, Xiaowei Huang, and Guangliang Cheng. SIDA: Social media image deepfake detection, localization and explanation with large multimodal model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 1991.

Shan Jia, Reilin Lyu, Kangran Zhao, Yize Chen, Zhiyuan Yan, Yan Ju, Chuanbo Hu, Xin Li, Baoyuan Wu, and Siwei Lyu. Can ChatGPT detect DeepFakes? A study of using multimodal large language models for media forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024.

Jordan Juravsky, Bradley Brown, Ryan Ehrlich, Daniel Y. Fu, Christopher Ré, and Azalia Mirhoseini. Hydragen: High-throughput LLM inference with shared prefixes. *arXiv preprint arXiv:2402.05099*, 2024.

Daniel Kahneman. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, 2011.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, 2022.

Christos Koutlis and Symeon Papadopoulos. Leveraging representations from intermediate encoder-blocks for synthetic image detection. In *Proceedings of the European Conference on Computer Vision*, 2024.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th ACM Symposium on Operating Systems Principles*, 2023.

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

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

Huan Liu, Zichang Tan, Chuangchuang Tan, Yunchao Wei, Jingdong Wang, and Yao Zhao. Forgery-aware adaptive transformer for generalizable synthetic image detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Ryan Liu, Jiayi Geng, Addison J. Wu, Ilia Sucholutsky, Tania Lombrozo, and Thomas L. Griffiths. Mind your step (by step): Chain-of-thought can reduce performance on tasks where thinking makes humans worse. In *International Conference on Machine Learning*, 2025.

Yuchen Luo, Yong Zhang, Junchi Yan, and Wei Liu. Generalizing face forgery detection with high-frequency features. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *Proceedings of the European Conference on Computer Vision*, 2020.

Qwen Team. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

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 *International Conference on Machine Learning*, 2021.

Jonas Ricker, Denis Lukovnikov, and Asja Fischer. AEROBLADE: Training-free detection of latent diffusion images using autoencoder reconstruction error. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Andreas Rössler, 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Zayne Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. To CoT or not to CoT? chain-of-thought helps mainly on math and symbolic reasoning. In *International Conference on Learning Representations*, 2025.

Chuangchuang Tan, Huan Liu, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Rethinking the up-sampling operations in CNN-based generative network for generalizable deepfake detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Mingxing Tan and Quoc V. Le. EfficientNetV2: Smaller models and faster training. In *International Conference on Machine Learning*, 2021.

Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. CNN-generated images are surprisingly easy to spot... for now. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. DIRE for diffusion-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, 2022.

Zhipei Xu, Xuanyu Zhang, Runyi Li, Zecheng Tang, Qing Huang, and Jian Zhang. FakeShield: Explainable image forgery detection and localization via multi-modal large language models. In *International Conference on Learning Representations*, 2025.

Shilin Yan, Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Weidi Xie. A sanity check for AI-generated image detection. In *International Conference on Learning Representations*, 2025a.

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*, 2023a.

Zhiyuan Yan, Yong Zhang, Xinhang Yuan, Siwei Lyu, and Baoyuan Wu. DeepfakeBench: A comprehensive benchmark of deepfake detection. In *Advances in Neural Information Processing Systems*, 2023b.

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

Zhiyuan Yan, Taiping Yao, Shen Chen, Yandan Zhao, Xinghe Fu, Junwei Zhu, Donghao Luo, Chengjie Wang, Shouhong Ding, Yunsheng Wu, and Li Yuan. DF40: Toward next-generation deepfake detection. In *Advances in Neural Information Processing Systems*, 2024b.

Zhiyuan Yan, Jiangming Wang, Peng Jin, Ke-Yue Zhang, Chengchun Liu, Shen Chen, Taiping Yao, Shouhong Ding, Baoyuan Wu, and Li Yuan. Orthogonal subspace decomposition for generalizable AI-generated image detection. In *International Conference on Machine Learning*, 2025b.

Songlin Yang, Jan Kautz, and Ali Hatamizadeh. Gated delta networks: Improving Mamba2 with delta rule. In *International Conference on Learning Representations*, 2025.

Lianmin Zheng, Liangsheng Yin, Zhiqiang Xie, Chuyue Sun, Jeff Huang, Cody Hao Yu, Shiyi Cao, Christos Kozyrakis, Ion Stoica, Joseph E. Gonzalez, Clark Barrett, and Ying Sheng. SGLang: Efficient execution of structured language model programs. In *Advances in Neural Information Processing Systems*, 2024.

Mingjian Zhu, Hanting Chen, Qiangyu Yan, Xudong Huang, Guanyu Lin, Wei Li, Zhijun Tu, Hailin Hu, Jie Hu, and Yunhe Wang. GenImage: A million-scale benchmark for detecting AI-generated image. In *Advances in Neural Information Processing Systems*, 2023.

Bojia Zi, Minghao Chang, Jingjing Chen, Xingjun Ma, and Yu-Gang Jiang. WildDeepfake: A challenging real-world dataset for deepfake detection. In *Proceedings of the 28th ACM International Conference on Multimedia*, 2020.

A Datasets and protocol details

Training corpus. The expert layer and the decision layer are trained on a curated corpus of aligned face crops. Each crop is square, 256×256 pixels, and stored as JPEG at quality 95. Real faces come from public photograph collections. Forged faces come from public releases of diffusion text-to-image models, GANs, face-swapping, and reenactment systems. Faces smaller than the native-size gate are rejected rather than upscaled. The quality gates on detection confidence, native size, sharpness (variance of the Laplacian), pose, and CLIP faceness are identical for real and forged faces. As a result, blur, resolution, and pose carry no label information. Splits are formed by group, so the same identity or near-duplicate never crosses the train, validation, and test partitions.

Decision-layer subsets. The decision layer trains on 60,000 faces (30,000 real, 30,000 forged). The forged faces are spread evenly over generator families, and the family quota is capped by availability. Validation uses 6,000 faces from the validation split, which also serve to fit the temperature. The in-domain test set holds 12,000 faces from the test split. We score a class-stratified sample of 4,000 of them, using the same rule as for the large external sets.

External benchmarks. Table 4 lists the six external benchmarks. None of their sources is used to train any component. Every face is cropped with the detector and crop rule of the training data. For 3rdBench, the StyleGAN set provides real and fake faces. The two face-swap collections contain forgeries only and are scored against the StyleGAN set's reals. The text-to-image subset contains diffusion-generated faces.

Table 4: External benchmarks (face crops after detection).

Benchmark	Faces	Real	Forged	Forgery types
Celeb-DF (Li et al., 2020)	1,109	660	449	face swap (video)
3rdBench	33,085	8,000	25,085	StyleGAN, face swap, text-to-image
GenImage faces (Zhu et al., 2023)	20,157	12,692	7,465	8 diffusion and GAN generators
DFDC (Dolhansky et al., 2020)	30,460	14,882	15,578	face swap (video)
FaceForensics++ (Rössler et al., 2019)	21,144	3,981	17,163	face swap, reenactment (video)
WildDeepfake (Zi et al., 2020)	7,746	2,849	4,897	internet deepfake videos

Evaluation samples. For the five large benchmarks and the in-domain test set we score a class-stratified sample of up to 4,000 faces, with 2,000 per class. The sample is drawn with a fixed random seed, so every method sees the same files. Celeb-DF is scored in full. All pairwise comparisons join scores by image path, never by row order.

B Implementation details

Single-query decision layer. The single-query decision layer uses the same backbone, evidence tokens, evidence dropout, data, and schedule as PDM-Face. It asks one question, "Is this face AI-generated or manipulated?", and its score is the margin between the "yes" and "no" logits of the shared three-way head. It is supervised with the authenticity label.

Hardware. All models were trained and evaluated on NVIDIA H100 80 GB GPUs, inside a container with PyTorch 2.10, Transformers 5.5, and vLLM 0.19. One PDM-Face training run takes about 20 minutes with prefix-state branching during training and about 1.5 hours with gradient checkpointing and full recomputation. The two paths compute the same function.

C Prompts and hypothesis questions

Shared prefix. The prefix is the model's chat template applied to one user turn that holds the image, followed by the text "Forensic evidence:", the k evidence tokens, and "Answer each question about the face above. Question:".

Table 5: Hyperparameters of PDM-Face (identical for the single-query decision layer unless noted).

Setting	Value
Backbone	Qwen3.5-4B (hybrid attention and gated delta-rule layers); Qwen3-VL-4B in Appendix H
Image resolution	224×224 (49 visual tokens after merging)
LoRA	rank 16, $\alpha = 32$, dropout 0.05, on attention, delta-rule, and MLP projections of the language model
Evidence adapter	LayerNorm, Linear($1315 \to 1024$), GELU, Linear($1024 \to 4d$); $k = 4$ tokens
Answer head	LayerNorm, Linear($d \to 3$), shared by all hypotheses
Fusion	Linear($3M + 3 \to 64$), GELU, Linear($64 \to 1$)
Trainable parameters	42.4 million
Optimizer	AdamW, weight decay 0.01, gradient clipping at 1.0
Learning rates	10^{-4} (LoRA), 10^{-3} (adapter, question embeddings, head, fusion)
Schedule	1,250 steps, batch size 16, 3% linear warm-up, cosine decay
Hypothesis loss weight	$\lambda = 0.5$ (single query: its one question replaces the fusion)
Evidence dropout	0.5 per image during training
Calibration	temperature fitted on the validation set with L-BFGS
Precision	bfloat16 autocast; float32 heads
Seeds	0, 1, 2

Table 6: The hypothesis panel and the source of each label.

	Question	Label source
H_1	Was this face produced by a diffusion or text-to-image generator?	generator family
H_2	Was this face produced by a GAN?	generator family
H_3	Was this face swapped or blended into another person's photo?	generator family
H_4	Is this an unedited photograph of a real person?	real or forged
H_5	Is the high-frequency spectral energy abnormal compared with real camera photos?	high-frequency ratio h, training-real percentiles
H_6	Does the spectrum show periodic up-sampling peaks?	peak score q, training-real percentiles
H_7	Does the forensic CNN evidence indicate blending or synthesis traces?	p_{cnn}, training percentiles
H_8	Does the semantic evidence indicate generated content?	p_{clip}, training percentiles

Hypothesis questions. Each branch appends one question and the text "Answer:".

Reasoning baseline. The system message reads: "You are a forensic analyst. Examine the face image for signs of AI generation or manipulation (blending boundaries, texture, lighting, eyes, teeth, hair, background, frequency artifacts). Reason step by step, then answer the question with Yes or No." The user turn holds the image and the question "Is this face AI-generated or manipulated?". In the variant with expert scores, the question is preceded by: "Forensic expert scores (probability the face is fake): CNN artifact detector a, CLIP semantic detector b; learned gate weight on the CNN detector g. The experts can be wrong on unseen generators; weigh them against what you see." The model runs in thinking mode with greedy decoding, up to 1,536 reasoning tokens. The closed reasoning is followed by "Answer:", and the score is $P(\text{Yes})/(P(\text{Yes}) + P(\text{No}))$ over the top-20 next-token log-probabilities.

D Equivalence and cost of prefix-state branching

We compare three computations on the same batch (Figure 5): branched (the prefix runs once and its state is copied into $M = 8$ branches), replicated (each of the M prefix-plus-question rows is recomputed from scratch), and sequential (eight separate single-hypothesis models share the weights

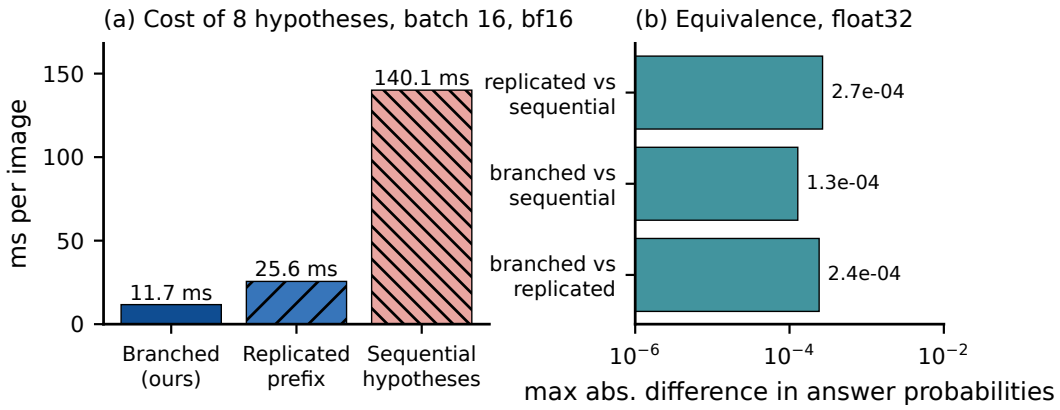

Figure 5: (a) Cost of answering the eight hypotheses for one image at batch size 16 in bfloat16 on one H100: branched prefix state, fully replicated prefix, and eight sequential single-hypothesis passes. (b) Maximum absolute difference between the three-way answer probabilities of the three computations in float32.

and run one after another). In float32, the maximum absolute difference in the three-way answer probabilities is 2.4×10^{-4} between branched and replicated, 1.3×10^{-4} between branched and sequential, and 2.7×10^{-4} between replicated and sequential. In bfloat16 the per-image cost at batch size 16 on one H100 is 11.7 ms (branched), 25.6 ms (replicated), and 140.1 ms (sequential).

Branching a gated delta-rule layer. A gated delta-rule layer applies a short causal convolution of width K to the projected inputs before the recurrence. To continue from a cached prefix with an L-token tail, we prepend the last $K - 1$ cached inputs to the new inputs, convolve, and keep the last L outputs. The recurrence is then initialized from the cached state. Both the convolution buffer and the recurrent state are copied M times along the batch dimension together with the attention key-value cache.

E PER-SEED RESULTS

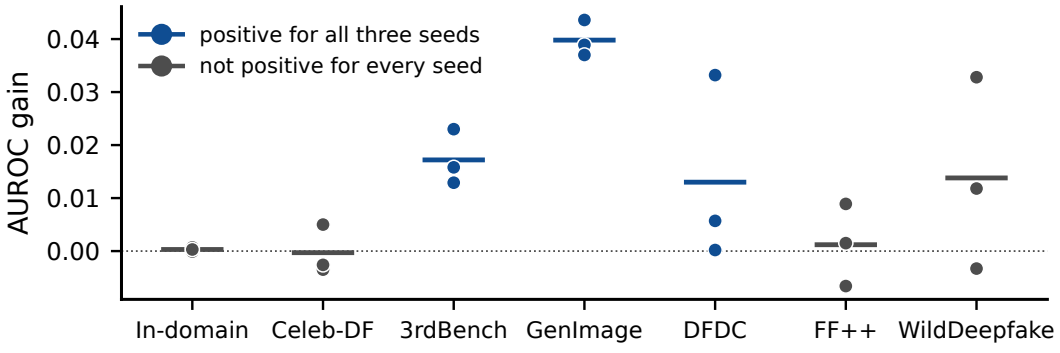

Figure 6: Seed-paired AUROC gain of PDM-Face over the single-query decision layer on identical images. Dots are seeds and bars are means; blue marks benchmarks where all three seeds are positive.

Table 7: Per-seed AUROC for the single-query decision layer and PDM-Face, and their seed-paired difference on identical images.

Set	Decision layer	Seed 0	Seed 1	Seed 2
In-domain	Single query, with evidence	0.997	0.997	0.997
In-domain	PDM-Face, 8 parallel hypotheses	0.997	0.998	0.997
	Paired gain	-0.000	+0.001	+0.000
Celeb-DF	Single query, with evidence	0.868	0.825	0.873
Celeb-DF	PDM-Face, 8 parallel hypotheses	0.864	0.830	0.870
	Paired gain	-0.004	+0.005	-0.003
3rdBench	Single query, with evidence	0.953	0.949	0.943
3rdBench	PDM-Face, 8 parallel hypotheses	0.966	0.964	0.966
	Paired gain	+0.013	+0.016	+0.023
GenImage	Single query, with evidence	0.807	0.789	0.808
GenImage	PDM-Face, 8 parallel hypotheses	0.851	0.828	0.845
	Paired gain	+0.044	+0.039	+0.037
DFDC	Single query, with evidence	0.631	0.587	0.622
DFDC	PDM-Face, 8 parallel hypotheses	0.637	0.620	0.622
	Paired gain	+0.006	+0.033	+0.000
FF++	Single query, with evidence	0.666	0.637	0.647
FF++	PDM-Face, 8 parallel hypotheses	0.659	0.639	0.656
	Paired gain	-0.007	+0.002	+0.009
WildDeepfake	Single query, with evidence	0.732	0.726	0.714
WildDeepfake	PDM-Face, 8 parallel hypotheses	0.743	0.723	0.747
	Paired gain	+0.012	-0.003	+0.033

F GENERATIVE REASONING: FULL RESULTS

Table 8: Reasoning baseline, expert layer, and PDM-Face on the same faces. The interval is the paired bootstrap 95% confidence interval of PDM-Face minus reasoning with expert scores.

Set	n	Reasoning (image)	Reasoning (+ scores)	Expert only	PDM-Face	PDM minus reasoning (+ scores), 95% CI
Celeb-DF	949	0.508	0.688	0.834	0.860	[+0.138, +0.207]
3rdBench	1000	0.598	0.864	0.983	0.973	[+0.086, +0.133]
GenImage	1000	0.690	0.737	0.832	0.858	[+0.094, +0.148]
DFDC	1000	0.502	0.499	0.608	0.629	[+0.089, +0.168]
FF++	1000	0.537	0.578	0.616	0.666	[+0.051, +0.124]
WildDeepfake	1000	0.562	0.647	0.748	0.741	[+0.057, +0.128]

On average the reasoning model generated 998 tokens per face. In 26.9% of requests it reached the 1,536-token limit, and its answer was then read immediately after the truncated reasoning. The batched throughput of reasoning was measured over all 11,898 requests in one continuous batch. Its batch-one latency is the median over 32 consecutive single requests, and PDM-Face's is the median over 32 consecutive single images, including image loading and preprocessing. Appendix L reproduces a complete reasoning trace.

G FULL TEST SETS

Table 9: Expert layer and PDM-Face (seed 0) on the complete test sets, with paired bootstrap intervals of the gain.

Set	n	Expert layer only	PDM-Face	Gain	95% CI
DFDC	30,460	0.613	0.634	+0.021	[+0.018, +0.024]
FF++	21,144	0.618	0.652	+0.034	[+0.030, +0.038]
WildDeepfake	7,746	0.746	0.742	-0.004	[-0.009, +0.002]

H SECOND BACKBONE

Qwen3-VL-4B is a full-attention VLM that injects multi-level visual features into the first language layers. For branching, these features and their positions are copied along with the prefix. Because the model has no recurrent layers, the key-value cache alone carries the prefix state. We train it with prefix-state branching in the forward pass and without gradient checkpointing. This matches the other runs in function but is 3.3 times faster.

Table 10: Backbone comparison, seed 0 (AUROC). The last row is the paired difference between PDM-Face and the single-query decision layer on Qwen3-VL-4B.

Backbone	Decision layer	In-domain	Celeb-DF	3rdBench	GenImage	DFDC	FF++	WildDeepfake
Qwen3.5-4B	Single query	0.997	0.868	0.953	0.807	0.631	0.666	0.732
Qwen3.5-4B	PDM-Face	0.997	0.864	0.966	0.851	0.637	0.659	0.743
Qwen3-VL-4B	Single query	0.998	0.889	0.957	0.816	0.629	0.684	0.732
Qwen3-VL-4B	PDM-Face	0.997	0.869	0.963	0.840	0.620	0.645	0.746
Qwen3-VL-4B	Paired gain	-0.000	-0.020	+0.006	+0.024	-0.009	-0.039	+0.014

I PER-SOURCE RESULTS, BRANCH ACCURACY, AND EXAMPLES

Figure 7 breaks the gain of the decision layer over its expert layer down by forgery source, and Table 11 lists the GenImage values. On all 8 GenImage generators the gain is between +0.010 and +0.034 AUROC. On FaceForensics++, AUROC rises from 0.644 to 0.692 on face swaps and from 0.578 to 0.596 on reenactment. Figure 4 in the main text shows the eight branch answers for single forged faces, and Table 12 reports how well each branch fits its own label on the validation set.

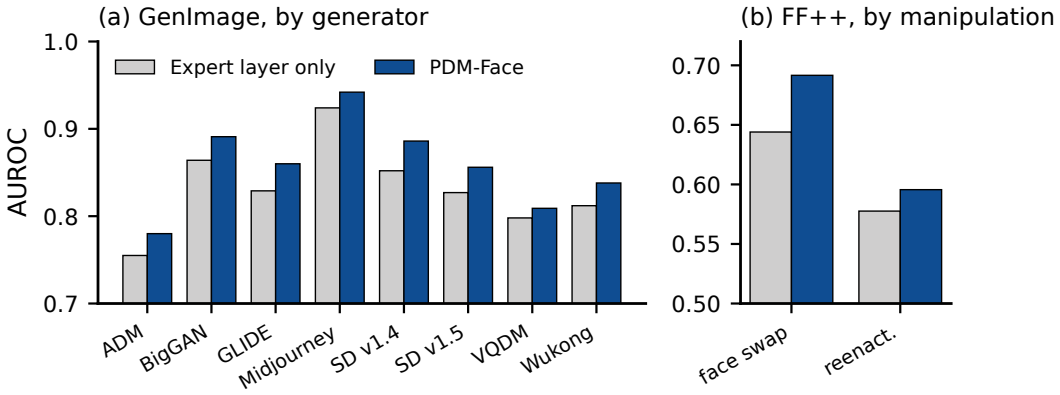

Figure 7: Expert layer and PDM-Face (seed 0) on the same images, broken down by forgery source. (a) GenImage faces, one generator's forgeries against all real faces of the sample. (b) FaceForensics++, face-swap and reenactment forgeries against all real faces. PDM-Face improves over its expert layer for every source.

Table 11: GenImage faces by generator: expert layer and PDM-Face (seed 0) on the same images. Each column scores one generator's forgeries against all real faces of the sample.

GenImage generator	ADM	BigGAN	GLIDE	Midjourney	SD v1.4	SD v1.5	VQDM	Wukong
Expert layer only	0.755	0.864	0.829	0.924	0.852	0.827	0.798	0.812
PDM-Face	0.780	0.891	0.860	0.942	0.886	0.856	0.809	0.838
Gain	+0.025	+0.028	+0.031	+0.017	+0.034	+0.029	+0.010	+0.027

Table 12: Three-way validation accuracy of each hypothesis branch (seed 0, 6,000 validation faces). The spectral hypotheses H_5 and H_6 have three nearly balanced label classes defined by percentile bands, so their chance level is about one third. At the level of the fused decision, the validation AUROC is 0.997, the expected calibration error is 0.008, and the fitted temperature is 0.917.

	H_1	H_2	H_3	H_4	H_5	H_6	H_7	H_8
Validation accuracy (three-way)	0.972	0.995	0.994	0.977	0.478	0.786	0.809	0.834

J Comparison with recent detectors: paired differences

Figure 8 shows, for each detector and benchmark of Table 2, the AUROC of PDM-Face minus that of the detector on the same images. Significance uses the paired, class-stratified bootstrap of Section 4 with 1,000 resamples. PDM-Face is significantly better in 19 of the 23 pairs.

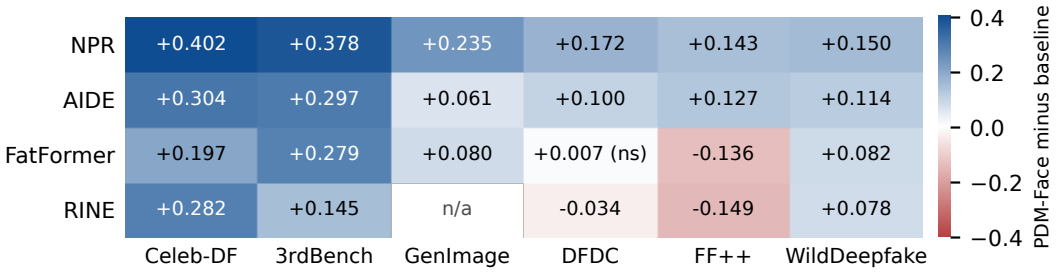

Figure 8: AUROC difference between PDM-Face (seed 0) and each detector on identical images. Every cell is significant under the paired bootstrap unless marked (ns). NPR and AIDE are retrained on our face corpus; FatFormer and RINE use released weights. AIDE on GenImage covers 454 images.

K Complementarity with CLIP-based detectors

To test whether PDM-Face and the CLIP-based detectors make different errors, we average their scores without any fitted parameter. On each benchmark, every score is replaced by its normalized rank among the scored images, and the two ranks are averaged. No label and no weight is used. Table 13 reports the mean over the three PDM-Face seeds, and significance uses the paired bootstrap of seed 0.

Table 13: Parameter-free rank-average fusion of PDM-Face with a CLIP-based detector on identical images (AUROC, mean over three PDM-Face seeds). Bold marks fusions that are significantly better than both of their inputs.

		Celeb-DF	3rdBench	GenImage	DFDC	FF++	WildDeepfake
RINE	RINE alone	0.582	0.821	–	0.671	0.808	0.666
	PDM-Face alone	0.855	0.965	–	0.626	0.651	0.738
	PDM-Face ⊕ RINE	0.792	0.969	–	**0.686**	0.778	0.742
FatFormer	FatFormer alone	0.667	0.687	0.771	0.630	0.796	0.662
	PDM-Face alone	0.855	0.965	0.841	0.626	0.651	0.738
	PDM-Face ⊕ FatFormer	0.832	0.920	**0.899**	**0.663**	0.785	0.744

On DFDC the fusion significantly exceeds both of its inputs for both detectors: 0.686 against 0.671 for RINE alone (95% CI of the difference [+0.008, +0.033]) and 0.663 against 0.630 for FatFormer alone ([+0.026, +0.051]). On FaceForensics++ it raises PDM-Face by +0.127, to 0.778 with RINE and 0.785 with FatFormer. On GenImage, fusion with FatFormer reaches 0.899, again significantly above both inputs. Where PDM-Face already leads a detector by a wide margin, an equal-weight average with the weaker detector can lower the score (Celeb-DF with either detector, and 3rdBench

with FatFormer). We therefore read the fixed fusion as evidence of complementary signals rather than as a detector of its own.

L A COMPLETE REASONING TRACE

The trace below is the full reasoning for the face of Figure 1, a Stable Diffusion v1.4 forgery from GenImage, with the expert scores in the prompt (CNN 0.900, CLIP 0.638, gate 0.82). It is regenerated with the baseline's prompt and greedy decoding; formatting marks are removed, lines are wrapped, and the closing answer-formulation step, which restates the conclusion, is omitted. The model reads the confident CNN score, attributes it to the image's grain, and concludes that the face is real, with $P(\text{fake}) = 0.001$. PDM-Face scores the same face at 0.997.

```
The user wants me to determine if the provided image is AI-generated or manipulated.
1. Analyze the Input Image:
*   Visual Style: The image is grainy, high-contrast, and monochromatic (sepia/brown tones).
  It looks like an old photograph, possibly from the early 20th century.
*   Subject: A man wearing a cap, looking slightly to the side.
*   Texture/Noise: There is significant film grain and noise throughout the image. This is a
  strong indicator of an old photograph.
*   Edges/Details:
    *   Eyes: The eyes look natural, with realistic reflections and depth. They don't have
      the "smiling eyes" or overly perfect symmetry often seen in AI generations.
    *   Hair/Beard: The facial hair is textured and irregular, consistent with a real
      person. The hair under the cap is also somewhat indistinct but plausible.
    *   Skin: The skin texture is rough and grainy, not the smooth, plastic-like look of
      modern AI faces. There are natural imperfections.
    *   Background: The background is indistinct and blurry, typical of old portraits where
      the focus is on the subject.
    *   Lighting: The lighting is soft and diffuse, consistent with flash photography or
      studio lighting of that era. Shadows fall naturally.
2. Evaluate the Forensic Scores:
*   CNN artifact detector: 0.900: This is a very high score for "fake." However, CNNs
  trained on modern datasets (like DeepFake or Midjourney datasets) often struggle with old,
  grainy images. They might interpret the noise or compression artifacts as "AI artifacts."
*   CLIP semantic detector: 0.638: This is a moderate score. CLIP measures semantic
  similarity. It might be flagging the image because it doesn't match the "real photo"
  distribution of modern datasets, or perhaps it's just a neutral score.
*   Learned gate weight: 0.82: This suggests the model is leaning heavily towards "fake"
  based on its training.
3. Synthesize Findings:
*   The Conflict: The AI detectors are screaming "Fake" (0.900), but the visual evidence
  screams "Old Photo."
*   Why the detectors might be wrong:
    *   Domain Shift: The detectors are likely trained on modern synthetic images. An old,
      grainy photo has a completely different statistical distribution (high frequency
      noise, low resolution, specific color palette) that might trigger false positives.
    *   Compression Artifacts: The image looks like it has been compressed or scanned from a
      physical print. These artifacts (blockiness, noise) can be mistaken for generative
      artifacts.
    *   Historical Context: This looks like a standard historical portrait. The
      "imperfections" are actually signs of age and film stock.
4. Conclusion:
*   The visual evidence strongly points to this being a genuine, old photograph. The grain,
  the specific type of noise, the natural imperfections in the face, and the overall
  aesthetic are consistent with early 20th-century photography.
*   The high CNN score is likely a false positive due to the image's age and texture
  characteristics confusing the modern detector.
*   Therefore, the face is *not* AI-generated or manipulated in the modern sense. It is a
  real, historical image.
```